arXiv:2603.07179v1 [cs.IR] 7 Mar 2026

# 利用图基础模型检索最小且充分的推理子图，实现路径感知 GraphRAG

Haonan Yuan[1], Qingyun Sun[1], Junhua Shi[1], Mingjun Liu[2], Jiaqi Yuan[1], Ziwei Zhang[1], Xingcheng Fu[3], Jianxin Li[1]1北京航空航天大学 [2]中国电子科技大学 [3]广西师范大学

## 摘要

基于图的检索增强生成（GraphRAG）利用结构化知识支持知识密集型推理。然而，大多数现有方法将图视为中间产物，而少数基于子图的检索方法则依赖启发式规则与领域特定分布的结合。在目标领域数据稀缺的典型冷启动场景中，这些方法表现不佳，因而得到的推理上下文要么信息不完整，要么结构冗余。在这项工作中，我们从结构视角重新审视检索，并提出**Gfm-Retriever**，直接以子图响应用户查询；其中，预训练**图** **基**础**模**型充当跨域检索器，用于多跳路径感知推理。基于这一视角，我们将预训练 GFM 从实体排序函数改造为通用检索器，以支持跨域检索。对于检索到的图，我们进一步基于原则性的信息瓶颈目标函数，构建无标签子图选择器，以识别查询条件化子图；该子图包含信息充分且结构最简的黄金证据，并形成一个独立完整的“核心集”。为将结构与生成相衔接，我们明确提取并重组关系路径，将其作为上下文内提示，从而实现可解释推理。在多跳问答基准测试上的大量实验表明，Gfm-Retriever 在检索质量和答案生成方面均达到最先进水平，同时保持了高效率。

## 1 引言

检索增强生成（RAG）已成为一种核心范式，使大语言模型（LLM）能够在广泛的现实应用中访问外部知识并进行推理 [13, 89, 91, 95]。为更好地支持知识密集型问答中的多跳推理，近期研究通过引入知识图谱扩展了 RAG 范式，由此形成了基于图的检索增强生成（GraphRAG）[

109]。尽管有这一动机，大多数 GraphRAG 系统最终仍将检索输出简化为实体和文档的排序列表 [21, 66]。这种基于排序的检索抽离了关系依赖，迫使下游答案生成从孤立证据中隐式重建推理路径。

我们在图 1 中展示了一个案例研究，其中提出了一个用户问题，其答案取决于追踪实体 [16] 之间的多重关系依赖。虽然实体排序可以遍历各自相关的节点，但它无法说明这些实体如何相互作用，也无法指出哪些关系路径对推理至关重要。因此，推理过程仍然隐含且脆弱。相比之下，查询特定子图能够明确保留关系结构，使多跳依赖清晰可见，并为生成提供核心结构上下文 [46]。这一比较

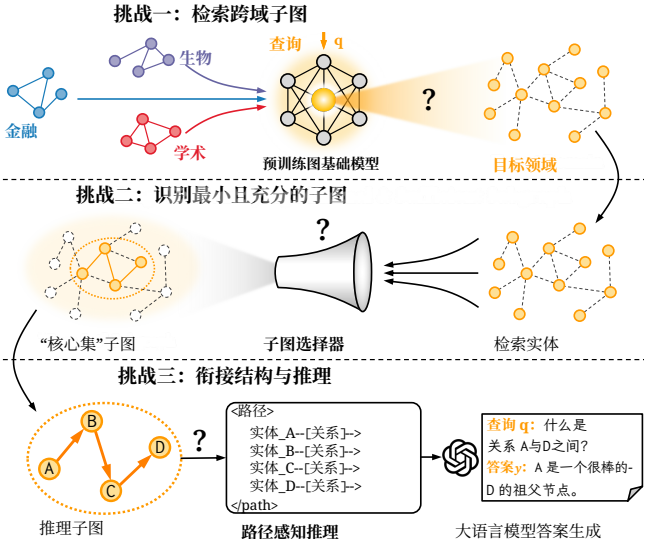


**图 1：基于子图的检索增强生成中的挑战。**

表明，对于复杂推理，针对子图而非扁平排序作答，能够提供信息更丰富的检索接口。

受这一关键观察的启发，越来越多的 GraphRAG 方法开始纳入子图级检索，以更好地支持复杂推理 [14, 48]。这些工作并非返回孤立的实体或文档，而是旨在构建能够揭示关系依赖的查询相关子结构。然而，它们通常依赖启发式规则来决定纳入哪些实体和关系 [108]。关键在于，这些启发式规则往往与特定领域分布隐式耦合，因而在目标域数据稀缺的典型冷启动场景中较为脆弱。在缺乏充分监督的情况下，它们缺少一种有原则的机制来平衡信息充分性与结构最小化。因此，它们有可能生成信息上不完整或结构上冗余的推理上下文。

上述观察表明：尽管子图相比已排序实体能够提供更忠实的中间表示，但仍有若干根本性挑战尚未得到解决：

**挑战一：检索跨域子图。**在现有 GraphRAG 系统中，检索与固定知识图谱 [24, 57] 上的查询特定传播紧密耦合。尽管这类设计在封闭场景中效果显著，但面对分布各异的未见领域时，往往难以泛化。在典型的冷启动场景中，由于数据稀缺，通常无法训练领域特定检索器；而直接应用预训练检索器又往往会因分布失配而失败。因此，关键挑战在于建立一种泛化检索范式，使学到的检索模式与特定领域脱钩。

**挑战二：识别最小且充分的子图**。检索子图带来了充分性与冗余之间的根本性权衡。较大的子图更有可能包含所有相关证据，但也会引入大量冗余；而较小的子图则可能遗漏关键推理路径。现有方法通常通过启发式预算、固定跳数限制或手动微调的剪枝策略 [48], 来约束这种权衡，但这些方法对查询难度不敏感，也难以泛化。如果没有针对何为“真实子图”的明确监督，要学习一种有原则且能适应查询的充分性定义仍然颇具挑战。

**挑战三：桥接结构与推理**。 即使已检索到查询相关子图，其结构信息也往往未被下游生成器充分利用。标准做法常常将检索子图坍缩为文档集或扁平化的三元组列表，从而无意中丢弃对多跳推理至关重要的关系依赖 [111]。因此，大语言模型被迫从碎片化的上下文中隐式重建推理链，使该过程不可解释。一个实际挑战是保留并重组这些关系路径，以显式引导大语言模型的推理过程。

为应对上述挑战，我们提出**Gfm-Retriever**，一种基于子图的检索增强生成框架，它从结构视角重新审视检索，直接以子图响应用户查询。与将检索同查询时的图遍历或启发式扩展紧密耦合的现有方法不同，Gfm-Retriever将子图提升为检索接口本身，并以原则性的方式学习如何检索它们。
Gfm-Retriever的核心是一个预训练的**G**图**F**基础**M**模型，它充当通用**检索器**，能够实现跨领域高效且可迁移的子图检索。为解决缺乏原则性子图选择的问题，我们推导出一个在信息瓶颈目标函数下优化的无标签子图选择器，显式平衡信息充分性与结构最小化。最后，为使检索到的结构能够服务于答案生成，Gfm-Retriever从子图中提取并重组关系路径，形成结构化的上下文内提示，从而使图结构能够直接引导多跳推理。**我们的贡献：**

- 我们提出了 Gfm-Retriever，这是一种基于子图的检索增强生成方法，用于检索查询特定子图。据我们所知，Gfm-Retriever 是首个利用跨领域图基础模型学习最小但充分的子图以进行答案生成的工作。
- 我们提出了一种无标签信息瓶颈的子图选择形式化方法，并推导出一个易于处理的优化目标，从而能够以有原则的方式识别查询相关结构，显式支持多跳、路径感知推理。
- 大量实验表明，与最先进的基线方法相比，Gfm-Retriever 在检索质量和端到端答案生成方面均取得了持续改进。

## 2 相关工作

### 2.1 基于图的检索增强生成

检索增强生成（RAG）通过引入外部检索知识，为大语言模型提供依据，从而提升其在知识密集型任务中的事实准确性 [28, 114]。大多数方法遵循“先检索、后生成”的范式，将检索到的文本作为上下文 [13], , 但通常将证据视为非结构化信息。基于图的检索增强生成（GraphRAG）将图引入为知识结构，以显式编码关系信息 [109]。通过将实体组织为由边连接的节点，GraphRAG 能够实现关系感知检索与推理，并已展现出对多步推理的益处 [12, 22, 66]。然而，现有 GraphRAG 方法依赖庞大或固定的结构，容易引入冗余或与任务无关的信息，这凸显了构建任务相关图结构所面临的挑战。

### 2.2 图基础模型

图基础模型（GFMs）旨在通过大规模自监督预训练学习可迁移的图表示 [52, 93]。尽管在有限领域内有效，大多数方法低估了预训练与下游之间的分布偏移，从而导致可迁移性下降 [29]。近期工作探索跨域图学习与大语言模型增强对齐，以提升泛化能力 [25, 81, 113], , 但往往依赖隐式表示相似性或受限的图假设。尽管一些研究将图基础模型用作检索增强生成系统的检索器 [54], , 其预训练仍是缺乏显式跨域对齐的语义“黑箱”，从而限制了其泛化能力。

### 2.3 关键子图挖掘

尽管图结构编码了丰富的关系语义，但只有一小部分子图真正具有信息量，并可能对下游任务具有决定性作用 [98]。早期工作依赖预定义的结构模式，例如核或模体，它们提供了较强的归纳偏置，但仍停留在启发式层面，且难以跨任务扩展或适配 [41, 59]。较新的方法则探索可学习准则或信息论准则，以识别任务相关子图 [76 , 104], 但通常缺乏明确的语义保持或最小性保证。在 GraphRAG 中，子图很少被用作检索单元，现有尝试关注的是覆盖范围，而非任务充分性，因此识别紧凑且关键的子图这一问题仍未得到解决 [48]。

## 3 预备知识

**符号说明**。知识图谱记为 $G = (V, R, E)$，其中 V 为实体集，R 表示关系，$E \subseteq V \times R \times V$ 表示关系三元组。G 从不同知识领域 {D} 的文档语料库 $C = \{c_i\}$ 中建立索引。每个 $c_i$ 都链接到一组实体 $\phi(c_i) \subseteq V$。将图基础模型表示为 $f_\theta(\cdot)$，其中参数 $\theta$ 在 G 上针对知识图谱补全任务进行预训练，然后针对子图检索任务进行微调。给定查询 $q$，$f_{\star\theta}(\cdot)$ 充当查询相关检索器，在目标知识图谱 $G^T$ 上检索子图 $G_q = f_{\star\theta}(G^T, q)$。

**基于子图的检索增强生成**。 给定查询 $q$，检索到的子图 $G_q$ 记为 $G_q = (V_q, R_q, E_q)$，其中同时指定了相关实体及其依赖关系。为将检索到的实体映射到文本中，我们将实体映射回其所属语料库，从而得到支持文档集 $C_q \subseteq C$，表示为 $C_q = \bigcup_{e \in V_q} \phi^{-1}(e) = \{c \in C \mid \phi(c) \cap V_q \neq \emptyset\}$，其中 $\phi^{-1}(e)$ 表示提及实体 $e$ 的文本块。下游答案生成器是一个大语言模型，它以查询和检索上下文为条件，生成答案 $y = \text{LLM}(q, G_q, C_q)$。

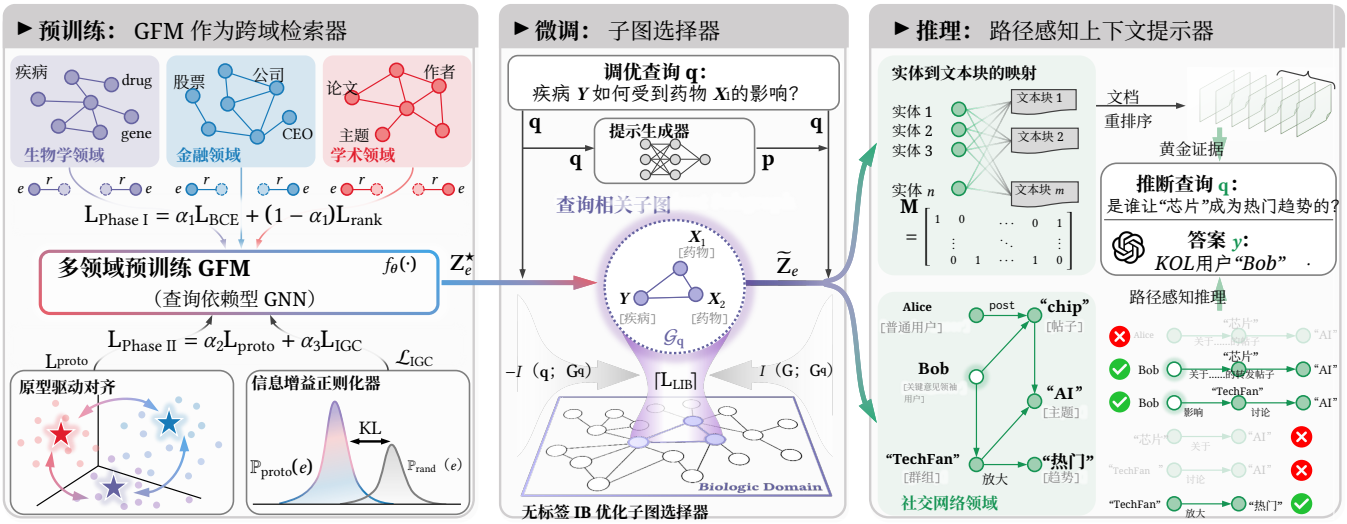


**图 2：Gfm-Retriever 框架概览。(1) 预训练：将查询条件化 GFM 训练为跨域检索器。(2) 微调：经无标签信息瓶颈优化的选择器识别出最小且充分的查询特定子图。(3) 推理：将检索到的子图转换为路径感知上下文提示，以引导多跳推理。点击“►”进行跳转。**

## 4 提出框架：Gfm-Retriever

本节将介绍所提出的 Gfm-Retriever。

### 4.1 广义图基础模型作为跨域检索器

为解决数据稀缺和冷启动场景下的跨域检索问题，我们构建了一种基于 GFM 的检索器。它通过在多领域知识图谱之间学习统一且具有区分性的表示空间，实现跨不同知识领域的泛化，并支持一致的检索。

**查询依赖型图基础模型消息传递**。与不依赖任务、独立进行聚合的传统图神经网络消息传递不同， [40], GFM 检索器采用查询依赖型消息传递机制，可根据输入查询动态调整 [54]。给定查询 $\mathbf{q}$，GFM 检索器 $f_\theta(\cdot)$ 首先为所有 $e \in \mathrm{V}$、 $r \in \mathrm{R}$ 中的实体（$e$）和关系（$r$）初始化嵌入：

$$\mathbf{Z}_e^{(0)} = \begin{cases} \mathbf{q}, & e \in \{\mathcal{V} \mid \mathbf{q}\}, \\ \mathbf{0}, & \text{otherwise}, \end{cases} \text{ and } \mathbf{Z}_r^{(0)} = \textsc{TextEmb}(r), \tag{1}$$

其中，$\mathbf{Z}_e$ 和$\mathbf{Z}_r$ 分别表示实体和关系的嵌入， $e \in \{\mathrm{V} \mid \mathbf{q}\}$表示查询$\mathbf{q}$中提及的实体集合。这种初始化会提高与查询相关实体的激活，为后续传播提供上下文起点。随后，第 $(l+1)$ 层的消息传递通过关系感知三元组编码、实体级聚合和表示更新进行。

$$\mathbf{m}_v^{(l+1)} = \textsc{Msg}\big(\mathbf{Z}_v^{(l)}, g^{(l+1)}(\mathbf{Z}_r^{(l)}), \mathbf{Z}_u^{(l)}\big),\ (u,r,v) \in \mathcal{E}, \tag{2}$$

$$\mathbf{Z}_v^{(l+1)} = \textsc{Update}\big(\mathbf{Z}_v^{(l)}, \textsc{Agg}(\{\mathbf{m}_u^{(l+1)} \mid u \in \mathcal{N}(v)\})\big), \tag{3}$$

其中，消息$\mathbf{m}$ 捕获相邻实体与关系之间的三元组级交互。具体而言，Msg(·)、投影函数 $g(\cdot)$、Update(·) 和 Agg(·) 是 GFM $f_\theta$(·) 的可学习组件，并在已索引的多域知识图谱上联合预训练。

查询依赖型图神经网络在理论上能够进行多跳推理，但它们假设知识图谱 101] [38, 具有语义一致性。然而，由于我们设定中的知识图谱 G 是从语义分布各异的多域文档中索引而来，这可能会打破该假设。接下来，我们通过推导 Qiu 等人 [68] 的方法来弥合这一对齐差距。

命题 *4.1*（查询条件化 *GFM* 的多域逻辑表达能力）。设 $G$ 为从多个领域索引得到的知识图谱。$\{Dom_d(e)\}$ 表示指示实体 $e$ 域属性的一元谓词。对于查询 $q$，当且仅当 $R(q\ e)$ 是 $G$ 上以 $q$ 为常量的分级模态逻辑公式时，查询条件化 *GFM* 才能学习规则 $R(q\ e)$，即 $\exists^{\geq N} e'(R(e', e) \wedge Dom_d(e') \wedge \psi(e'))$，其中 $\psi(\cdot)$ 是一个子公式。[▷证明]

通俗地说，命题 4.1 表明，在多领域知识图谱上，查询条件化 GFM 能够表示的恰好是那些可由分级模态逻辑定义的查询-实体关系，其中查询 $\mathbf{q}$ 充当常量。这涵盖了复杂查询所需的多跳关系推理和领域感知推理。

**原型驱动的多领域预训练**。 理想的 GFM 应在不同任务间呈现同质化，并涌现可迁移知识 [51],，从而能够在分布偏移下充当跨域检索器。为此，我们使用知识图谱补全任务对 $f_\theta$(·) 进行预训练。对于从不同领域的 G 中随机采样的每个掩码三元组（$e$，$r$，?）或（? $r$，$e$），$f_\theta(\cdot)$ 接收一个由已知元素构造的伪查询 $\mathbf{q}$，并预测缺失实体及其相关性分数 [54]:

$$\mathbb{P}_{\mathbf{q}}(e) = \sigma\big(\mathrm{MLP}(\mathbf{Z}_e^{(L)})\big) \in \mathbb{R}^{|\mathcal{V}|\times 1}, \tag{4}$$

从而得到二元交叉熵（BCE）损失:

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{|\mathcal{V}_{\mathbf{q}}^{+}|} \sum_{e \in \mathcal{V}_{\mathbf{q}}^{+}} \log \mathbb{P}_{\mathbf{q}}(e) - \frac{1}{|\mathcal{V}_{\mathbf{q}}^{-}|} \sum_{e \in \mathcal{V}_{\mathbf{q}}^{-}} \log(1 - \mathbb{P}_{\mathbf{q}}(e)), \tag{5}$$

其中$V_q^+$表示与查询$q$相关的实体集合，$V_q^-$表示其余实体。为进一步缓解正实体稀疏性所导致的梯度消失问题，我们额外引入排序损失，以在正实体与负实体之间强制保持间隔 [3, 54]:

$$\mathcal{L}_{\text{rank}} = -\frac{1}{|\mathcal{V}_q^+|}\sum_{e\in\mathcal{V}_q^+}\frac{\mathbb{P}_q(e)}{\sum_{e'\in\mathcal{V}_q^-}\mathbb{P}_q(e')}. \tag{6}$$

为提升跨域语义一致性，我们引入一种轻量级的基于原型的对比式持续预训练机制。对于每个预训练领域 $d \in \{D\}$，我们均匀采样一部分查询-实体对（$q_i$，$e_i$）$\in Q_d$，并将领域 $d$ 的语义原型$c_d$定义为相应实体的平均嵌入$\{e_i\}$。原型驱动目标如下:

$$\mathcal{L}_{\text{proto}} = \sum_{(q_i,e_i,d^+)\in\bigcup_d\{Q_d\}} \log\frac{\exp\left(\text{sim}(Z_{e_i}^{(L)}, c_{d^+})/\tau\right)}{\sum_{d\in\mathcal{D}}\exp\left(\text{sim}(Z_{e_i}^{(L)}, c_d)/\tau\right)}, \tag{7}$$

其中，$\{c_d\}_{d\in\{\mathcal{D}\}}$ 表示所有领域的原型，$c_{d^+}$ 对应于（$q_i$，$e_i$）的真实域，$\tau$为温度超参数。该损失促使检索到的实体表示与其对应的领域原型对齐，同时与无关的领域中心保持分离。优化过程仅针对少量采样子集进行，从而确保持续预训练期间的效率。

**基于信息增益的正则化**。虽然 $L_{\text{proto}}$ 鼓励领域间的语义可区分性，但并未明确区分有信息量的语义对齐与虚假相关性或随机扰动。为进一步强调与领域相关的信息，我们引入一种基于信息增益的正则项 [39, 110]。给定预测分布 $\mathbb{P}_q(e)$，信息增益对比（IGC）正则项定义如下:

$$\mathcal{L}_{\text{IGC}} = \text{KL}(\mathbb{P}_q(e) \,\|\, \mathbb{P}_{\text{proto}}(e)) - \text{KL}(\mathbb{P}_q(e) \,\|\, \mathbb{P}_{\text{rand}}(e)), \tag{8}$$

其中，KL(·) 表示 Kullback-Leibler 散度。$\mathbb{P}_{\text{proto}}(e) \propto \exp(\text{sim}(Z_e, c_{d^+}))$ 是一种原型诱导分布，反映了领域一致的语义；而 $\mathbb{P}_{\text{rand}}(e)$ 则是通过打乱域原型构造的随机参考分布。通过明确对比结构化语义对齐与随机性，$L_{\text{IGC}}$ 促使 $\mathbb{P}_q(e)$ 对以域相关语义为基础的实体赋予更高置信度，从而增强跨域一致性。

**整体预训练目标**。完整的预训练损失将检索监督与语义正则化相结合:

$$\mathcal{L}_{\text{pre}} = \underbrace{\alpha_1\mathcal{L}_{\text{BCE}} + (1-\alpha_1)\mathcal{L}_{\text{rank}}}_{\text{Phase I}} + \underbrace{\alpha_2\mathcal{L}_{\text{proto}} + \alpha_3\mathcal{L}_{\text{IGC}}}_{\text{Phase II}}, \tag{9}$$

其中 $\alpha_1$ 是权衡超参数，$\alpha_2$ 和 $\alpha_3$ 是拉格朗日乘数。鉴于目标项的多样性，我们提出一种两阶段预训练策略以增强稳定性。在第一阶段，我们在混合域三元组上的知识图谱补全任务中，使用$L_{\text{BCE}}$ 和$L_{\text{rank}}$ 对 $f_\theta(\cdot)$进行固定轮次数的预训练，以建立其通用检索能力。在第二阶段，我们在一个子集上采用连续预训练机制，使用$L_{\text{proto}}$ 和$L_{\text{IGC}}$ 进行细粒度语义对齐。该分阶段训练方案稳定了优化过程，并避免了由原型级正则化引入的异质损失信号之间的冲突。我们将在附录 B.1 中详细说明。

## 4.2 无标签 IB 优化子图选择器

尽管预训练检索器 $f_\theta(\cdot)$ 能识别相关实体，但直接在整个图上进行推理的计算成本高昂，且结构可解释性有限。我们不返回实体排序列表，而是学习一个查询条件化子图选择器，以识别用于推理的最小且充分的 $G_q$。

**查询条件化子图选择**。为使预训练检索器 $f_\theta^\star(\cdot)$ 高效适配查询特定子图选择，我们采用参数高效微调（PEFT）策略，冻结骨干网络 GFM 参数，并引入轻量级的查询条件化适配机制以进行子图选择。给定知识图谱 G 上的查询$q$，我们生成共享查询提示$p = \text{MLP}(q)$，并将其广播以构造查询条件化实体嵌入:

$$\widetilde{Z}_e = \left[Z_e^\star \,\|\, W_1 p \,\|\, W_2 q\right], \tag{10}$$

其中，$Z_e^\star$是从 $f_\theta^\star(\cdot)$ 初始化后冻结的实体嵌入。$W_1, W_2$ 是可学习权重。$\|$ 表示拼接。

为使微调过程中的子图采样可微且可优化，我们采用 Gumbel-Sigmoid 松弛来近似二元选择门。具体而言，对于每个实体 $e$，选择概率 $S_e(q) \in [0,1]$ 采样如下:

$$S_e(q) = \left[1 + \exp\left(-\sigma\big(\big(\text{MLP}(\widetilde{Z}_e)\big) + g_e\big)/\tau\right)\right]^{-1}, \tag{11}$$

其中，变量 $g_e$ 从标准 Gumbel 分布中采样，$\tau$ 为温度参数。该公式为离散节点选择提供了连续且可微的替代目标，使得在端到端优化框架中能够进行随机但可训练的子图采样。给定阈值$\epsilon$，查询特定子图 $G_q$ 由实体集 $V_q = \{e \in V^T \mid S_e(q) > \epsilon\}$实例化，并据此生成相应的关系集 $R_q$ 和边集 $E_q$。重要的是，Gumbel-Sigmoid 可确保梯度能够通过选择过程传播，从而使子图构建完全可微，并能够针对下游目标进行优化。接下来，我们将讨论其优化过程。

**无标签信息瓶颈**。我们从信息论视角来构建这一选择器。信息瓶颈（IB）原理 [83, 84] 为优化子图选择器提供了理想的方法：在压缩完整知识图谱（最小）的同时，保留查询相关子图（充分）：

$$\mathcal{L}_{\text{IB}} = -I(y; \mathcal{G}_q) + \beta I(\mathcal{G}; \mathcal{G}_q), \tag{12}$$

其中，$I(\cdot)$ 表示互信息，$\beta$ 是用于权衡压缩（第一项）与保留（第二项）的拉格朗日乘子。最小化 $L_{\text{IB}}$ 会促使选择器在 G 中找到一个最小但充分的“核心集”。

然而，由于生成答案 $y$ 不可微，直接优化式（12）难以实现，这也使得寻找 $G_q$面临主要挑战。为解决这一困境，我们通过用查询本身替代标签 $y$，建立了一个用于无标签信息瓶颈优化的替代目标:

$$\mathcal{L}_{\text{LIB}} = -I(q; \mathcal{G}_q) + \beta I(\mathcal{G}; \mathcal{G}_q). \tag{13}$$

接下来我们将证明 $L_{\text{LIB}}$ 是 $L_{\text{IB}}$的忠实代理。

命题 4.2（**L 的误差界$\mathcal{L}_{\text{LIB}}$ 近似**）。设$\mathcal{G}_{\mathbf{q}}$ 为子图，$y$ 为真实答案，$\mathbf{q}$ 为查询。假设马尔可夫性质$\langle y \leftrightarrow \mathbf{q} \leftrightarrow \mathcal{G}_{\mathbf{q}} \rangle$ 存在，则$\mathcal{L}_{\text{IB}}$ 相对于$\mathcal{L}_{\text{LIB}}$ 的近似误差以上界受给定 $y$时$\mathbf{q}$ 的条件熵约束：

$$\left| I(y;\mathcal{G}_{\mathbf{q}}) - I(\mathbf{q};\mathcal{G}_{\mathbf{q}}) \right| \leqslant H(\mathbf{q} \mid y), \tag{14}$$

*where $H(\mathbf{q} \mid y)$ is a data-dependent constant.* [▷Proof]

命题 4.2 指出，近似误差的上界由真实标签 $y$ 在给定查询 $\mathbf{q}$ 时的条件熵确定。一旦数据分布给定并固定，近似差距便可由查询的信息量控制在一个常数界内，从而保证近似质量在理论上受查询信息量的控制。由于公式 (13) 给出的是微调目标的理论形式，其中两项互信息在计算上均不可行。接下来我们将给出它们的可处理推导。

命题 4.3（**下界：** $I(\mathbf{q};\mathcal{G}_{\mathbf{q}})$）。将平均池化子图表示记为 $\mathbf{G}_{\mathbf{q}} = \sum_{e \in \mathcal{V}_{\mathbf{q}}} S_e(\mathbf{q}) \cdot \widetilde{\mathbf{Z}}_e$。由 *Alemi* 等人 *[1],*推导，该下界为：

$$I(\mathbf{q};\mathcal{G}_{\mathbf{q}}) \geqslant -\mathcal{L}_{\text{NCE}} = \log \frac{\exp\left(\text{sim}(\mathbf{G}_{\mathbf{q}}, \mathbf{q})/\tau\right)}{\sum_{\mathbf{q}' \in \mathcal{B}} \exp\left(\text{sim}(\mathbf{G}_{\mathbf{q}'}, \mathbf{q}')/\tau\right)}, \tag{15}$$

*where $(\mathbf{G}_{\mathbf{q}'}, \mathbf{q}')$ is the negative sample pair.* [▷Proof]

命题 4.4（**的上界** $I(\mathcal{G};\ \mathcal{G}_{\mathbf{q}})$）。将子图规模和连通性惩罚项分别记为：

$$\mathcal{L}_{\text{size}} = \sum\nolimits_{e \in \mathcal{V}} S_e(\mathbf{q}), \quad \mathcal{L}_{\text{con}} = \text{Tr}\left(\mathbf{S}_e(\mathbf{q})^{\top} \mathbf{L} \mathbf{S}_e(\mathbf{q})\right), \tag{16}$$

其中 $S_e(\mathbf{q}) \in [0,1]$ 表示实体选择概率，$\mathbf{S}_e(\mathbf{q}) \in \mathbb{R}^{|\mathcal{V}|}$ 表示软选择向量，$\mathbf{L}$ 表示归一化图拉普拉斯矩阵，$\text{Tr}(\cdot)$ 表示计算矩阵的迹。上界为：

$$I(\mathcal{G};\mathcal{G}_{\mathbf{q}}) \leqslant \beta_1 \mathcal{L}_{\text{size}} + \beta_2 \mathcal{L}_{\text{con}}, \tag{17}$$

其中 $\beta_1$ 和 $\beta_2$ 为超参数。 [▷Proof]

命题 4.3 和命题 4.4 提供了一种有理论依据的方法，可通过可处理的优化目标近似 $\mathrm{L}_{\text{LIB}}$；然而，在无标签设置下推导此类界并非易事。通过建立这种联系，我们既确保了理论上的严谨性，也为多领域知识图谱检索中的子图识别提供了实践指导。

**整体微调目标**。在预训练阶段之后，我们在有监督文档检索任务上微调 GFM 检索器，其中每个查询 $\mathbf{q}$ 都与一个带标签的目标实体集 $\mathrm{V}_{\mathbf{q}}^{+}$相关联。尽管标准检索目标旨在根据带标签实体进行准确排序，但并未明确约束检索子图的结构属性。为了解决这一局限，我们将子图级正则化纳入微调目标，得到如下结果：

$$\mathcal{L}_{\text{ftn}} = \alpha_1 \mathcal{L}_{\text{BCE}} + (1-\alpha_1)\mathcal{L}_{\text{rank}} + \lceil \mathcal{L}_{\text{LIB}} \rceil, \tag{18}$$

其中，$\mathrm{L}_{\text{BCE}}$ 和 $\mathrm{L}_{\text{rank}}$ 为实体级相关性和相对排序提供互补监督。附加项$\lceil \mathrm{L}_{\text{LIB}} \rceil$ 充当结构正则项，促使生成依赖于查询的最小且充分的子图。 $\lceil \mathrm{L}_{\text{LIB}} \rceil$ 通过将公式 (15) 和公式 (17) 的右侧代入公式 (13) 构造而成：

$$\lceil \mathcal{L}_{\text{LIB}} \rceil = \mathcal{L}_{\text{NCE}} + \beta_1 \mathcal{L}_{\text{size}} + \beta_2 \mathcal{L}_{\text{con}}. \tag{19}$$

## 4.3 上下文提示器诱导的关系路径

现有方法主要检索相关文档，却将关系隐含处理。为显露 $\mathrm{G}_{\mathbf{q}}$ 中编码的关键推理路径，我们不仅将检索子图视为实体“核心集”，还将其视为用于生成的结构化推理支架。

**实体到文档映射**。 遵循倒排索引范式 [54],，我们构建一个二元映射矩阵 $\mathbf{M} \in \{0,1\}^{|\mathcal{E}_{\mathbf{q}}| \times |\mathcal{C}_{\mathbf{q}}|}$，其中 $\mathrm{C}_{\mathbf{q}}$ 表示支撑文档块集合。矩阵元素 $\mathbf{M}[e,c] = 1$ 表示实体 $e$出现在文档块 $c$中。在推理时，我们不再依赖启发式的 top-$K$ 实体选择，而是将检索范围限制在子图实体 $\mathrm{V}_{\mathbf{q}}$ 中，并为每个实体赋予一个复合重要性权重，定义如下：

$$w_e = S_e(\mathbf{q}) \cdot \frac{1}{\sum_c \mathbf{M}[e,c]} \cdot \left|\phi_{\text{path}}(e)\right|, \tag{20}$$

其中，$S_e(\mathbf{q})$表示式 (11) 中引入的软选择概率，$|\phi_{\text{path}}(e)|$ 表示涉及实体 $e$的最短推理路径数量。该加权方案强调与查询高度相关且在推理子图中具有结构中心性的实体。随后，通过映射矩阵汇总实体权重来计算文档相关性，表示为 $R_c = \mathbf{M}^{\top} w_e$，并据此选取 top-$K$ 个文档。

**推理路径提取**。由于候选路径数量呈指数级增长，在稠密子图中枚举所有可能的路径在计算上不可行。为显式呈现可解释的关系结构，我们使用 $l$ 跳截断深度优先搜索（DFS），从查询特定子图 $\mathrm{G}_{\mathbf{q}}$ 中提取一组紧凑的推理路径 $\Pi_{\mathbf{q}} = \{\pi_1, \ldots\ldots, \pi_l\}$。每条提取出的路径形式为 $\pi = (e_0 \xrightarrow{r_1} e_1 \xrightarrow{r_2} \cdots \xrightarrow{r_k} e_k)$，其中 $k \leqslant l$。为量化每条推理路径与关系感知的相关性，我们定义如下评分函数：

$$S_r(\pi) = \sum_{i=0}^{k} S_{e_i}(\mathbf{q}) + \sum_{j=1}^{k} \omega(r_j), \tag{21}$$

其中，$\omega(r_j)$ 通过关系深度惩罚长程依赖。该评分倾向于选择经由高置信度节点和语义简洁关系的路径。

**上下文提示器构建**。 为避免子图在答案生成过程中退化为文档和实体的无序集合，我们构建了一个上下文提示器，以进行显式的路径感知大语言模型推理：

| 模式 | 值 |
|---|---|
| <系统> | 作为一名高级阅读理解助手，<br>你的任务是仔细分析文本段落及相应问<br>题。你的回答从 ... 之后开始 |
| <文档> | 文档：{title} \n {content} \n |
| <路径> | 推理路径 {path_id}（得分：{path_score}）： \n |
| | 实体 {entity}（类型：{entity_type}）： \n |
| | ---（关系：{relation}）---> \n |
| | 实体 {entity}（类型：{entity_type}）： \n |
| | ---（关系：{relation}）---> \n |
| | 实体 {entity}（类型：{entity_type}）： \n |
| <查询> | 问题：{query} \n |

我们在附录D.3中提供了详细的上下文提示器配置和案例研究。

# 5 实验

本节中，我们进行了大量实验，以评估 Gfm- Retriever[1] 在以下研究问题（研究问题复数）上的表现：

- 研究问题***1***： Gfm-Retriever 能否提升检索性能?
- 研究问题***2***： Gfm-Retriever 在问答任务中的表现如何?
- 研究问题***3***： Gfm-Retriever 能否在不同领域之间实现泛化?
- 研究问题***4***： Gfm-Retriever 是否更省时且更有效?
- 研究问题***5***： 子图如何支持多跳问答推理?
- 研究问题***6***： 各个组件和超参数如何影响 Gfm-Retriever 的整体性能?

## 5.1 实验设置

我们简要介绍数据集、基线和指标的实验设置。详细设置见附录D。

**数据集**。 主要实验基于三**个** 具有代表性的多跳问答基准测试：HotpotQA [100],MuSiQue [85], 和 2WikiMultiHopQA [26],，这些基准广泛用于评估多跳检索和推理能力，涵盖多样的领域和复杂的推理模式。为进一步评估 Gfm-Retriever 在不同领域的泛化能力，我们还纳入**七个** 基准测试，包括 PubMedQA [36], DelucionQA [71], TechQA [8], ExpertQA [56], EManual [61], MS Marco [62], 和 HAGRID [37]。统计数据见表 D.1 和表 D.2。

**基线**。 我们将 Gfm-Retriever 与**18**种最先进的基线方法（及其组合）进行比较，并将其分为**四** 组：

- 基础大语言模型： GPT-4o-mini [31]。
- 单步检索增强生成： 包括 BM25 [70], Contriever [32], GTR [63], ColBERTv2 [72], RAPTOR [73], 和 Proposition [10]。
- 图增强检索增强生成： 包括 GraphRAG [12], G-Retriever[24], LightRAG [17], HippoRAG [18], HippoRAG 2 [19], SubgraphRAG [48], PropRAG [88], 以及与本文最相关的 GFM-RAG [54]。
- 多步检索增强生成： IRCoT [86], FLARE [35],Adaptive-RAG [33]。

值得注意的是，IRCoT [86] 是一种通用的多步推理封装器，可与非迭代检索器结合，使单步检索增强生成和基于图的方法能够通过交错进行检索与生成来完成多跳推理。我们在表 D.1 中对所有基线方法与 Gfm-Retriever 进行了全面比较。

**指标**。 为评估检索质量，我们报告检索实体和文档的 Recall@2 和 Recall@5（$R@2/5_E$、$R@2/5_D$）。此外，我们在部分检索实验中报告平均倒数排名（MRR），以进一步反映排名质量。对于端到端问答，我们在主要实验中采用标准指标，包括精确匹配（EM）、F1 分数、精确率（P）和召回率（R），全面分析答案覆盖率和正确性 [18]。在部分情况下，我们还报告 ROUGE-L，以评估生成的答案与参考答案之间的语义重合度。

**实现细节**。 检索器骨干网络是一个具有 6 层、隐藏维度为 512 的查询条件化图神经网络，使用 all-mpnet-v2 句子编码器进行初始化 [74]。预训练在包含 70 万份文档的大规模语料库上进行，并据此构建了 60 个领域专用知识图谱，其中包含超过 1,400 万个三元组 [54]。模型在 8 个 NVIDIA A100 GPU（80GB）上进行训练，批量大小为 4。其他超参数和实现细节详见附录 E.1。

[1]论文正式发表后，我们将公开代码。

## 5.2 研究问题***1***： 实体检索性能

我们在标准多跳问答基准上评估实体级和文档级检索性能。

**结果**。 如表 1 所示，Gfm-Retriever 在大多数指标上始终优于所有基线，平均排名也最高。它在 $R@2_E$ 和 $R@5_E$ 得分上均排名第一，这验证了子图选择器在提取紧凑且信息丰富的实体集合方面的有效性，也体现了 GFM 检索器所具备的跨域泛化能力。相比之下，Contriever 和 GTR 等稠密检索器忽略了实体间结构，而 GraphRAG 和 G-Retriever 等基于图的方法则倾向于过度选择无关邻居。SubgraphRAG 和 GFM-RAG 利用了图结构，但缺乏查询特定过滤。Gfm-Retriever 通过联合建模语义相关性和结构最小化，改进了上述不足。

## 5.3 研究问题***2***： 问答性能

我们将检索到的证据输入同一个基础大语言模型，以评估端到端多跳问答。

**结果**。 如表2所示，我们的 Gfm-Retriever 已展现出强大的单步问答能力，优于大多数单步基线方法，并超过了多种多步方法。这表明，即使不进行迭代检索，检索最小但充分的子图也能为多跳问答提供信息丰富的证据。当结合 IRCoT 后，Gfm-Retriever 在各数据集和指标上均取得稳定提升，达到最佳总体性能，表明多步推理能够从结构化证据中获益匪浅。Gfm-Retriever 与 Gfm-Retriever+IRCoT 的比较进一步揭示了子图对迭代推理的补充作用：子图提供紧凑的推理支架，而 IRCoT 则逐步优化证据选择，从而在不牺牲正确性（精确率）的情况下，提高答案覆盖率（召回率）。总体而言，实验结果证实，我们的以子图为中心的检索既是一种有效的独立单步问答方法，也为多步推理框架提供了坚实基础。

## 5.4 研究问题***3***： 跨领域泛化能力

我们在**七个** 涵盖生物医学、客户支持和一般知识领域的问答数据集上评估跨域泛化能力，这些数据集均与预训练数据不重叠。我们将 Gfm-Retriever 与三**种** 在研究问题***1*** 和研究问题***2***中确定的强基线进行比较。在不进行领域特定微调的零样本设置下，我们报告所有数据集上的 $R@5_D$ 检索性能，以及三个支持答案生成的数据集上的 ROUGE-L 问答性能。

**结果**。 如图 3 所示，Gfm-Retriever 在检索和问答任务上始终取得最佳跨领域性能。在七个数据集的文档检索任务中，它均优于所有基线方法；在 PubMedQA、DelucionQA 和 HAGRID 等生物医学与客户支持基准测试中，表现尤为突出，显示出其在领域偏移下具备稳健的泛化能力。相比之下，PropRAG 和 HippoRAG 2 的性能明显下降，反映出它们对领域特定索引较为敏感。在生成式问答任务中，Gfm-Retriever 也取得了最高的 ROUGE-L 分数，表明其跨域检索优势能够直接转化为更好的答案生成效果。更多领域特定微调结果见附录 F.1。

**表 1：检索性能结果。我们报告实体级（R@$2_E$、R@$5_E$）和文档级（R@$2_D$、R@$5_D$）在前 2 和前 5 项中的召回率（%）。最佳结果以粗体显示，次优结果带下划线。单元格颜色越深，结果越好。**

| 数据集 | HotpotQA | | | | MuSiQue | | | | 2WikiMultiHopQA | | | | 平均排名 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 方法 | 召回率@$2_E$ | 召回率@$5_E$ | 召回率@$2_D$ | 召回率@$5_D$ | 召回率@$2_E$ | 召回率@$5_E$ | 召回率@$2_D$ | 召回率@$5_D$ | 召回率@$2_E$ | 召回率@$5_E$ | 召回率@$2_D$ | 召回率@$5_D$ | |
| BM25 (▷*SIGIR'94*) | 40.0 | 53.2 | 55.4 | 72.2 | 19.5 | 23.6 | 32.3 | 41.2 | 46.9 | 57.9 | 51.8 | 61.9 | 18.3 |
| Contriever (▷*TMLR'22*) | 41.7 | 56.0 | 57.2 | 75.5 | 19.8 | 28.8 | 34.8 | 46.6 | 39.7 | 51.7 | 46.6 | 57.5 | 17.7 |
| GTR (▷*EMNLP'22*) | 46.5 | 57.1 | 59.4 | 73.3 | 20.3 | 27.7 | 37.4 | 49.1 | 56.7 | 65.7 | 60.2 | 67.9 | 14.1 |
| ColBERTv2 (▷北美计算语言学协会2022年) | 49.3 | 62.3 | 64.7 | 79.3 | 24.4 | 28.1 | 37.9 | 49.2 | 54.7 | 64.7 | 59.2 | 68.2 | 11.6 |
| RAPTOR (▷国际学习表征会议2024年) | 44.6 | 56.7 | 58.1 | 71.2 | 22.3 | 28.6 | 35.7 | 45.3 | 41.0 | 50.0 | 46.3 | 53.8 | 17.2 |
| 命题 (▷*EMNLP2024*年) | 45.5 | 56.4 | 58.7 | 71.1 | 22.2 | 30.1 | 37.6 | 49.3 | 50.7 | 58.7 | 56.4 | 63.1 | 14.6 |
| GraphRAG (▷*arXiv'24*) | 46.6 | 60.6 | 58.3 | 76.6 | 21.1 | 30.1 | 35.4 | 49.3 | 57.0 | 72.0 | 61.6 | 77.3 | 11.7 |
| G-Retriever (▷神经信息处理系统大会2024年) | 42.9 | 52.6 | 53.3 | 65.5 | 21.6 | 28.8 | 38.8 | 45.1 | 54.3 | 62.3 | 60.8 | 67.8 | 15.6 |
| LightRAG (▷*arXiv'24*) | 39.7 | 42.5 | 38.8 | 54.7 | 12.0 | 20.6 | 24.8 | 34.7 | 40.9 | 54.9 | 45.1 | 59.1 | 20.5 |
| HippoRAG (▷神经信息处理系统大会2024年) | 47.3 | 60.8 | 60.1 | 78.5 | 25.7 | 33.5 | 41.2 | 53.2 | 65.1 | 83.1 | 68.4 | 87.0 | 8.6 |
| HippoRAG 2 (▷国际机器学习会议2025年) | 63.4 | 70.0 | 80.5 | 88.1 | 30.3 | 35.9 | 47.0 | 56.7 | 83.6 | 86.5 | 88.9 | 90.1 | 3.1 |
| SubgraphRAG (▷国际学习表征会议2025年) | 48.2 | 56.5 | 61.5 | 73.0 | 27.5 | 32.0 | 42.1 | 49.3 | 64.3 | 80.3 | 70.7 | 85.5 | 9.3 |
| PropRAG (▷*EMNLP2025*年) | 63.4 | 70.7 | 81.9 | 88.0 | 30.9 | 36.9 | 47.7 | 57.9 | 82.0 | 86.0 | 87.9 | 90.1 | 2.8 |
| GFM-RAG (▷神经信息处理系统大会2025年) | 60.7 | 70.9 | 75.6 | 89.6 | 29.8 | 35.0 | 43.5 | 57.6 | 72.6 | 87.6 | 79.1 | 92.4 | 3.7 |
| FLARE (▷*EMNLP2023*年) | 56.4 | 65.0 | 73.1 | 81.3 | 29.6 | 35.9 | 44.3 | 55.1 | 63.0 | 69.0 | 67.1 | 73.1 | 6.8 |
| Adaptive-RAG (▷北美计算语言学协会2024年) | 47.9 | 61.0 | 61.0 | 76.4 | 22.5 | 27.5 | 35.1 | 44.7 | 41.7 | 58.7 | 44.7 | 61.4 | 15.3 |
| BM25 + IRCoT (▷*ACL'23*) | 46.9 | 59.5 | 65.6 | 79.0 | 19.2 | 28.6 | 34.2 | 44.7 | 53.8 | 68.8 | 61.2 | 75.6 | 13.3 |
| Contriever + IRCoT (▷*ACL'23*) | 45.8 | 63.7 | 65.9 | 81.6 | 23.9 | 29.6 | 39.1 | 52.2 | 44.8 | 56.8 | 51.6 | 63.8 | 12.3 |
| ColBERTv2 + IRCoT (▷*ACL'23*) | 52.4 | 66.6 | 67.9 | 82.0 | 24.8 | 32.9 | 41.7 | 53.7 | 56.9 | 67.9 | 64.1 | 74.4 | 8.1 |
| HippoRAG + IRCoT (▷*ACL'23*) | 56.9 | 67.3 | 67.0 | 83.0 | 27.2 | 35.7 | 45.3 | 57.6 | 69.3 | 86.3 | 75.8 | **93.9** | 4.7 |
| **Gfm-Retriever** (我们的) | **65.9** | **72.4** | **83.4** | **90.5** | **33.1** | **37.8** | **51.3** | **59.8** | **85.0** | **88.3** | **91.3** | 93.8 | **1.1** |

**表 2：多跳问答（问答）性能结果。我们报告精确匹配（精确匹配）、F1 分数、精确率（精确率）和召回率（召回率），均以百分比（%）表示。最佳结果以粗体标出，次优结果加下划线。单元格颜色越深，结果越好。**

| 数据集 | HotpotQA | | | | MuSiQue | | | | 2WikiMultiHopQA | | | | 平均排名 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 方法 | EM | F1 | P | R | EM | F1 | P | R | EM | F1 | P | R | |
| GPT-4o-mini (▷*arXiv2024*年) | 40.0 | 53.2 | 55.4 | 72.2 | 19.5 | 23.6 | 32.3 | 41.2 | 46.9 | 57.9 | 51.8 | 61.9 | 19.3 |
| Contriever (▷*TMLR'22*) | 34.9 | 51.2 | 51.8 | 52.2 | 16.3 | 21.6 | 21.9 | 22.0 | 24.3 | 33.9 | 34.1 | 34.8 | 18.5 |
| GTR (▷*EMNLP'22*) | 33.8 | 51.9 | 52.5 | 52.9 | 15.1 | 25.2 | 25.5 | 25.7 | 33.7 | 42.5 | 42.8 | 43.6 | 16.9 |
| ColBERTv2 (▷北美计算语言学协会'22) | 43.4 | 57.7 | 58.4 | 58.8 | 15.5 | 26.4 | 26.7 | 27.0 | 33.4 | 43.3 | 43.6 | 44.3 | 14.8 |
| RAPTOR (▷国际学习表征会议'24) | 48.2 | 59.2 | 59.9 | 60.3 | 17.6 | 28.9 | 29.2 | 29.5 | 30.6 | 42.0 | 42.3 | 43.1 | 13.3 |
| GraphRAG (▷*arXiv'24*) | 35.3 | 54.6 | 55.3 | 55.6 | 13.4 | 29.5 | 29.8 | 30.1 | 28.3 | 46.9 | 47.2 | 48.1 | 14.3 |
| G-Retriever (▷神经信息处理系统大会'24) | 33.2 | 50.3 | 50.6 | 51.5 | 18.0 | 25.9 | 26.1 | 26.6 | 42.3 | 45.6 | 45.9 | 46.8 | 15.7 |
| LightRAG (▷*arXiv'24*) | 36.8 | 48.3 | 48.9 | 49.2 | 18.1 | 27.5 | 27.8 | 28.1 | 45.1 | 49.5 | 49.8 | 50.8 | 14.4 |
| HippoRAG (▷神经信息处理系统大会'24) | 41.8 | 55.0 | 55.7 | 56.0 | 19.2 | 29.8 | 30.1 | 30.5 | 46.6 | 59.5 | 60.0 | 61.1 | 11.6 |
| HippoRAG 2 (▷国际机器学习会议'25) | 57.3 | 69.6 | 70.1 | 71.5 | 33.9 | 40.5 | 40.8 | 41.6 | 75.4 | 78.2 | 78.7 | 80.3 | 3.8 |
| SubgraphRAG (▷国际学习表征会议'25) | 52.6 | 65.8 | 66.2 | 67.5 | 33.0 | 39.2 | 39.5 | 40.2 | 71.0 | 76.2 | 76.7 | 78.3 | 6.7 |
| PropRAG (▷*EMNLP2025*年) | 57.2 | 70.1 | 69.6 | 71.0 | 38.9 | 41.3 | 40.6 | 40.4 | 73.0 | 79.9 | 78.4 | 79.8 | 4.0 |
| GFM-RAG (▷神经信息处理系统大会2025年) | 51.6 | 66.9 | 67.7 | 68.2 | 30.2 | 40.4 | 40.8 | 41.3 | 69.8 | 77.7 | 78.2 | 79.8 | 5.8 |
| FLARE (▷*EMNLP 2023*年) | 48.7 | 60.6 | 61.3 | 61.8 | 16.2 | 28.4 | 28.7 | 29.1 | 46.7 | 65.4 | 65.8 | 67.2 | 10.0 |
| Adaptive-RAG (▷北美计算语言学协会2024年) | 45.5 | 59.6 | 60.3 | 60.7 | 13.8 | 25.6 | 25.9 | 26.1 | 48.9 | 62.8 | 63.2 | 64.5 | 11.8 |
| ColBERTv2 + IRCoT (▷*ACL'23*) | 45.5 | 58.4 | 59.1 | 59.5 | 19.1 | 30.5 | 30.8 | 31.2 | 35.4 | 45.1 | 45.4 | 46.3 | 11.9 |
| HippoRAG + IRCoT (▷*ACL'23*) | 45.7 | 59.2 | 59.9 | 60.3 | 21.9 | 33.3 | 33.7 | 34.0 | 47.7 | 62.7 | 63.1 | 64.3 | 9.3 |
| GFM-RAG + IRCoT (▷*ACL'23*) | 56.0 | 71.8 | 70.7 | 70.2 | 36.6 | **49.2** | 47.7 | 47.3 | 72.5 | **80.8** | 77.1 | 78.0 | 3.4 |
| **Gfm-Retriever** （我们的方法） | 55.4 | 70.4 | 72.5 | 72.5 | 34.8 | 41.5 | 42.3 | 42.5 | 74.3 | 78.8 | 79.1 | 80.4 | 3.0 |
| **Gfm-Retriever** （我们的方法） + IRCoT （▷*ACL'23*） | **60.5** | **72.9** | **75.5** | **74.5** | **42.8** | 48.7 | **48.9** | **50.4** | **78.8** | 79.7 | **79.4** | **82.3** | **1.3** |

## 5.5 研究问题*4*： 效率与效果分析

我们比较了各代表性基线方法的检索延迟和文档级召回率（R@$5_D$）。

**结果。** 图4展示了不同检索范式之间明显的效率—效果权衡。单步稠密检索器（例如 ColBERTv2）能够实现较低的延迟，但召回率有限；而 HippoRAG+IRCoT 等多步方法则提升了效果但代价是显著的延迟。相比之下，Gfm-Retriever 在所有数据集上始终位于左上区域，以亚秒级检索时间实现了最高或接近最高的 R@$5_D$。与 SubgraphRAG 和 GFM-RAG 等基于图的基线相比，它无需迭代图遍历，便能提供更高的检索质量；这得益于信息瓶颈优化的子图选择器，它能通过单次前向传播提取紧凑的查询条件化子图。

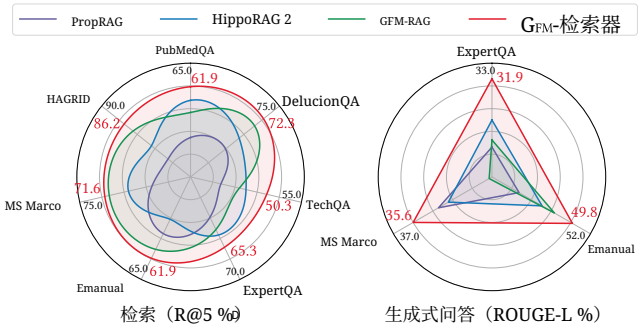


图3：跨域泛化能力分析。

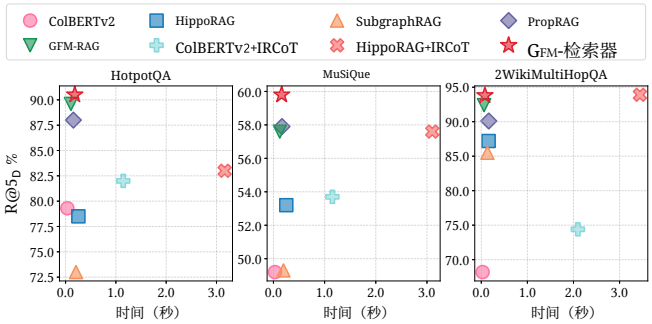


图4：检索时间效率与效果。

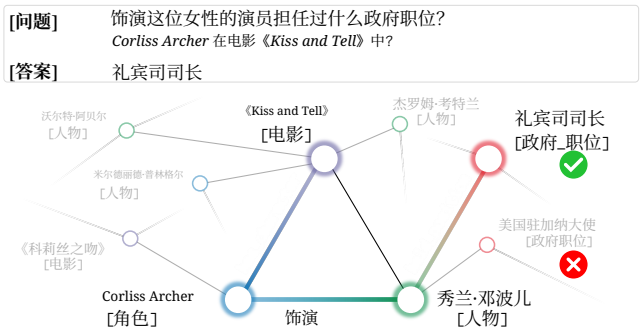


图5：检索子图的可视化。

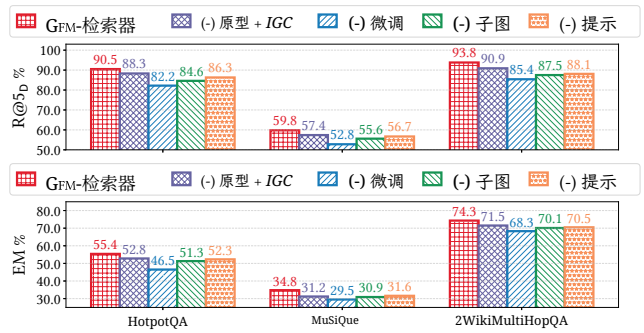


图6：检索与多跳问答的消融实验。

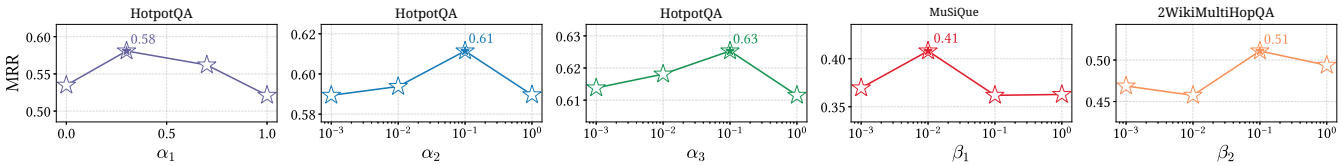


图7：超参数敏感性分析。结果以 MRR 指标报告

ic.

## 5.6 研究问题*5*： 子图可视化与案例研究

图5展示了一个来自 HotpotQA 的典型多跳问答示例。Gfm-Retriever 检索到一个紧凑的子图，其中形成了一条清晰的推理链，从“Corliss Archer”到《Kiss and Tell》、秀兰·邓波儿，再到正确的政府职务“原型委员会主席”。重要的是，与目标推理路径语义相关但无关的实体被排除在外。这个案例表明，Gfm-Retriever 能识别出最小化且与查询对齐的子图，其显式关系路径直接支持多跳推理，而非聚合松散相关的证据。

## 5.7 研究问题*6*： 消融与敏感性分析

**消融**。 我们对 Gfm-Retriever 的**五个** 核心模块进行消融：原型对齐（原型，公式（7））、信息增益正则化器（*IGC*，公式（8））、微调（微调，公式（18））、子图选择器（子图，第 4.2 节），以及上下文内关系提示策略（提示，第 4.3 节）。如图 6 所示，移除任何一个组件都会导致检索和问答性能持续下降。具体而言，移除原型+*IGC* 会导致 $R@5_D$ 明显下降，使模型退化为 GFM-RAG 中的检索器，并牺牲跨域对齐。禁用微调 会导致性能下降，尤其是问答性能，这凸显了适应查询条件化子图选择的重要性。

移除子图 也会导致性能大幅下降，这相当于用通用检索策略取代学习得到的子图。移除提示 则会进一步降低问答性能，而检索受到的影响较小，这证实了将关系路径明确组织为上下文内提示，对于将结构性证据转化为有效的多跳推理至关重要。

**敏感性**。 我们研究了**五个** 加权超参数：$\alpha_1$、$\alpha_2$和 $\alpha_3$（见公式 (9)），以及 $\beta_1$和 $\beta_2$（见公式 (17)）。图 7 报告了以 MRR 衡量的结果。总体而言，性能对超参数的不同取值较为敏感，且存在一个合理的取值范围。

## 6 结论

在这项工作中，我们从结构视角重新审视基于图的检索，并提出 Gfm-Retriever，直接学习查询特定的子图作为推理证据。通过将预训练的 GFM 改造为跨域检索器，并引入无标签的信息瓶颈来优化最小但充分的子图，我们的框架实现了高效且可解释的多跳推理。大量实验表明，该方法在检索质量、问答性能、泛化能力和效率方面表现出色。

# 参考答案

[1] Alexander A. Alemi、Ian Fischer、Joshua V. Dillon 和 Kevin Murphy。2017年。深度变分信息瓶颈。载于《国际学习表征会议》。[2] Emily Alsentzer、Samuel Finlayson、Michelle Li 和 Marinka Zitnik。2020年。子图神经网络。载于《神经信息处理系统进展》第33卷（2020年），8017–8029。[3] Aijun Bai、Rolf Jagerman、Zhen Qin、Le Yan、Pratyush Kar、Bing-Rong Lin、Xu-anhui Wang、Michael Bendersky 和 Marc Najork。2023年。用于二元相关性校准排序的回归兼容列表式目标。载于《第32届ACM国际信息与知识管理会议论文集》。4502–4508。[4] Pablo Barceló、Egor V Kostylev、Mikael Monet、Jorge Pérez、Juan Reutter 和 Juan Pablo Silva。2020年。图神经网络的逻辑表达能力。载于《国际学习表征会议》。[5] Sebastian Borgeaud、Arthur Mensch、Jordan Hoffmann、Trevor Cai、Eliza Ruther-ford、Katie Millican、George Bm Van Den Driessche、Jean-Baptiste Lespiau、Bog-dan Damoc、Aidan Clark 等。2022年。通过从数万亿个词元中检索来改进语言模型。载于《国际机器学习会议》。PMLR，2206–2240。[6] Karsten M Borgwardt 和 Hans-Peter Kriegel。2005年。图上的最短路径核。载于《第五届IEEE国际数据挖掘会议论文集》。IEEE，74–81。[7] Yuxuan Cao、Jiarong Xu、Carl Yang、Jiaan Wang、Yunchao Zhang、Chunping Wang、Lei Chen 和 Yang Yang。2023年。何时预训练图神经网络？从数据生成的视角出发！载于《第29届ACM SIGKDD知识发现与数据挖掘会议论文集》。142–153。[8] Vittorio Castelli、Rishav Chakravarti、Saswati Dana、Anthony Ferritto、Radu Florian、Martin Franz、Dinesh Garg、Dinesh Khandelwal、J Scott McCarley、Michael McCawley 等。2020年。TechQA数据集。载于《第58届计算语言学协会年会论文集》。1269–1278。[9] Ke-Jia Chen、Jiajun Zhang、Linpu Jiang、Yunyun Wang 和 Yuxuan Dai。2022年。动态图神经网络上的预训练。《神经计算》500（2022年），679–687。[10] Tong Chen、Hongwei Wang、Sihao Chen、Wenhao Yu、Kaixin Ma、Xinran Zhao、Hongming Zhang 和 Dong Yu。2024年。Dense X检索：应采用何种检索粒度？载于《2024年自然语言处理实证方法会议论文集》。15159–15177。[11] Maarten De Rijke。2000年。关于分级模态逻辑的说明。《Studia Logica》64卷第2期（2000年），271–283。[12] Darren Edge、Ha Trinh、Newman Cheng、Joshua Bradley、Alex Chao、Apurva Mody、Steven Truitt、Dasha Metropolitansky、Robert Osazuwa Ness 和 Jonathan Larson。2024年。从局部到全局：一种面向查询摘要的图检索增强生成方法。arXiv预印本 arXiv:2404.16130（2024年）。[13] Wenqi Fan、Yujuan Ding、Liangbo Ning、Shijie Wang、Hengyun Li、Dawei Yin、Tat-Seng Chua 和 Qing Li。2024年。RAG与大语言模型结合综述：迈向检索增强大语言模型。载于《第30届ACM SIGKDD知识发现与数据挖掘会议论文集》。6491–6501。[14] Jie Gao、Jia Hu、Geyong Min 和 Fei Hao。2025年。XDGNN：通过解释引导的子图扩展实现高效的分布式图神经网络训练。《IEEE并行与分布式系统汇刊》（2025年）。[15] Thomas Gärtner、Peter Flach 和 Stefan Wrobel。2003年。关于图核：困难性结果与高效替代方案。载于《学习理论与核机器：第16届学习理论年会暨第7届核研讨会》。Springer，129–143。[16] Abdellah Ghassel、Ian Robinson、Gabriel Tanase、Hal Cooper、Bryan Thomp-son、Zhen Han、Vassilis Ioannidis、Soji Adeshina 和 Huzefa Rangwala。2025年。用于增强多跳检索的层次词汇图。载于《第31届ACM SIGKDD知识发现与数据挖掘会议论文集》。4457–4466。[17] Zirui Guo、Lianghao Xia、Yanhua Yu、Tu Ao 和 Chao Huang。2024年。LightRAG：简单快速的检索增强生成。arXiv预印本 arXiv:2410.05779（2024年）。[18] Bernal Jiménez Gutiérrez、Yiheng Shu、Yu Gu、Michihiro Yasunaga 和 Yu Su。2024年。HippoRAG：面向大语言模型、受神经生物学启发的长期记忆。载于《神经信息处理系统进展》第37卷（2024年），59532–59569。[19] Bernal Jiménez Gutiérrez、Yiheng Shu、Weijian Qi、Sizhe Zhou 和 Yu Su。2025年。从RAG到记忆：面向大语言模型的非参数持续学习。载于《第四十二届国际机器学习会议》。[20] Kelvin Guu、Kenton Lee、Zora Tung、Panupong Pasupat 和 Mingwei Chang。2020年。检索增强语言模型预训练。载于《国际机器学习会议》。PMLR，3929–3938。[21] Haoyu Han、Yu Wang、Harry Shomer、Kai Guo、Jiayuan Ding、Yongjia Lei、Mahantesh Halappanavar、Ryan A Rossi、Subhabrata Mukherjee、Xianfeng Tang 等。2024年。基于图的检索增强生成（GraphRAG）。arXiv预印本 arXiv:2501.00309（2024年）。[22] Haoyu Han, Yaochen Xie, Hui Liu, Xianfeng Tang, Sreyashi Nag, William Headden, Yang Li, Chen Luo, Shuiwang Ji, 等。2025年。图推理：构建隐式知识以增强大语言模型推理能力。载于计算语言学协会论文集。[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie 和 Ross Girshick。2020年。用于无监督视觉表示学习的动量对比。载于IEEE/CVF计算机视觉与模式识别会议论文集。9729–9738。[24] Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh Chawla, Thomas Laurent, Yann Le-Cun, Xavier Bresson 和 Bryan Hooi。2024年。G-Retriever：用于文本图理解和问答的检索增强生成。神经信息处理系统进展37（2024年），132876–132907。[25] Yufei He 和 Bryan Hooi。2024年。UniGraph：从自然语言中学习跨域图基础模型。arXiv预印本arXiv:2402.13630（2024年）。[26] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara 和 Akiko Aizawa。2020年。构建多跳问答数据集，以全面评估推理步骤。载于第28届国际计算语言学会议论文集。6609–6625。[27] Ziniu Hu, Yuxiao Dong, Kuansan Wang, Kai-Wei Chang 和 Yizhou Sun。2020年。GPT-GNN：图神经网络的生成式预训练。载于第26届ACM SIGKDD知识发现与数据挖掘国际会议论文集。1857–1867。[28] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin 等。2025年。大语言模型幻觉综述：原理、分类体系、挑战与开放问题。ACM信息系统汇刊43，2（2025年），1–55。[29] Qian Huang, Hongyu Ren 和 Jure Leskovec。2022年。通过连接子图预训练进行少样本关系推理。神经信息处理系统进展35（2022年），6397–6409。[30] Xingyue Huang, Miguel Romero, Ismail Ceylan 和 Pablo Barceló。2023年。知识图谱上基于关系Weisfeiler-Leman的链接预测理论。神经信息处理系统进展36（2023年），19714–19748。[31] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford 等。2024年。GPT-4o系统卡片。arXiv预印本arXiv:2410.21276（2024年）。[32] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin 和 Edouard Grave。2022年。通过对比学习进行无监督稠密信息检索。《机器学习研究汇刊》（2022年）。[33] Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang 和 Jong C Park。2024年。Adaptive-RAG：通过问题复杂度学习如何自适应检索增强大语言模型。载于2024年北美计算语言学协会分会年会：人类语言技术论文集。7029–7043。[34] Xunqiang Jiang, Yuanfu Lu, Yuan Fang 和 Chuan Shi。2021年。异构图上的图神经网络对比式预训练。载于第30届ACM国际信息与知识管理会议论文集。803–812。[35] Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan 和 Graham Neubig。2023年。主动检索增强生成。载于2023年自然语言处理实证方法会议论文集。7969–7992。[36] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen 和 Xinghua Lu。2019年。PubMedQA：生物医学研究问答数据集。载于2019年自然语言处理实证方法会议暨第9届自然语言处理国际联合会议（EMNLP-IJCNLP）论文集。2567–2577。[37] Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur 和 Jimmy Lin。2023年。HAGRID：用于生成式信息寻求并带有归因的人类与大语言模型协作数据集。arXiv预印本arXiv:2307.16883（2023年）。[38] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen 和 Wen-tau Yih。2020年。面向开放域问答的稠密段落检索。载于2020年自然语言处理实证方法会议论文集。6769–6781。[39] John T Kent。1983年。信息增益与相关性的一般度量。《Biometrika》70，1（1983年），163–173。[40] Thomas N Kipf 和 Max Welling。2017年。利用图卷积网络进行半监督分类。载于国际学习表征会议。[41] Nils M Kriege, Fredrik D Johansson 和 Christopher Morris。2020年。图核综述。《Applied Network Science》5，1（2020年），6。[42] John Boaz Lee, Ryan A Rossi, Xiangnan Kong, Sungchul Kim, Eunyee Koh 和 Anup Rao。2019年。具有基于模体的注意力的图卷积网络。载于第28届ACM国际信息与知识管理会议论文集。499–508。[43] O-Joun Lee 等。2025年。通过子图条件图信息瓶颈对分子图神经网络进行预训练。载于AAAI人工智能会议论文集，第39卷。17204–17213。

[44] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel 等。2020年。面向知识密集型自然语言处理任务的检索增强生成。神经信息处理系统进展33（2020年），9459–9474。[45] Dawei Li, Shu Yang, Zhen Tan, Jae Baik, Sukwon Yun, Joseph Lee, Aaron Chacko, Bojian Hou, Duy Duong-Tran, Ying Ding 等。2024年。DALK：通过科学文献动态协同增强大语言模型和知识图谱，以回答阿尔茨海默病相关问题。载于计算语言学协会论文集。2187–2205。[46] Haoyang Li, Xin Wang, Xueling Zhu, Weigao Wen 和 Wenwu Zhu。2025年。在分布偏移下通过方差对比估计解耦不变子图。载于国际机器学习会议。[47] Jianxin Li, Qingyun Sun, Hao Peng, Beining Yang, Jia Wu 和 Philip S Yu。2023年。通过强化关键结构挖掘实现自适应子图神经网络。《IEEE模式分析与机器智能汇刊》45，7（2023年），8063–8080。[48] Mufei Li, Siqi Miao 和 Pan Li。2025年。简单即有效：图和大语言模型在基于知识图谱的检索增强生成中的作用。载于国际学习表征会议。[49] Fangru Lin, Emanuele La Malfa, Valentin Hofmann, Elle Michelle Yang, Anthony G Cohn 和 Janet B Pierrehumbert。2024年。异步计划推理中的图增强大语言模型。载于第四十一届国际机器学习大会。[50] Jingzhe Liu, Haitao Mao, Zhikai Chen, Wenqi Fan, Mingxuan Ju, Tong Zhao, Neil Shah 和 Jiliang Tang。2024年。一图一模型：跨域图预训练的新视角。arXiv预印本arXiv:2412.00315（2024年）。[51] Jiawei Liu, Cheng Yang, Zhiyuan Lu, Junze Chen, Yibo Li, Mengmei Zhang, Ting Bai, Yuan Fang, Lichao Sun, Philip S Yu 等。2025年。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年）。[52] Yixin Liu, Ming Jin, Shirui Pan, Chuan Zhou, Yu Zheng, Feng Xia 和 S Yu Philip。2022年。图自监督学习：综述。《IEEE知识与数据工程汇刊》35，6（2022年），5879–5900。[53] Yuanfu Lu, Xunqiang Jiang, Yuan Fang 和 Chuan Shi。2021年。学习预训练图神经网络。AAAI人工智能会议论文集35，5（2021年），4276–4284。[54] Linhao Luo, Zicheng Zhao, Gholamreza Haffari, Dinh Phung, Chen Gong 和 Shirui Pan。2025年。GFM-RAG：用于检索增强生成的图基础模型。载于神经信息处理系统进展。[55] Yao Ma 和 Jiliang Tang。2021年。《图上的深度学习》。剑桥大学出版社。[56] Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar 和 Dan Roth。2024年。ExpertQA：专家策划的问题与带属性的答案。载于2024年北美计算语言学协会分会年会：人类语言技术论文集。3025–3045。[57] Costas Mavromatis 和 George Karypis。2025年。GNN-RAG：通过图神经网络检索实现知识图谱上的高效大语言模型推理。载于计算语言学协会论文集。16682–16699。[58] Fatma Miladi, Valéry Psyché 和 Daniel Lemire。2024年。利用GPT-4提升教育准确性：针对大规模开放在线课程中检索增强生成的比较研究。载于国际人工智能教育会议。Springer，427–434。[59] Federico Monti, Karl Otness 和 Michael M Bronstein。2018年。MotifNet：用于有向图的基于模体的图卷积网络。载于IEEE数据科学研讨会。IEEE，225–228。[60] Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan 和 Martin Grohe。2019年。Weisfeiler和Leman走向神经网络：高阶图神经网络。载于AAAI人工智能会议论文集，第33卷。4602–4609。[61] Abhilash Nandy, Soumya Sharma, Shubham Maddhashiya, Kapil Sachdeva, Pawan Goyal 和 Niloy Ganguly。2021年。电子设备问答：新的基准数据集和基于多任务学习的问答框架。载于计算语言学协会论文集：2021年EMNLP。4600–4609。[62] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder 和 Li Deng。2016年。MS Marco：由人工编写的机器阅读理解数据集。载于认知计算研讨会论文集：整合神经与符号方法（与第30届神经信息处理系统年会同期举行，2016年）。[63] Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang 等。2022年。大型双编码器是可泛化检索器。载于2022年自然语言处理实证方法会议论文集。9844–9855。[64] Aaron van den Oord, Yazhe Li 和 Oriol Vinyals。2018年。利用对比预测编码进行表征学习。arXiv预印本arXiv:1807.03748（2018年）。[65] Martin Otto。2019。分级模态逻辑与计数双模拟。arXiv预印本 arXiv:1910.00039（2019）。[66] Boci Peng、Yun Zhu、Yongchao Liu、Xiaohe Bo、Haizhou Shi、Chuntao Hong、Yan Zhang 和 Siliang Tang。2024。图检索增强生成：综述。ACM信息系统汇刊（2024）。[67] Hao Peng、Jianxin Li、Qiran Gong、Yuanxin Ning、Senzhang Wang 和 Lifang He。2020。基于模体匹配的图分类子图级注意力卷积网络。载于《AAAI人工智能会议论文集》第34卷。5387–5394。[68] Haiquan Qiu、Yongqi Zhang、Yong Li 等。2024。理解图神经网络在规则学习中的表达能力。载于《国际学习表征会议》。[69] Jiezhong Qiu、Qibin Chen、Yuxiao Dong、Jing Zhang、Hongxia Yang、Ming Ding、Kuansan Wang 和 Jie Tang。2020。GCC：用于图神经网络预训练的图对比编码。载于《第26届ACM SIGKDD知识发现与数据挖掘国际会议论文集》。1150–1160。[70] Stephen E Robertson 和 Steve Walker。1994。对概率加权检索二泊松模型的一些简单有效近似。载于《第十七届ACM-SIGIR信息检索研究与发展国际大会论文集》。Springer，232–241。[71] Mobashir Sadat、Zhengyu Zhou、Lukas Lange、Jun Araki、Arsalan Gundroo、Bingqing Wang、Rakesh R Menon、Md Parvez 和 Zhe Feng。2023。DelucionQA：检测特定领域问答中的幻觉。载于《计算语言学协会论文集：EMNLP 2023》。822–835。[72] Keshav Santhanam、Omar Khattab、Jon Saad-Falcon、Christopher Potts 和 Matei Zaharia。2022。ColBERTv2：通过轻量级后期交互实现高效检索。载于《2022年计算语言学协会北美分会会议：人类语言技术论文集》。3715–3734。[73] Parth Sarthi、Salman Abdullah、Aditi Tuli、Shubh Khanna、Anna Goldie 和 Christopher D Manning。2024。RAPTOR：用于树组织检索的递归式抽象处理。载于《国际学习表征会议》。[74] SBERT。2021。all-mpnet-base-v2。https://huggingface.co/sentence-transformers/all-mpnet-base-v2[75] Nino Shervashidze、Pascal Schweitzer、Erik Jan Van Leeuwen、Kurt Mehlhorn 和 Karsten M Borgwardt。2011。Weisfeiler–Lehman图核。《机器学习研究杂志》12卷9期（2011）。[76] Yongduo Sui、Xiang Wang、Jiancan Wu、Min Lin、Xiangnan He 和 Tat-Seng Chua。2022。用于可解释且可泛化图分类的因果注意力。载于《第28届ACM SIGKDD知识发现与数据挖掘会议论文集》。1696–1705。[77] Jiashuo Sun、Chengjin Xu、Lumingyuan Tang、Saizhuo Wang、Chen Lin、Yeyun Gong、Lionel Ni、Heung-Yeung Shum 和 Jian Guo。2024。Think-on-Graph：大语言模型在知识图谱上的深度负责任推理。载于《国际学习表征会议》。[78] Mingchen Sun、Kaixiong Zhou、Xin He、Ying Wang 和 Xin Wang。2022。GPPT：图预训练与提示调优，助力图神经网络实现泛化。载于《第28届ACM SIGKDD知识发现与数据挖掘会议论文集》。1717–1727。[79] Qingyun Sun、Jianxin Li、Hao Peng、Jia Wu、Yuanxing Ning、Philip S Yu 和 Lifang He。2021。SUGAR：采用强化池化和自监督互信息机制的子图神经网络。载于《ACM网络会议论文集》。2081–2091。[80] Yanchao Tan、Zihao Zhou、Hang Lv、Weiming Liu 和 Carl Yang。2024。WalkLM：用于属性图嵌入的统一语言模型微调框架。《神经信息处理系统进展》36卷（2024）。[81] Jiabin Tang、Yuhao Yang、Wei Wei、Lei Shi、Long Xia、Dawei Yin 和 Chao Huang。2024。HiGPT：异构图语言模型。载于《第30届ACM SIGKDD知识发现与数据挖掘会议论文集》。2842–2853。[82] Wenzhuo Tang、Haitao Mao、Danial Dervovic、Ivan Brugere、Saumitra Mishra、Yuying Xie 和 Jiliang Tang。2024。跨域图数据扩展：扩散模型的一个示例。arXiv预印本 arXiv:2406.01899（2024）。[83] Naftali Tishby、Fernando C Pereira 和 William Bialek。2000。信息瓶颈方法。arXiv预印本 physics/0004057（2000）。[84] Naftali Tishby 和 Noga Zaslavsky。2015。深度学习与信息瓶颈原理。载于《IEEE信息论研讨会》。IEEE，1–5。[85] Harsh Trivedi、Niranjan Balasubramanian、Tushar Khot 和 Ashish Sabharwal。2022。MuSiQue：通过单跳问题组合构造多跳问题。《计算语言学协会会刊》10卷（2022），539–554。[86] Harsh Trivedi、Niranjan Balasubramanian、Tushar Khot 和 Ashish Sabharwal。2023。将检索与思维链推理交替用于知识密集型多步问题。载于《计算语言学协会第61届年会论文集》。10014–10037。[87] Heng Wang、Shangbin Feng、Tianxing He、Zhaoxuan Tan、Xiaochuang Han 和 Yulia Tsvetkov。2024。语言模型能否用自然语言解决图问题？《神经信息处理系统进展》36卷（2024）。[88] Jingjin Wang 和 Jiawei Han。2025。PropRAG：通过在命题路径上进行束搜索来引导检索。载于《2025年自然语言处理实证方法会议论文集》。

自然语言处理方法。6223–6238。[89] Shuting Wang、Jiejun Tan、Zhicheng Dou 和 Ji-Rong Wen。2025。OmniEval：面向金融领域的全方位自动化检索增强生成评估基准。载于《2025年自然语言处理实证方法会议论文集》。5737–5762。[90] Yu Wang、Nedim Lipka、Ryan A Rossi、Alexa Siu、Ruiyi Zhang 和 Tyler Derr。2024。用于多文档问答的知识图谱提示。载于《AAAI人工智能会议论文集》第38卷。19206–19214。[91] Nirmalie Wiratunga、Ramitha Abeyratne、Lasal Jayawardena、Kyle Martin、Stewart Massie、Ikechukwu Nkisi-Orji、Ruvan Weerasinghe、Anne Liret 和 Bruno Fleisch。2024。CBR-RAG：基于案例的推理在大语言模型法律问答检索增强生成中的应用。载于《基于案例推理国际会议》。Springer，445–460。[92] Yaozu Wu、Yankai Chen、Zhishuai Yin、Weiping Ding 和 Irwin King。2023。生物医学数据图嵌入技术综述：方法与应用。《信息融合》100卷（2023），101909。[93] Yaochen Xie、Zhao Xu、Jingtun Zhang、Zhengyang Wang 和 Shuiwang Ji。2022。图神经网络的自监督学习：统一综述。《IEEE模式分析与机器智能汇刊》45卷2期（2022），2412–2429。[94] Jiarong Xu、Renhong Huang、Xin Jiang、Yuxuan Cao、Carl Yang、Chunping Wang 和 Yang Yang。2023。少即是多：从数据主动视角看图神经网络预训练。《神经信息处理系统进展》36卷（2023），56946–56978。[95] Ran Xu、Wenqi Shi、Yue Yu、Yuchen Zhuang、Bowen Jin、May Dongmei Wang、Joyce Ho 和 Carl Yang。2024。RAM-EHR：检索增强与电子健康记录临床预测相结合。载于《计算语言学协会第62届年会论文集（第2卷：短篇论文）》。754–765。[96] Shicheng Xu、Liang Pang、Huawei Shen 和 Xueqi Cheng。2024。检索增强生成中词元级协调的理论。载于《国际学习表征会议》。[97] Qi Xuan、Jinhuan Wang、Minghao Zhao、Junkun Yuan、Chenbo Fu、Zhongyuan Ruan 和 Guanrong Chen。2019。子图网络及其在结构特征空间扩展中的应用。《IEEE知识与数据工程汇刊》33卷6期（2019），2776–2789。[98] Carl Yang、Mengxiong Liu、Vincent W Zheng 和 Jiawei Han。2018年。节点、模体与子图：通过结构卷积利用网络功能块。载于《IEEE/ACM社交网络分析与挖掘进展国际会议》。IEEE，47–52。[99] Yaming Yang、Ziyu Guan、Zhe Wang、Wei Zhao、Cai Xu、Weigang Lu 和 Jianbin Huang。2022。基于结构聚类的异构图自监督预训练。《神经信息处理系统进展》35卷（2022），16962–16974。[100] Zhilin Yang、Peng Qi、Saizheng Zhang、Yoshua Bengio、William Cohen、Ruslan Salakhutdinov 和 Christopher D Manning。2018年。HotpotQA：用于多样化、可解释多跳问答的数据集。载于《2018年自然语言处理实证方法会议论文集》。2369–2380。[101] Michihiro Yasunaga、Hongyu Ren、Antoine Bosselut、Percy Liang 和 Jure Leskovec。2021。QA-GNN：利用语言模型和知识图谱进行问答推理。载于《2021年计算语言学协会北美分会会议：人类语言技术论文集》。535–546。[102] Zixuan Yi、Iadh Ounis 和 Craig Macdonald。2023。用于跨域推荐的对比式图提示调优。《ACM信息系统汇刊》42卷2期（2023），1–28。[103] Jun Yin、Chaozhuo Li、Hao Yan、Jianxun Lian 和 Senzhang Wang。2023。一次训练，处处解释：可解释图神经网络预训练。《神经信息处理系统进展》36卷（2023），35277–35299。[104] Junchi Yu、Tingyang Xu、Yu Rong、Yatao Bian、Junzhou Huang 和 Ran He。2021。用于子图识别的图信息瓶颈。载于《国际学习表征会议》。[105] Junchi Yu、Tingyang Xu、Yu Rong、Yatao Bian、Junzhou Huang 和 Ran He。2021。利用子图信息瓶颈识别预测性子结构。《IEEE模式分析与机器智能汇刊》46卷3期（2021），1650–1663。[106] Xingtong Yu、Chang Zhou、Yuan Fang 和 Xinming Zhang。2024。无文本多域图预训练：迈向图基础模型。arXiv预印本 arXiv:2405.13934（2024）。[107] Jintao Zhang、Guoliang Li 和 Jinyang Su。2025。SAGE：面向检索增强生成的精确检索框架。载于《国际数据工程会议》。[108] Minjia Zhang、Wenhan Wang 和 Yuxiong He。2022。Grasp：通过子图采样与剪枝优化基于图的最近邻搜索。载于《第15届ACM网络搜索与数据挖掘国际会议论文集》。1395–1405。[109] Qinggang Zhang、Shengyuan Chen、Yuanchen Bei、Zheng Yuan、Huachi Zhou、Zijin Hong、Hao Chen、Yilin Xiao、Chuang Zhou、Junnan Dong 等。2025。A面向定制化大语言模型的图检索增强生成综述。*arXiv*预印本 *arXiv:2501.13958*（2025年）。[110] Wentao Zhang、Yexin Wang、Zhenbang You、Meng Cao、Ping Huang、Jiulong Shan、Zhi Yang 和 Bin Cui。2022年。信息增益传播：一种基于软标签进行图主动学习的新方法。国际学习表征会议（2022年）。[111] Ziwei Zhang、Haoyang Li、Zeyang Zhang、Yijian Qin、Xin Wang 和 Wenwu Zhu。2023年。图遇见大语言模型：迈向大型图模型。载于神经信息处理系统大会*2023*年研讨会：图学习新前沿。[112] Zaixi Zhang、Qi Liu、Hao Wang、Chengqiang Lu 和 Chee-Kong Lee。2021年。用于分子性质预测的基于模体的图自监督学习。神经信息处理系统进展 34（2021年），15870–15882。[113] Haihong Zhao、Aochuan Chen、Xiangguo Sun、Hong Cheng 和 Jia Li。2024年。兼顾整体与个体：一种简单而有效的跨域图预训练方法。载于第*30*届*ACM SIGKDD*知识发现与数据挖掘会议论文集。4443–4454。[114] Wayne Xin Zhao、Kun Zhou、Junyi Li、Tianyi Tang、Xiaolei Wang、Yupeng Hou、Yingqian Min、Beichen Zhang、Junjie Zhang、Zican Dong等。2023年。大语言模型综述。*arXiv*预印本 *arXiv:2303.18223* 1，2（2023年）。[115] Yingli Zhou、Yaodong Su、Youran Sun、Shu Wang、Taotao Wang、Runyuan He、Yongwei Zhang、Sicong Liang、Xilin Liu、Yuchi Ma等。2025年。在统一框架下深入分析基于图的检索增强生成。*arXiv*预印本 *arXiv:2503.04338*（2025年）。[116] Tanli Zuo、Yukun Qiu 和 Wei-Shi Zheng。2021年。用于关键结构挖掘的邻居组合注意力。载于第二十九届国际人工智能联合会议论文集。3299–3305。

# A 符号说明

| 符号 | 说明 |
|---|---|
| 知识图谱与语料库 | |
| $G = (V, 召回率, E)$ | 包含实体集 V、关系集 R 和三元组集 E 的知识图谱。 |
| $(u, r, v) \in E$ | 从实体 $u$ 到实体 $v$、通过关系 $r$ 表示的关系三元组。 |
| $\{D\}, d \in D$ | 知识领域集合及一个领域标识符。 |
| $\mathrm{Dom}_d(e)$ | 表示实体 $e$ 属于领域 $d$的谓词。 |
| $C = \{c_i\}$ | 文档块语料库。 |
| $\phi(c_i)$ $\phi^{-1}(e)$ | 文档块与其中提及的实体之间的映射关系。 |
| 查询、答案和 *RAG* 接口 | |
| $\mathbf{q}, \mathcal{V} \mid \mathbf{q}$ | 输入查询，以及查询中明确提及的实体。 |
| 大语言模型$(\mathbf{q}, G_\mathbf{q}, C_\mathbf{q}), y$ | 答案生成器，以及生成的答案。 |
| $C_\mathbf{q}$ | 由检索到的实体引出的支持文档。 |
| 子图检索 | |
| $G_\mathbf{q} = (V_\mathbf{q}, 召回率_\mathbf{q}, E_\mathbf{q})$ | 查询特定的检索子图。 |
| $V_\mathbf{q}, 召回率_\mathbf{q}, E_\mathbf{q}$ | 检索子图的实体、关系和边集。 |
| 图基础模型（*GFM*） | |
| $f_\theta(\cdot)$、 $f_\theta^\star(\cdot)$、 $\theta$ | 图基础模型及其带参数的预训练变体。 |
| $\mathbf{Z}_e^{(l)}, \mathbf{Z}_r^{(l)}$ | GFM 层 $l$中的实体和关系嵌入。 |
| $N(v)$， $\mathbf{m}_v^{(l+1)}$ | 实体 $v$ 的邻域及聚合消息。 |
| $\mathrm{Msg}(\cdot), g(\cdot), \mathrm{Agg}(\cdot), \mathrm{Update}(\cdot)$ | 可学习的 GFM 消息传递和更新函数。 |
| $l, L$ | GNN层索引和总层数。 |
| 预训练与原型学习 | |
| $(e, r, ?), (?, r, e)$ | 用于知识图谱补全预训练的掩码三元组。 |
| $\mathbb{P}_\mathbf{q}(e)$ | 实体 $e$ 对于查询$\mathbf{q}$的相关性概率。 |
| $\mathcal{V}_\mathbf{q}^+, \mathcal{V}_\mathbf{q}^-$ | 正负实体集。 |
| $L_{\mathrm{BCE}}, L_{\mathrm{rank}}$ | 二元交叉熵和排序损失。 |
| $Q_d, (\mathbf{q}_i, e_i)$ | 从领域 $d$中采样的查询-实体对。 |
| $\mathbf{c}_d, \{\mathbf{c}_d\}$ | 领域原型嵌入和所有原型。 |
| $L^{\mathrm{proto}}, L_{\mathrm{IGC}}$ | 原型对比损失和信息增益对比正则项。 |
| $L^{\mathrm{pre}}, \alpha_1, \alpha_2, \alpha_3$ | 整体预训练目标和损失权重。 |
| 子图选择与信息瓶颈 | |
| $\mathbf{p}$ | 查询条件化提示。 |
| $\tilde{\mathbf{Z}}_e^\star, \mathbf{Z}_e$ | 冻结实体嵌入，以及查询条件化实体表示。 |
| $\mathbf{W}_1, \mathbf{W}_2$ | 可学习投影矩阵。 |
| $S_e(\mathbf{q})$， $g_e, \epsilon$ | 软实体选择概率、Gumbel噪声和选择阈值。 |
| $I(\cdot;\cdot)$， $H(\cdot \mid \cdot)$ | 互信息和条件熵。 |
| $\mathbf{S}_e(\mathbf{q})$， $\mathbf{L}$ | 软选择向量和归一化图拉普拉斯矩阵。 |
| $\beta, \beta_1, \beta_2$ | 信息瓶颈正则化系数。 |
| $L_{\mathrm{IB}}, L_{\mathrm{LIB}}$ | 信息瓶颈目标函数及其无标签变体。 |
| $G_\mathbf{q}$， $L_{\mathrm{NCE}}$ | 池化子图表示和对比下界损失。 |
| $L_{\mathrm{size}}, L_{\mathrm{con}}$ | 子图大小与连通性正则化项。 |
| $\mathcal{L}_{\mathrm{ftn}}$ | 总体微调目标。 |
| 路径推理与提示构建 | |
| $\mathbf{M}$ | 二元实体-文档关联矩阵。 |
| $w_e, R_c$ | 实体重要性权重与文档相关性得分。 |
| $\Pi_\mathbf{q}$， $\pi$ | 推理路径集合和单条关系路径。 |
| $S_r(\pi)$ 、 $\phi_{\mathrm{path}}(e)$ 、 $\omega(r)$ | 路径相关性得分、基于路径的实体中心性和关系深度惩罚。 |

**算法1：** GFM 检索器 $f_\theta(\cdot)$ 的预训练。

**输入：** 多领域索引知识图谱 $\mathcal{G} = (\mathcal{V}, \mathcal{R}, \mathcal{E})$，覆盖领域 $\mathcal{D}$；查询条件化 GFM $f_\theta(\cdot)$，具有 $L$ 层；超参数 $\alpha_1, \alpha_2, \alpha_3, \tau$。

**输出：** 预训练的 GFM 检索器 $f_\theta^\star(\cdot)$。

初始化模型参数 $\theta$;

// # 第一阶段：知识图谱补全预训练

**对** 每个预训练周期**执行**

  **对** 从$\mathcal{G}$中采样的每个三元组$(e, r, ?)$或$(?, r, e)$**执行**

    根据三元组构造伪查询$\mathbf{q}$；

    初始化嵌入：$\mathbf{Z}_e^{(0)} \leftarrow$式 (1);

    **为** $l = 0$ 到 $L-1$ **做**

      消息传递：$\mathbf{m}_v^{(l+1)} \leftarrow$ 式（2）；

      实体更新：$\mathbf{Z}_v^{(l+1)} \leftarrow$式（3）；

    **end**

    预测实体相关性 $\mathbb{P}_\mathbf{q}(e) \leftarrow$式（4）；

    损失：$\mathrm{L_{BCE}} \leftarrow$式（5），$\mathrm{L_{rank}} \leftarrow$式（6）；

    $\mathrm{L_{Phase\ I}} = \alpha_1 \mathrm{L_{BCE}} + (1-\alpha_1)\mathrm{L_{rank}}$ （式（9））；

  **end**

**结束**

// # 第二阶段：连续预训练

**对于**每个预训练周期**执行**采样配对 $(\mathbf{q}_i, e_i) \in Q_d$ 从每个领域 $d \in \mathcal{D}$；计算领域原型$\mathbf{c}_d$；计算原型对比损失：$\mathcal{L}_{\text{proto}} \leftarrow$式（7）；计算信息增益对比$\mathcal{L}_{\text{IGC}} \leftarrow$式（8）；$\mathcal{L}_{\text{Phase II}} = \alpha_2 \mathcal{L}_{\text{proto}} + \alpha_3 \mathcal{L}_{\text{IGC}}$ （式（9））。；**结束**

## B 算法与复杂度分析

Gfm-Retriever 的预训练、微调和推理（问答任务）流程如算法 1、算法 2 和算法3所示。接下来，我们将分析它们的计算复杂度。

### B.1 预训练阶段的计算复杂度

预训练阶段分为两个阶段：**(1)** 混合域知识图谱补全预训练（第一阶段），以及**(2)** 原型驱动的持续预训练（第二阶段）。除非另有说明，我们关注相对于图规模的渐近复杂度，并忽略与嵌入维度及神经模块内部实现（例如，Msg(·)、Update(·)、Agg(·)、MLP(·)等）相关的常数因子。

**第一阶段：混合域知识图谱补全预训练**。第一阶段使用查询条件化的GFM推理，在多域索引图 $\mathcal{G} = (\mathcal{V}, \mathcal{R}, \mathcal{E})$ 上执行知识图谱补全。针对每个掩码三元组，构造一个伪查询$\mathbf{q}$。

• **嵌入初始化**。为所有 $e \in \mathcal{V}$ 和 $r \in \mathcal{R}$ 初始化实体和关系嵌入的复杂度为 $O(|\mathcal{V}| + |\mathcal{R}|)$。

• **查询条件化消息传递**。对于每个 GFM 层，都会在所有边（$u$, $r$, $v$）$\in \mathcal{E}$ 上计算消息，并在每个节点处进行聚合。每层的复杂度为 $O(|\mathcal{E}| + |\mathcal{V}|)$，堆叠 $L$ 层后的复杂度为 $O(L(|\mathcal{E}| + |\mathcal{V}|))$。

• **实体相关性评分**。要计算所有实体的 $\mathbb{P}_\mathbf{q}(e)$，需要执行一次多层感知机的前向传播，其计算量为 $O(|\mathcal{V}|)$。

• **知识图谱补全损失**。二元交叉熵损失和排序损失都需要对正负实体集求和，而这些实体集的并集等于 $\mathcal{V}$。因此，它们的计算成本为 $O(|\mathcal{V}|)$。

综上，第一阶段的时间复杂度为 $O(L|\mathcal{E}| + L|\mathcal{V}| + |\mathcal{E}| + |\mathcal{V}|)$。由于 $|\mathcal{E}| \geqslant |\mathcal{V}|$ 和 $L \geqslant 1$，可将其简化为 $O(L|\mathcal{E}| + |\mathcal{V}|)$。

其中 $O(|\mathcal{E}|)$ 是主导项。若每轮对 $\mathcal{E}$ 中的所有三元组进行一次掩码处理，则每轮复杂度为：

$$O\big(L|\mathcal{E}|^2 + |\mathcal{E}||\mathcal{V}| + |\mathcal{E}||\mathcal{R}|\big), \tag{B.1}$$

其主导项为 $O(L|\mathcal{E}|^2)$。

**第二阶段：原型驱动的持续预训练**。第二阶段使用从每个领域中采样的部分查询-实体对 $\bigcup_{d\in\mathcal{D}} Q_d$ 进行语义对齐。

• **原型构建**。计算域原型需要对实体表示取平均，计算量为 $O(\sum_{d\in\mathcal{D}} |Q_d|)$。

• **原型对比损失**。对于每一对，softmax 分母对所有领域求和，得到 $O((\sum_{d\in\mathcal{D}} |Q_d|) \cdot |\mathcal{D}|)$。

• **信息增益对比正则化**。KL 散度基于实体分布 $\mathbb{P}_\mathbf{q}(e) \in \mathbb{R}^{|\mathcal{V}|\times 1}$ 计算。对于每个查询，这需要 $O(|\mathcal{V}|)$ 的时间，因此得到 $O((\sum_{d\in\mathcal{D}} |Q_d|) \cdot |\mathcal{V}|)$。

• **GFM 前向推理**。由于 $\mathbb{P}_\mathbf{q}(e)$ 取决于查询条件化的 GFM 输出，每个查询都需要进行一次 $L$ 层消息传递前向传播，耗费 $O(L(|\mathcal{E}| + |\mathcal{V}|))$。对于所有采样查询，这部分的复杂度为 $O((\sum_{d\in\mathcal{D}} |Q_d|) \cdot L|\mathcal{E}|)$，这是第二阶段的主要开销。

综上，第二阶段每轮的复杂度为：

$$O\Big(\big(\sum_{d\in\mathcal{D}} |Q_d|\big) \cdot (L|\mathcal{E}| + |\mathcal{V}| + |\mathcal{D}|)\Big). \tag{B.2}$$

**总体预训练复杂度**。结合第一阶段与第二阶段，预训练的每轮次总时间复杂度为：

$$O\big(L|\mathcal{E}|^2 + |\mathcal{E}||\mathcal{V}| + |\mathcal{E}||\mathcal{R}|\big) + O\Big(\sum_{d\in\mathcal{D}} |Q_d|(L|\mathcal{E}| + |\mathcal{V}| + |\mathcal{D}|)\Big). \tag{B.3}$$

由于第二阶段在子集上执行，而第一阶段在大规模知识图谱中占主导，总体复杂度渐近地由 $O(L|\mathcal{E}|^2)$ 支配；这表明预训练的计算成本主要由在整个索引图上反复进行的查询条件化消息传递决定，从而促使我们在下游阶段采用参数高效微调与子图级优化。

### B.2 微调阶段的计算复杂度

微调阶段通过可处理的替代目标，利用 Proposition 4.3（公式 (15)）和 Proposition 4.4（公式 (17)）优化无标签 IB 目标（公式 (13)），并结合有监督检索损失（公式 (5) 和公式 (6)）。遵循附录 B.1 中相同的协议，我们省略与隐藏维度和神经模块内部实现（例如，MLP(·)）相关的常数因子，重点关注对图和批次大小的渐近依赖关系。

**算法 2：** 使用子图选择器进行微调。

**输入：** 多领域索引知识图谱 $G = (V, R, E)$；冻结实体嵌入 $\mathbf{Z}_e^\star$ 来自 $f_\theta^\star(\cdot)$；训练查询 $\{(q, V_q^+)\}$；超参数 $\alpha_1, \beta_1, \beta_2, \tau, \epsilon$。

**输出：** 训练得到的子图选择器，生成 $G_q$。

**对** 每个查询小批量 $\mathcal{B}$ **执行**
  **对**每个查询 $q \in \mathcal{B}$ **执行**
    构造查询提示 $\mathbf{p}$ =多层感知机(q);
    **对** 每个实体 $e \in \mathcal{V}$ **执行**
      // 学习查询条件化子图
      查询条件化实体：$\tilde{\mathbf{Z}}_e \leftarrow$ 公式 (10);
      样本选择门：$S_e(q) \leftarrow$ 式（11）；
    **end**
    通过对 $\epsilon$进行阈值化来诱导子图 $G_q$；
    // 无标签信息瓶颈正则化
    检索损失：$L_{BCE} \leftarrow$式（5），$L_{rank} \leftarrow$式（6）；
    池化子图表示：$\mathbf{G}_q = \sum_{e \in V_q} S_e(q) \cdot \tilde{\mathbf{Z}}_e$;
    $I$（q；$G_q$）的下界：$L_{NCE} \leftarrow$ 式（15）；
    惩罚项：$L_{size} \leftarrow$式（16），$L_{con} \leftarrow$式（16）；
    结合信息瓶颈代理目标：$\lceil L_{LIB} \rceil \leftarrow$式（19）；
    计算最终的微调损失：$L_{ftn} \leftarrow$式（18）；
  **end**
  通过梯度下降更新选择器参数;
**结束**

**每查询复杂度（小批量 B）**。对于查询 $q$，微调阶段包含以下计算模块。

• **查询提示构造**。 提示 $\mathbf{p}$ = 多层感知机(q)对每个查询计算一次。将多层感知机(·)视为常数代价算子相对于 图规模，该步骤需要 $O(1)$。

• **查询条件实体嵌入构造**。 对于每个实体 $e \in V$，我们构造 $\tilde{\mathbf{Z}}_e = [\mathbf{Z}_{\star e} \| \mathbf{W}_1 \mathbf{p} \| \mathbf{W}_2 q]$，这是一种逐实体操作。因此，复杂度为 $O(|V|)$。

• **可微子图门采样**。 对于每个 $e \in V$，式 (11) 中的 Gumbel-Sigmoid 门 $S_e(q)$ 计算一次。该步骤同样与实体数量呈线性关系：$O(|V|)$。

• **基于阈值的子图归纳**。 构造 $V_q$ 需要对所有实体进行一次遍历，其复杂度为 $O(|V|)$。

• **子图生成**。 由 $V_q$ 诱导 $R_q$ 和 $E_q$ 可通过扫描边完成，复杂度至多为 $O(|E|)$。若采用完整边扫描，我们在下文将其明确列为可选的主导项。

• **监督检索损失**。 LBCE 和 Lrank 都对正样本 $V^+_q$ 以及负样本 $V^-_q$ 求和。在最坏情况下，它们会在所有实体上进行计算，产生 $O(|V|)$ 的复杂度，这与公式 (5) 和公式 (6) 对实体级分布的依赖相匹配。

• **池化子图表示**。 池化后的$G_q$ 需要遍历选定的实体集 $V_q$：$O(|V_q|) \leqslant O(|V|)$。 • **噪声对比估计项**。 式（15）计算一个分母，对负样本对（$G_{q'}, q'$）求和，其中$q' \in B$。假设所有$q' \in B$对应的$G_{q'}$均已预先计算（如算法 2 所示），则每个查询计算 Softmax 归一化得分的开销为 $O(|B|)$。

（如果$G_{q'}$ 未缓存，开销还需计入 $O(|V_{q'}|)$ 的构建成本，这部分已由池化表示的开销涵盖。） • **规模惩罚**。 规模项 $L_{size}$ 的开销为 $O(|V_q|) \leqslant O(|V|)$。

• **连通性惩罚**。 连通性项 Lcon 涉及与归一化拉普拉斯矩阵 $L$ 的乘法。计算 $LS_e(q)$ 可以通过稀疏矩阵-向量乘法实现，其稀疏模式由 $E$ 给出，计算成本为 $O(|E|)$；这通常是各正则项中的主导项。

**每批次复杂度**。 综上，单个查询$q$的计算复杂度为：

$$\underbrace{O(|\mathcal{V}|)}_{\text{Eq. (10)}} + \underbrace{O(|\mathcal{V}|)}_{\text{Eq. (11)}} + \underbrace{O(|\mathcal{V}|)}_{\text{thresholding}} + \underbrace{O(|\mathcal{V}|)}_{\mathcal{L}_{\text{BCE}},\, \mathcal{L}_{\text{rank}}} + \underbrace{O(|\mathcal{V}_q|)}_{\mathbf{G}_q,\, \mathcal{L}_{\text{size}}} + \underbrace{O(|\mathcal{B}|)}_{\mathcal{L}_{\text{NCE}}} + \underbrace{O(|\mathcal{E}|)}_{\mathcal{L}_{\text{con}}}. \tag{B.4}$$

使用 $|V_q| \leqslant |V|$ 并对各项进行归类，可得 $O(|V|+|E|+|B|)$。

对于一个小批量 B，总成本为 $O(|B| \cdot (|V| + |E| + |B|))$。展开乘积可得 $O(|B||V| + |B||E| + |B|^2)$。在典型的大型知识图谱中，$|E|$ 占主导地位，高于 $|V|$ 和 $|B| \ll |E|$。因此，微调的渐近复杂度由连通性正则化项主导，为 $O(|B||E|)$，其中每个查询的主导成本为 $O(|E|)$。

值得注意的是，如果每次迭代都通过扫描完整边集显式生成 $E_q$ 和 $R_q$，则每个查询的额外开销至多为 $O(|E|)$，且已包含在上述 $O(|E|)$ 项中。相比之下，若使用仅限于选定节点集 $V_q$ 的边索引，实际运行时间可以缩短，但最坏情况下的渐近上界保持不变。

## B.3 推理阶段（问答）的计算复杂度

给定目标知识图谱 $G^T = (V^T, R^T, E^T)$ 上的查询 $q$，推理阶段包括**(1)** 通过训练好的选择器进行子图检索（式 (11)），以及 **(2)** 在上下文提示中生成关系路径，包括实体到文档映射和截断深度优先搜索路径提取。我们省略与隐藏维度和神经模块内部结构相关的常数因子。

**逐步复杂度分析**。对于单个查询$q$，算法 3 包含以下模块。

• **选择概率**。选择器计算 $S_e(q)$（式 (11)），这需要遍历所有实体一次，因此复杂度为 $O(|V^T|)$。
• **基于阈值的子图归纳**。构建 $V_q$ 的成本为 $O(|V^T|)$。如果通过扫描 $E^T$ 中的所有边来显式地从 $V_q$ 诱导 $E_q$，则最坏情况下的额外成本为 $O(|E^T|)$。

• **实体加权**。对于每个实体 $e \in V_q$，推理会计算 $w_e$(式 (20))，其时间复杂度与所选实体集的大小呈线性关系：$O(|V_q|)$。（项 $\sum_c \mathbf{M}[e, c]$ 可根据固定映射矩阵 $\mathbf{M}$ 为每个实体预先计算。）

• **文档相关性计算**。文档得分由 $R_c = \mathbf{M}^\top w_e$ 给出。将 $\mathbf{M} \in \{0,1\}^{|\mathcal{E}_q| \times |\mathcal{C}_q|}$ 视为稀疏矩阵后，计算 $\mathbf{M}^\top w_e$ 相当于一次稀疏矩阵-向量乘法。其时间复杂度与限制在所选实体上的 $\mathbf{M}$ 中的非零元素数（nnz）成正比，即 $O(\text{nnz}(\mathbf{M}_{\mathcal{V}_q}))$，其中 $\mathbf{M}_{\mathcal{V}_q}$ 表示由 $V_q$诱导的子矩阵。最坏情况下，时间复杂度的上界为扫描查询范围内所有实体到文本块链接，即 $O(|E_q||C_q|)$；但在实际应用中，由于稀疏性，通常远小于此值。

**算法3：** 基于上下文内提示的推理。

**输入：** 查询 $\mathbf{q}$；目标知识图谱 $\mathrm{G}^T$；训练好的选择器生成 $S_e(\mathbf{q})$；实体到文档映射矩阵 $\mathbf{M}$；深度优先搜索搜索跳数限制 $l$；前$K$ 个文档 $\mathrm{C}_\mathbf{q}^{(K)}$。**输出：** 生成的答案 $\hat{y}$。

// 子图检索
所有 $e \in \mathrm{V}^T$ 的选择概率 $S_e(\mathbf{q}) \leftarrow$，公式 (11)；
诱导子图 $\mathrm{G}_\mathbf{q} = (\mathrm{V}_\mathbf{q}, \mathrm{R}_\mathbf{q}, \mathrm{E}_\mathbf{q})$；
// 实体到 -文档评分
**对**每个实体 $e \in \mathcal{V}_\mathbf{q}$ **执行**
  计算重要性权重 $w_e \leftarrow$ 公式 (20)；
**结束**
计算相关性 $R_c = \mathbf{M}^\top w_e$ 并选择排名前$K$ 的文档；
// 推理路径提取。
通过 $l$跳截断深度优先搜索，从 $\mathrm{G}_\mathbf{q}$ 中提取关系路径 $\Pi_\mathbf{q}$；
**对**每条路径 $\pi$ **执行**
  计算路径得分 $S_r(\pi)$，参见公式（21）。
**结束**
选择排名靠前的路径 $\Pi_\mathbf{q}^\star$;
// 上下文内提示与生成
根据 $\Pi_\mathbf{q}^\star$ 和检索到的文档构建提示；
生成答案 $\hat{y} = \mathrm{LLM}(\mathbf{q}, \Pi_\mathbf{q}^\star, \mathrm{C}_\mathbf{q}^{(K)})$;

• **Top-$K$ 文档选择**。从 $\{R_c\}$ 中选出前 $K$ 个文档，可以在 $O(|C_\mathbf{q}|)$ 时间内完成（例如，使用线性时间选择）；若采用完全排序，则需要 $O(|C_\mathbf{q}| \log |C_\mathbf{q}|)$ 时间。
• **关系路径提取**。在最坏情况下，由于深度优先搜索中的分支，枚举所有路径（$\leqslant l$）的数量可能呈指数级增长。设诱导子图 $\mathrm{G}_\mathbf{q}$ 的最大度为 $\Delta$，则深度优先搜索扩展的最坏情况数量为 $O(\Delta^l)$。

• **路径评分**。对于每条提取的路径 $\pi$，推理计算 $S_r(\pi)$（式 (21)），每条路径的代价为 $O(k) \leqslant O(l)$，因此对所有提取路径需要 $O(|\Pi_\mathbf{q}| \cdot l)$。

• **路径排序和子集选择**。按 $\{S_r(\pi)\}$ 对路径排序的代价为 $O(|\Pi_\mathbf{q}| \log |\Pi_\mathbf{q}|)$，或在 top-$K$ 选择下为 $O(|\Pi_\mathbf{q}|)$。• **大语言模型答案生成**。答案生成 $\hat{y} = \mathrm{LLM}(\cdot)$ 取决于大语言模型的推理代价和提示长度。为简洁起见，我们将其省略。

**总体推理复杂度**。排除大语言模型生成代价后，每次查询的推理复杂度为：

$$O(|\mathcal{V}^T|) + \underbrace{O(|\mathcal{E}^T|)}_{\text{(optional, if inducing } \mathcal{E}_\mathbf{q} \text{ by edge scan)}} + O(|\mathcal{V}_\mathbf{q}|) + O\big(\mathrm{nnz}(\mathbf{M}_{\mathcal{V}_\mathbf{q}})\big) + O(|C_\mathbf{q}|) + O(\Delta^l) + O(|\Pi_\mathbf{q}| \cdot l) + O(|\Pi_\mathbf{q}| \log |\Pi_\mathbf{q}|). \quad \text{(B.5)}$$

由于 $|\mathrm{V}_\mathbf{q}| \leqslant |\mathrm{V}^T|$，我们得到简化界：

$$O\Big(|\mathcal{V}^T| + |\mathcal{E}^T| + \mathrm{nnz}(\mathbf{M}_{\mathcal{V}_\mathbf{q}}) + \Delta^l + |\Pi_\mathbf{q}| \log |\Pi_\mathbf{q}|\Big). \quad \text{(B.6)}$$

在典型设置中，主要开销来自截断深度优先搜索路径提取，其相对于跳数预算 $l$ 在最坏情况下可能呈指数级增长。

# C 证明

## C.1 命题 4.1 的证明

我们首先重述命题 4.1 以供参考。

> 命题 *4.1*（查询条件化 *GFM* 的多域逻辑表达能力）设 $G$ 为从多个领域索引构建的知识图谱。$\{Dom_d(e)\}$ 表示指示实体 $e$ 域属性的一元谓词。对于查询 $q$，查询条件化 *GFM* 能够学习规则 $R(q\ e)$ 当且仅当 $R(q\ e)$ 是 $G$ 上以 $q$ 为常量的分级模态逻辑公式，即 $\exists^{\geqslant N} e'(R(e', e) \wedge Dom_d(e') \wedge \psi(e'))$，其中 $\psi(\cdot)$ 是子公式。

证明。设 $\mathrm{G} = (\mathrm{V}, \mathrm{R}, \mathrm{E})$ 为从多领域文档索引构建的知识图谱。$\mathrm{D} : \mathrm{V} \to \{0, 1\}^{|\mathbb{D}|}$ 将每个实体节点映射到其域属性。我们将一元谓词 $\mathrm{Dom}_d(v)$ 定义为：当 $\mathrm{D}(v)$ 的第 $d$ 个条目为 1 时为真。我们假设查询条件化 GFM $f_\theta(\cdot)$ 遵循式 (2) 和式 (3) 中定义的消息传递方案。为证明命题 4.1，我们将该 GFM 可学习的逻辑公式与扩展了域谓词的分级模态逻辑 (CML) [11, 65, 68] 之间建立等价关系。

**(1) 充分性（构造性证明）**。我们证明，对于多领域 CML 逻辑中的任意公式 $\varphi$，都存在一种 GFM 参数化方式，使其能够学习 $\varphi$。我们对公式的构造进行结构归纳。基础情形是原子公式（例如，$\mathrm{Dom}_d(e)$）直接包含在初始实体嵌入 $\mathbf{Z}_e^{(0)}$ 中。简单的线性投影（1 层 GFM）即可对这些嵌入进行编码，因此基础谓词是可学习的。对于归纳步骤，假设 GFM 能够通过 $l-1$ 层学习子公式 $\psi(e')$，且 $\mathbf{Z}_e^{(l-1)}$ 的维度 $j$ 可作为 $\mathbf{Z}_e^{(l-1)}$，$[j] = \mathbb{I}(\mathrm{G}\ e' \models \psi)$ 的指示，其中 $\mathbb{I}(\cdot)$ 是指示函数。

接下来，我们证明公式

$$\Phi\ e\ = \exists^{\geqslant N} e'\quad e'\ e\ \wedge\quad {}_d\ e'\ \wedge (\ )\ (\mathrm{R}(,\ )\ \mathrm{Dom}\ (\ )\ (\ ))$$
$$\psi\ e' \qquad\qquad l \qquad\qquad \Phi\ e$$

可由第 层学习。( ) 的语义要求对通过关系 R 连接、且同时满足领域约束 $\mathrm{Dom}_d$ (·) 和性质 $\psi$ 的邻居 $e'$ 进行计数。我们按照引理 C.3 [68] 构造第 $l$ 层的聚合：

**消息构造**。令消息函数 $\mathrm{Msg}^{(l)}{}_\mathrm{R}(\cdot)$ 选择并组合相关特征。由于 $\mathrm{Dom}_d(e')$ 是输入特征中可用的一元谓词（并且在第 $(l-1)$ 层得到保留或传播），令 $\mathbf{Z}_e^{(l-1)}, [k]$ 对应于 $\mathrm{Dom}_d(e')$。我们定义消息仅在两个条件都满足时才输出标量值 1，实质上通过软连续近似实现逻辑“与” ($\wedge$)（例如，使用求和组件或在极限情况下使用多层感知机近似）：

$$\mathbf{m}_{e',e} \doteq \mathbb{I}\big(\mathbf{Z}_{e'}^{(l-1)}[k] = 1 \wedge \mathbf{Z}_{e'}^{(l-1)}[j] = 1\big). \quad \text{(C.1)}$$

或者，在标准线性图神经网络中，$\mathbf{W}_{\mathrm{MSG}}$ 会为 $\psi$ 维度赋予单位权重，前提是 $e'$ 被领域 $d$屏蔽或过滤。更稳健的构造是利用多层感知机的通用逼近定理（包含在 $\mathrm{Msg}(\cdot)$ 中），将 $\mathbf{Z}_e^{(l-1)}$，映射为：当领域和 $\psi$ 位均处于激活状态时映射为 1，否则映射为 0。

**聚合**。GFM 对邻居发送的消息求和：

$$S_e = \sum_{e' \in \mathcal{N}_\mathrm{R}(e)} \mathbf{m}_{e',e} = |\{e' \mid \mathrm{R}(e', e) \wedge \mathrm{Dom}_d(e') \wedge \psi(e')\}|. \quad \text{(C.2)}$$

该求和 $S_e$ 精确计算特定领域中有效逻辑见证的数量。

**阈值化（量词实现）**。为实现量词 $\exists^{\geqslant N}$，我们设置偏置项 $\mathbf{b}^{(l)}$，并使用阶跃型激活函数 $\sigma$（例如，陡峭的 Sigmoid 函数）。假设 $\mathbf{b} = -N + \delta$，则：

$$\mathbf{Z}_e^{(l)} = \sigma(S_e - N + \delta), \quad \delta > 0. \tag{C.3}$$

如果 $S_e \geqslant N$， $\sigma$的输入为正，则输出 1（真）。如果 $S_e < N$，则输出 0（假）。

因此，GFM 可以在结构上实现公式 $\Phi(e)$。

**（2）必要性（表达能力限制）**。根据 Qiu 等人中的定理 C.2 和定理 C.7 [68],，我们证明，查询条件化 GFM 能够学习的任何规则都对应于 G 上的一个 CML 公式。我们借助消息传递神经网络与一维 Weisfeiler-Leman（1-WL）测试之间已确立的对应关系 [30]。众所周知，标准消息传递神经网络的判别能力受限于 1-WL。此外，图上的 1-WL 稳定着色等价于由带计数量词的 CML 定义的逻辑等价类 [4, 60]。在我们的具体情形中：

**图增强**。多域图 G 可以视为一个标准着色图，其初始节点颜色 $C(v)$ 是如下元组：$(\mathbb{I}(v = \mathbf{q}),\ \mathrm{Dom}_{d_1}(v),\ \cdots,\ \mathrm{Dom}_{d_k}(v),\ \cdots)$。

**双模拟不变性**。 GFM 是置换不变的。如果两个实体 $u$ 与 $v$ 无法被 CML 公式区分（即，在增广图上双模拟），则它们在每一步都会接收到相同的消息聚合，从而得到相同的最终嵌入。

综上所述，如果 GFM 学习到一个分类器，能够基于 $\mathrm{R}(\mathbf{q}\ e)$ 区分实体 $e$，则计算树中必然存在一种可区分结构。由于计算树与 CML 中展开的逻辑公式同构，GFM 所学到的决策边界必须能够由 CML 公式描述。具体而言，“多领域”性质只是扩展了逻辑中原子一元谓词的集合。如果某条规则依赖于检查 $\mathrm{Dom}_d$，这在 CML 中作为检查原子命题是有效的。如果该规则无法用 CML 表达（例如，它需要计数环，或检查超出树展开范围的连通性），则 1-WL 测试将无法区分相关子图，GFM 也因此无法学习该规则。因此，一条规则只有在能够用带域谓词的 CML 表达时才是可学习的。 □

## C.2 命题 4.2 的证明

我们先重述命题 4.2，以供参考。

命题 4.2 (**L 近似的误差界 $\mathcal{L}_{\mathrm{LIB}}$ 近似**) 令 $\mathcal{G}_\mathbf{q}$ 为子图， $y$ 为真实答案，$\mathbf{q}$ 为查询。假设马尔可夫性质 $\langle y \leftrightarrow \mathbf{q} \leftrightarrow \mathcal{G}_\mathbf{q}\rangle$ 存在，则 $\mathcal{L}_{\mathrm{IB}}$ 相对于 $\mathcal{L}_{\mathrm{LIB}}$ 的近似误差以上界受给定 $y$时$\mathbf{q}$ 的条件熵所限：

$$\left|I(y;\mathcal{G}_\mathbf{q}) - I(\mathbf{q};\mathcal{G}_\mathbf{q})\right| \leqslant H(\mathbf{q} \mid y),$$

其中， $H(\mathbf{q} \mid y)$ 是一个依赖于数据的常数 .

证明。我们利用交互信息和互信息链式法则，对无标签目标引入的近似误差进行形式化分析。令随机变量$y$、$\mathbf{q}$和 $\mathrm{G}_\mathbf{q}$ 分别表示真实答案、查询和检索子图。考虑联合互信息 $I(y$、$\mathbf{q}$; $\mathrm{G}_\mathbf{q})$。根据对称性，我们可以将该项分解为

基于链式法则的两种等价形式：

$$I(y, \mathbf{q}; \mathcal{G}_\mathbf{q}) = I(y; \mathcal{G}_\mathbf{q}) + I(\mathbf{q} \mid y; \mathcal{G}_\mathbf{q}), \tag{C.4}$$

$$I(y, \mathbf{q}; \mathcal{G}_\mathbf{q}) = I(\mathbf{q}; \mathcal{G}_\mathbf{q}) + I(y \mid \mathbf{q}; \mathcal{G}_\mathbf{q}). \tag{C.5}$$

从式 (C.5) 中减去式 (C.4)，即可分离监督目标 $I(y; \mathrm{G}_\mathbf{q})$ 与我们提出的无标签代理目标 $I(\mathbf{q}; \mathrm{G}_\mathbf{q})$ 之间的差异：

$$I(y; \mathcal{G}_\mathbf{q}) - I(\mathbf{q}; \mathcal{G}_\mathbf{q}) = I(y \mid \mathbf{q}; \mathcal{G}_\mathbf{q}) - I(\mathbf{q} \mid y; \mathcal{G}_\mathbf{q}). \tag{C.6}$$

为推导上界，我们回顾互信息受其变量的熵所界定这一性质。具体而言，对于条件互信息，有：

$$I(X; Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z). \tag{C.7}$$

由于熵非负，且条件化会降低熵或使其保持不变，因此我们知道：

$$I(X; Y \mid Z) \leqslant H(Y \mid Z), \tag{C.8}$$

其中， $H(\cdot)$ 表示香农熵。将此不等式应用于式 (C.6) 右侧的两项：

$$I(y \mid \mathbf{q}; \mathcal{G}_\mathbf{q}) \leqslant H(y \mid \mathbf{q}), \tag{C.9}$$

$$I(\mathbf{q} \mid y; \mathcal{G}_\mathbf{q}) \leqslant H(\mathbf{q} \mid y). \tag{C.10}$$

将这些界代回绝对差误差：

$$\begin{aligned}\left|I(y; \mathcal{G}_\mathbf{q}) - I(\mathbf{q}; \mathcal{G}_\mathbf{q})\right| &\leqslant \max\left(H(y \mid \mathbf{q}) + H(\mathbf{q} \mid y)\right) \\ &\leqslant H(y \mid \mathbf{q}) + H(\mathbf{q} \mid y).\end{aligned} \tag{C.11}$$

在检索场景中，查询旨在指定答案（即，$\mathbf{q}$ 与 $y$不独立）。若假设检索子图 $\mathrm{G}_\mathbf{q}$ 仅通过$\mathbf{q}$的语义包含关于 $y$ 的信息（马尔可夫性质 $y \leftrightarrow \mathbf{q} \leftrightarrow \mathrm{G}_\mathbf{q}$），则该界可收紧为：

$$\left|I(y; \mathcal{G}_\mathbf{q}) - I(\mathbf{q}; \mathcal{G}_\mathbf{q})\right| \leqslant H(\mathbf{q} \mid y). \tag{C.12}$$

虽然 $H(\mathbf{q} \mid y)$ 表示指向同一答案的查询所具有的内在多样性，但需要注意，这一项仅由数据分布 $\mathbb{P}(\mathbf{q}\ y)$ 决定，且相对于子图选择器中可学习的参数保持不变。因此，最小化上界等价于在一个常数加性因子的范围内最小化代理间隙。优化充分性的梯度方向基本不受这一项影响。

总之，命题 4.2 表明，在查询歧义 $H(y \mid \mathrm{G}_\mathbf{q})$ 较低的情况下，最大化查询 $I(\mathrm{G}_\mathbf{q}; \mathbf{q})$ 的互信息，是 $I(\mathrm{G}_\mathbf{q}; y)$ 的有效替代指标。由于查询的设计目标是指向答案（歧义较低），最小化无标签损失便能在可控的误差界限内有效地最小化监督损失。□

## C.3 命题 4.3 的证明

我们首先重述命题 4.3，以供参考。

命题 4.3 (**下界 $I(\mathbf{q}; \mathcal{G}_\mathbf{q})$**) 将平均池化的子图表示记为 $\mathrm{G}_\mathbf{q} = \sum_{e \in \mathcal{V}_\mathbf{q}} S_e(\mathbf{q}) \cdot \tilde{\mathbf{Z}}_e$。根据 Alemi 等人 [1],，下界为：

$$I(\mathbf{q}; \mathcal{G}_\mathbf{q}) \geqslant -\mathcal{L}_{\mathrm{NCE}} = \log \frac{\exp\left(\mathrm{sim}(\mathrm{G}_\mathbf{q}, \mathbf{q})/\tau\right)}{\sum_{\mathbf{q}' \in \mathcal{B}} \exp\left(\mathrm{sim}(\mathrm{G}_{\mathbf{q}'}, \mathbf{q}')/\tau\right)},$$

其中 （$\mathrm{G}_{\mathbf{q}'}$,$\mathbf{q}'$）是负样本对 .

证明。主要目标是最大化互信息 $I(\mathbf{q}; \mathrm{G_q})$，以确保检索子图 $\mathrm{G_q}$ 保留与查询 $\mathbf{q}$ 相关的充分语义信息。由于精确计算需要对联合分布进行高维积分，我们通过变分估计推导出一个易于处理的下界。

**变分下界推导**。根据定义，互信息表示给定 $\mathrm{G_q}$ 后，$\mathbf{q}$ 的不确定性降低量：

$$I(\mathbf{q}; \mathcal{G}_\mathbf{q}) = H(\mathbf{q}) - H(\mathbf{q} \mid \mathcal{G}_\mathbf{q}). \tag{C.13}$$

由于查询分布 $H(\mathbf{q})$ 的熵相对于子图选择器是常数，因此，最大化互信息等价于最小化条件熵 $H(\mathbf{q} \mid \mathrm{G_q})$。然而，真实的后验分布 $\mathbb{P}(\mathbf{q} \mid \mathrm{G_q})$ 未知。我们引入变分分布 $\mathbb{Q}(\mathbf{q} \mid \mathrm{G_q})$ 来近似该后验分布。利用 Kullback-Leibler 散度的非负性，$\mathrm{KL}(\mathbb{Q}(\mathbf{q} \mid \mathrm{G_q}) \,||\, \mathbb{Q}(\mathbf{q} \mid \mathrm{G_q})) \geqslant 0$，我们得到以下界：

$$\begin{aligned} I(\mathbf{q}; \mathcal{G}_\mathbf{q}) &= \mathbb{E}_{\mathbb{P}(\mathbf{q}, \mathcal{G}_\mathbf{q})} \left[ \log \frac{\mathbb{P}(\mathbf{q} \mid \mathcal{G}_\mathbf{q})}{\mathbb{P}(\mathbf{q})} \right] \\ &= \mathbb{E}_{\mathbb{P}(\mathbf{q}, \mathcal{G}_\mathbf{q})} \left[ \log \mathbb{Q}(\mathbf{q} \mid \mathcal{G}_\mathbf{q}) \right] + \mathrm{KL}(\mathbb{P} \,||\, \mathbb{Q}) \\ &\geqslant \mathbb{E}_{\mathbb{P}(\mathbf{q}, \mathcal{G}_\mathbf{q})} \left[ \log \mathbb{Q}(\mathbf{q} \mid \mathcal{G}_\mathbf{q}) \right], \end{aligned} \tag{C.14}$$

这意味着，最大化变分分布的对数似然是互信息的一种有效替代目标。

**语义对齐建模**。为了弥合查询与图结构之间的差距，我们使用基于能量的形式对变分分布 $\mathbb{Q}$ 进行建模。其概率与池化后的子图表示 $\mathrm{G_q}$ 和查询嵌入 $\mathbf{q}$ 之间的语义相似度成正比：

$$\mathbb{Q}(\mathbf{q} \mid \mathcal{G}_\mathbf{q}) = \frac{\exp(\mathrm{sim}(\mathbf{G_q}, \mathbf{q})/\tau)}{\Pi(\mathcal{G}_\mathbf{q})}, \tag{C.15}$$

其中，$\tau$ 是温度超参数，$\Pi(\cdot)$ 是配分函数。该公式有效地促进相关子图与查询的嵌入空间对齐。

**可处理的 NCE 近似**。直接计算配分函数 $\Pi(\mathrm{G_q})$ 需要对整个查询空间求和，这在计算上不可行。遵循 InfoNCE 原理 [23, 64]，我们使用一批负样本 B 来近似这一项。因此，学习问题可转化为一个判别任务：从一组干扰项（负样本）中识别真实查询 $\mathbf{q}$：

$$\mathbb{E}\left[\log \mathbb{Q}(\mathbf{q} \mid \mathcal{G}_\mathbf{q})\right] \doteq \mathbb{E}_{\mathcal{B}} \left[ \log \frac{\exp(\mathrm{sim}(\mathbf{G_q}, \mathbf{q})/\tau)}{\sum_{\mathbf{q}' \in \mathcal{B}} \exp(\mathrm{sim}(\mathbf{G_{q'}}, \mathbf{q}')/\tau)} \right] = -\mathcal{L}_{\mathrm{NCE}}. \tag{C.16}$$

证明到此结束。 □

## C.4 命题 4.4 的证明

我们首先重述命题 4.4，以供参考。

命题 4.4 (**上界** $I(\mathcal{G}; \mathcal{G}_\mathbf{q})$) 将子图规模和连通性惩罚项记为：

$$\mathcal{L}_{\mathrm{size}} = \sum\nolimits_{e \in \mathcal{V}} S_e(\mathbf{q}), \quad \mathcal{L}_{\mathrm{con}} = \mathrm{Tr}\left(\mathbf{S}_e(\mathbf{q})^\top \mathbf{L} \mathbf{S}_e(\mathbf{q})\right),$$

其中，$S_e(\mathbf{q}) \in [0, 1]$ 是实体选择概率，$\mathbf{S}_e(\mathbf{q}) \in \mathbb{R}^{|\mathcal{V}|}$ 是软选择向量，$\mathbf{L}$ 是归一化图拉普拉斯矩阵，并且 $\mathrm{Tr}(\cdot)$ *calculates the trace of a matrix. The upper bound is:*

$$I(\mathcal{G}; \mathcal{G}_\mathbf{q}) \leqslant \beta_1 \mathcal{L}_{\mathrm{size}} + \beta_2 \mathcal{L}_{\mathrm{con}},$$

*where* $\beta_1$ *and* $\beta_2$ *are hyper-parameters.*

证明。该压缩项可防止子图选择器检索整个图，并促使其识别出一个“最小”子图。我们证明，最小化所提出的两个结构惩罚项，等价于最小化该互信息的一个上界。

**变分上界推导**。令 $\mathbb{P}(\mathrm{G_q} \mid \mathrm{G})$ 表示编码分布（即子图选择过程）。通过在可能的子图空间上引入任意先验分布 $r(\mathrm{G_q})$，我们构造一个上界：

$$\begin{aligned} I(\mathcal{G}; \mathcal{G}_\mathbf{q}) &= \mathbb{E}_{\mathbb{P}(\mathcal{G})} \left[ \mathrm{KL}(\mathbb{P}(\mathcal{G}_\mathbf{q} \mid \mathcal{G}) \,||\, \mathbb{P}(\mathcal{G}_\mathbf{q})) \right] \\ &\leqslant \mathbb{E}_{\mathbb{P}(\mathcal{G})} \left[ \mathrm{KL}(\mathbb{P}(\mathcal{G}_\mathbf{q} \mid \mathcal{G}) \,||\, r(\mathcal{G}_\mathbf{q})) \right]. \end{aligned} \tag{C.17}$$

最小化 KL 散度会促使学习到的子图分布符合先验 $r(\mathrm{G_q})$ 所规定的属性。

**图结构的吉布斯先验**。为了注入有效推理路径应当稠密且连通这一归纳偏置，我们将先验 $r(\mathrm{G_q})$ 明确定义为吉布斯分布（或玻尔兹曼分布）。能量函数 $E(\mathrm{G_q})$ 对子图配置的代价进行编码：

$$\begin{aligned} r(\mathcal{G}_\mathbf{q}) &= \frac{1}{Z} \exp\left(-E(\mathcal{G}_\mathbf{q})\right), \\ \text{where} \quad E(\mathcal{G}_\mathbf{q}) &= \beta_1 E_{\mathrm{size}} + \beta_2 E_{\mathrm{con}}. \end{aligned} \tag{C.18}$$

常数 $Z$ 表示 Zustandssumme（德语，意为“对状态求和”）。其数学定义是对所有可能状态的能量指数进行归一化求和：

$$Z = \sum_{\mathcal{G}' \in \Omega} \exp(-E(\mathcal{G}')). \tag{C.19}$$

能量项的具体形式源自两个先验：

**(1) 通过拉普拉斯先验实现稀疏性**。假设边选择遵循以零为中心的拉普拉斯分布，其负对数似然对应于选择概率的 $L_1$ 范数，得到 $E_{\mathrm{size}} = \sum S_e(\mathbf{q})$，记为 $\mathrm{L_{size}}$。

**（2）通过 GMRF 先验实现连通性**。假设选择变量在结构上服从高斯马尔可夫随机场（GMRF），则能量由图拉普拉斯算子的二次型定义，得到 $E_{\mathrm{con}} = \mathrm{Tr}(\mathbf{S}_e(\mathbf{q})^\top \mathbf{L} \mathbf{S}_e(\mathbf{q}))$，记为 $\mathrm{L_{con}}$。

**等价于损失最小化**。我们将吉布斯先验代回 KL 散度项。请注意，编码器 $H(\mathrm{G_q} \mid \mathrm{G})$ 的熵由子图选择器的置信度决定。假设选择过程是确定性的或低方差的，则该熵项可视为常数。优化目标可简化如下：

$$\begin{aligned} \min \mathrm{KL}(\mathbb{P}(\cdot \mid \mathcal{G}) \,||\, r(\cdot)) &\Leftrightarrow \min \mathbb{E}_{\mathbb{P}(\mathcal{G}_\mathbf{q} \mid \mathcal{G})} \left[ -\log r(\mathcal{G}_\mathbf{q}) \right] \\ &\Leftrightarrow \min \mathbb{E}_{\mathbb{P}(\mathcal{G}_\mathbf{q} \mid \mathcal{G})} \left[ -\log \left( \frac{1}{Z} \exp(-E(\mathcal{G}_\mathbf{q})) \right) \right] \\ &\Leftrightarrow \min \mathbb{E}_{\mathbb{P}(\mathcal{G}_\mathbf{q} \mid \mathcal{G})} \left[ \beta_1 \mathcal{L}_{\mathrm{size}} + \beta_2 \mathcal{L}_{\mathrm{con}} + \log Z \right] \\ &\Leftrightarrow \min \left( \beta_1 \mathcal{L}_{\mathrm{size}} + \beta_2 \mathcal{L}_{\mathrm{con}} \right). \end{aligned} \tag{C.20}$$

我们结束证明。 □

表 D.1：通用多跳问答数据集的统计数据。

| 数据集 | #问题-文档对 | #文档 | #知识图谱 | # 实体 | # 关系 | # 三元组 |
|---|---|---|---|---|---|---|
| HotpotQA [100] | 20,000 | 204,822 | 20 | 1,930,362 | 967,218 | 6,393,342 |
| MuSiQue [85] | 20,000 | 410,380 | 20 | 1,544,966 | 900,338 | 4,848,715 |
| 2WikiMultiHopQA [26] | 20,000 | 122,108 | 20 | 916,907 | 372,554 | 2,883,006 |
| 总计 | 60,000 | 737,310 | 60 | 4,392,235 | 2,240,110 | 14,125,063 |

表 D.2：特定领域问答数据集的统计数据。

| 数据集 | 领域 | #测试 | #文档 | #实体 | # 关系 | # 三元组 |
|---|---|---|---|---|---|---|
| HotpotQA [100] | 多跳 | 1,000 | 9,221 | 87,768 | 45,112 | 279,112 |
| MuSiQue [85] | 多跳 | 1,000 | 11,656 | 100,853 | 55,944 | 319,618 |
| 2WikiMultiHopQA [26] | 多跳 | 1,000 | 6,119 | 48,779 | 20,748 | 160,950 |
| PubMedQA [36] | 生物医学 | 2,450 | 5,932 | 42,389 | 20,952 | 149,782 |
| DelucionQA [71] | 客户支持 | 184 | 235 | 2,669 | 2,298 | 6,183 |
| TechQA [8] | 客户支持 | 314 | 769 | 10,221 | 4,606 | 57,613 |
| ExpertQA [56] | 客户支持 | 203 | 808 | 11,079 | 6,810 | 16,541 |
| EManual [61] | 客户支持 | 132 | 102 | 695 | 586 | 1,329 |
| MS Marco [62] | 一般知识 | 423 | 3,481 | 24,740 | 17,042 | 63,995 |
| HAGRID [37] | 一般知识 | 1,318 | 1,975 | 23,484 | 18,653 | 48,969 |

# D 实验详情

在本节中，我们提供额外的实验细节。

## D.1 数据集详情

*D.1.1* 通用多跳问答数据集。 遵循先前的工作 [54],, 我们在三**个**用于多跳问答任务的基准数据集上进行实验，分别是 HotpotQA [100],、MuSiQue [85], 和 2WikiMultiHopQA [26]。统计数据列于表 D.1。下面简要介绍各个数据集。

• HotpotQA [100]: 一个大规模多跳问答数据集，包含 97k 个样例。每个问题配有最多两篇金标准支持文档和若干干扰项。回答问题通常需要结合两篇文档中的证据，以检验细粒度推理能力。

• MuSiQue [85]: 包含 2.5 万个经过精心整理的多跳问题，需要进行两到四步推理。这些问题旨在鼓励对多个文档进行组合推理，因此是评估多步语义理解能力的一项颇具挑战性的基准。

• 2WikiMultiHopQA [26]: 一个基于维基百科的多跳问答数据集，包含 19.2 万个问题。每个问题都需要基于两篇或四篇不同的维基百科条目进行推理。该数据集突显了事实内容方面多样的推理模式，为评估提供了全面的测试平台。

*D.1.2* 特定领域问答数据集。为评估 Gfm-Retriever 在不同应用场景中的泛化能力，我们进一步在**七个**涵盖多个领域的多跳问答基准上对其进行评估，包括：**(1) 生物医学：** PubMedQA [36]; **(2) 客户支持：** DelucionQA [71], TechQA [8],ExpertQA [56],和 EManual [61] ; 以及 **(3) 一般知识：** MS Marco [62] 和 HAGRID [37]。数据集的统计数据见表 D.2。

## D.2 基线详情

我们将其与**18 种**最先进的基线方法（及其组合）进行比较，并将它们分为**四**组：**(1) 基础大语言模型：** GPT-4o-mini [31]; **(2) 单步检索增强生成：** 包括 BM25 [70], Contriever [32], GTR[63], ColBERTv2 [72], RAPTOR [73], Proposition [10]; **(3) 图增强检索增强生成：** GraphRAG [12], G-Retriever [24], LightRAG[17], HippoRAG [18], HippoRAG 2 [19], SubgraphRAG [48], PropRAG[88], GFM-RAG [54]; **(4) 多步检索增强生成：** 包括 IRCoT [86], FLARE [35], Adaptive-RAG [33]。下面将作简要介绍。

**（1）基础大语言模型：** 使用未经微调或检索增强的预训练大语言模型。它作为所有基于检索的方法的通用生成器，确保比较公平。我们采用 GPT-4o-mini [31], , 它仅根据参数化知识回答查询。

**（2）单步检索增强生成：** 通过单次检索完成任务，因其效率高且适用性广而得到广泛应用。此类别包括稀疏检索器和稠密检索器:

• BM25 [70]: 一种基于词频、文档长度归一化和逆文档频率的经典概率稀疏检索器，作为强大的零样本基线。

• Contriever [32]: 一种通过对比学习训练的无监督稠密检索器，在没有标注数据的情况下展现出强大的零样本能力和跨域泛化能力。

• GTR [63]: 一种基于 T5 的双编码器，通过结合点积评分扩展模型容量来提升跨域检索能力，并实现较高的数据效率。

• ColBERTv2 [72]: 一种后期交互检索器，利用压缩表示执行词元级匹配，在检索质量和内存效率之间取得平衡。

• RAPTOR [73]: 一种层次化检索方法，将文档组织成递归摘要，从而支持针对复杂查询跨多个抽象层级进行检索。

**表 D.1：基线比较。实心圆（●）、半实心圆（◑）和空心圆（○）分别表示对各项特性的完全支持、部分支持和不支持。Gfm-Retriever 是唯一涵盖所有所列比较维度的方法。**

| 方法 | 检索器设计 | | | | | | 训练与泛化 | | | 推理 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 外部检索 | 密集神经检索器 | 图-增强的检索 | 结构化输出 | 子图输出 | 自适应预算 | 预训练 | 跨领域 | 微调 | 结构-感知提示 | 单步 |
| BM25 | ● | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Contriever | ● | ● | ○ | ○ | ○ | ○ | ● | ◑ | ◑ | ○ | ○ |
| GTR | ● | ● | ○ | ○ | ○ | ○ | ● | ◑ | ● | ○ | ○ |
| ColBERTv2 | ● | ● | ○ | ○ | ○ | ◑ | ○ | ● | ● | ○ | ○ |
| RAPTOR | ● | ● | ◑ | ● | ○ | ◑ | ○ | ○ | ○ | ◑ | ○ |
| Proposition | ● | ● | ● | ● | ● | ○ | ○ | ○ | ○ | ● | ○ |
| GraphRAG | ● | ○ | ● | ● | ○ | ○ | ○ | ◑ | ○ | ● | ● |
| G-Retriever | ● | ● | ● | ● | ● | ◑ | ○ | ◑ | ◑ | ● | ● |
| LightRAG | ● | ● | ● | ● | ◑ | ◑ | ○ | ◑ | ○ | ○ | ● |
| HippoRAG | ● | ◑ | ● | ● | ◑ | ◑ | ○ | ◑ | ○ | ○ | ● |
| HippoRAG 2 | ● | ◑ | ● | ● | ◑ | ◑ | ○ | ◑ | ○ | ○ | ● |
| SubgraphRAG | ● | ● | ● | ● | ● | ○ | ○ | ● | ◑ | ● | ● |
| PropRAG | ● | ◑ | ● | ● | ● | ○ | ○ | ○ | ○ | ○ | ● |
| GFM-RAG | ● | ● | ● | ● | ○ | ○ | ● | ◑ | ● | ○ | ● |
| FLARE | ● | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Adaptive-RAG | ● | ○ | ○ | ○ | ○ | ● | ○ | ○ | ◑ | ○ | ○ |
| IRCoT | ● | ○ | ○ | ○ | ○ | ○ | ○ | ● | ○ | ○ | ○ |
| **Gfm-Retriever** | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● |

• Proposition [10]：一种稠密检索器，使用细粒度命题作为原子单元对文本建立索引，在有限预算下提升检索精确率和下游问答效果。

**(3) 图增强检索增强生成：**以图的形式纳入结构化知识，通过显式建模实体与关系来增强检索和推理，从而实现超越平面文本检索的多跳依赖建模。

• GraphRAG [12]：从文档构建实体图，并利用大语言模型生成社区级摘要，从而实现语料库 - 级问答，超越标准检索。

• G-Retriever [24]：将检索表述为软奖赏收集斯坦纳树问题，以在大型图中识别查询相关子图，在减少幻觉的同时提升推理能力。

• LightRAG [17]：通过双层方案将图感知索引与平面文本检索相结合，高效捕获细粒度和粗粒度语义依赖关系。

• HippoRAG [18]：将知识图谱与个性化PageRank相结合，以支持受海马体记忆启发的高效单步检索，并具有出色的多跳问答性能。

• HippoRAG 2 [19]：通过更深的段落整合和改进的大语言模型交互扩展HippoRAG，增强事实记忆、联想记忆和意义建构记忆。

• SubgraphRAG [48]：利用三元组评分和结构距离编码检索紧凑的知识图谱子图，但依赖固定启发式规则，从而限制了查询自适应推理。

• PropRAG [88]：用富含上下文的命题替换三元组，并在命题路径上执行束搜索，从而改进零样本多跳检索。

• GFM-RAG [54]：采用图基础模型作为大规模异构图上的检索器，但缺乏显式的跨域语义对齐，且在推理时未能充分利用知识图谱推理结构。

**（4）多步检索增强生成：**迭代地交替进行检索与推理，以支持复杂的多跳问题，并逐步构建推理链。这些方法在很大程度上与检索器无关，并可与不同的检索增强生成架构相结合。

• IRCoT [86]：将检索与思维链推理交错进行，其中每一步推理都会引导下一次检索。该设计可减少幻觉并提升多跳问答效果。值得注意的是，IRCoT与检索器无关，可与现有检索增强生成方法结合，以实现多步推理。

• FLARE [35]：在生成过程中执行前瞻式主动检索，通过预判低置信度内容并检索证据加以优化，从而提升长文本生成中的事实一致性。

• Adaptive-RAG [33]：根据预测的查询复杂度，在单步与多步检索策略之间动态选择，从而在多样化查询中平衡效率与推理能力。

## D.3 实验设置详情

在本节中，我们介绍详细的实验设置[2]。

**第5.2节（研究问题*1*）的详细设置。**我们在统一协议下评估实体级和文档级检索。对于输出排序列表的基线方法，使用 Recall@2/5 评估实体和文档的召回率（$R@2/5_E$、$R@2/5_D$）。相比之下，Gfm-Retriever 返回查询条件化子图 $G_q$，并将其节点集 $V_q$作为检索实体

[2]由于采用了标准基准数据集，并使用了相同的基础大语言模型设置，除非另有说明，本论文报告的基线结果均直接取自原始出版物。

，其中检索规模由训练得到的选择器决定，而非显式的 top-$K$ 截断值（我们设置 $K = 5$ 以进行比较）。所有方法均使用相同的实体到文本块倒排索引执行文档检索。对于多步基线方法，我们汇总各步骤中不重复的实体和文档，并在最终排序结果上评估召回率。此设置将检索质量与下游生成解耦，同时保持实体-文档链接的一致性。

**第 5.3 节的详细设置**（研究问题***2***）。我们将各方法检索到的证据输入同一个基础大语言模型 GPT-4o-mini [31]，以评估端到端问答。单步检索增强生成使用按与研究问题***1***相同预算检索到的排名靠前的文档块，而图增强方法则按照各自的原始设计，将检索到的实体或子图序列化。IRCoT 等多步方法交替进行检索与推理，并串联各步骤中的证据。对于 Gfm-Retriever，输入由针对查询构建的最小充分子图组成，其中包含选定的实体、关系路径和关联的文档块。所有方法均采用相同的解码设置，不进行问答微调，并在测试集划分上使用精确匹配、F1 分数、精确率和召回率进行评估。

**第 5.4 节的详细设置**（研究问题***3***）。我们在全部七个跨领域问答数据集上评估检索性能，并在其中提供自由形式答案标注的三个数据集上报告生成式问答结果。对于检索，我们采用文档级 Recall@5（$R@5_D$），该指标体现了实际检索预算，并能更好地衡量领域偏移下的证据覆盖情况。对于生成式问答，我们采用 ROUGE-L，而非精确匹配或 F1 分数，因为跨领域答案通常更长且经过改述，此时序列级重叠能更可靠地反映语义忠实度。

**第 5.5 节（研究问题*4*）的详细设置**。我们通过测量端到端检索延迟以及文档级 Recall@5（$R@5_D$）来评估检索器的效率。延迟从输入查询开始，至文档评分完成为止，不包括大语言模型生成。对于多步方法（例如基于 IRCoT 的方法），延迟包括所有迭代检索步骤；对于基于图的方法，则不计离线索引时间，只统计在线遍历时间。所有结果均取 1,000 次运行的平均值，以确保不同检索范式之间的公平比较。

**第 5.6 节（研究问题*5*）的详细设置**。我们针对 HotpotQA 中的一个多跳问题开展案例研究。该问题询问在电影《Kiss and Tell》中饰演“Corliss Archer”的女演员担任过什么政府职位。相关证据分散在多份文档中，包括电影《Kiss and Tell (1945)》、女演员秀兰·邓波儿，以及介绍她后来担任公职的传记描述。我们将 Gfm-Retriever 应用于该实例，并将子图选择器生成的查询条件化子图可视化。可视化展示了选中的实体、关系和明确的推理路径，并将它们与原始上下文中语义相关但与查询无关的实体进行对照。该案例有助于定性评估检索子图是否足够精简且能充分支持多跳推理，完整元数据见图 D.1。

**第5.7节的详细设置**（研究问题***6***）。为进行消融研究，以分析 Gfm-Retriever 中核心组件的贡献，我们在保持其他设置不变的情况下，通过移除特定模块构建了***五个***消融变体（及其组合）。这些变体如下：

• **Gfm-Retriever**（无原型）：移除公式 (7) 中基于原型的语义对齐损失 Lproto，从而在检索器训练期间禁用跨域原型监督。

• **Gfm-Retriever(*w/o IGC*):** 移除式（8）中的信息增益对比损失 LIGC，避免检索器受到正则化而趋向信息性语义原型。

• **Gfm-Retriever (*w/o Tune*):** 禁用式（18）定义的微调阶段，该阶段会使检索器适应查询特定子图选择。

• **Gfm-Retriever (*w/o SubG*):** 移除第4.2节中的整个子图选择器模块，使模型退化为实体级检索，且不再学习最小充分子图。

• **Gfm-Retriever (*w/o Prompt*):** 移除第4.3节所述的上下文内关系提示策略。仍会使用检索实体和文档，但不会将关系路径显式重组为面向大语言模型的结构化提示。

所有变体均在相同设置下进行评估。对于超参数敏感性分析，我们针对***五***个关键加权超参数进行敏感性分析：

• $\alpha_1$：控制式（7）中二元交叉熵检索损失 LBCE 的权重，有助于预测查询-实体对的相关性。

• $\alpha_2$：控制式（7）中基于排序的检索损失 Lrank 的贡献，通过在相关实体与不相关实体之间施加间隔来提升检索鲁棒性。

• $\alpha_3$：为式（7）中的基于原型的语义对齐损失 Lproto 加权，以调节训练期间跨域语义一致性的强度。

• $\beta_1$：控制极小性正则化项，该项会惩罚式（18）中所选子图的大小。

• $\beta_2$：控制连通性正则化项，该项旨在增强式（18）中所选子图的结构连贯性。

每次独立调整一个超参数，并使用验证集上的 MRR 报告性能。

## E 实现细节

本节介绍实现细节。

## E.1 Gfm-Retriever 实现细节

*E.1.1* 预训练数据。我们遵循 Luo 等人 [54] 发布的预训练数据处理流程，准备大规模预训练数据。具体而言，我们从 HotpotQA、MuSiQue 和 2WikiMultiHopQA 的训练划分中采样了共 60K 个预训练实例，并合并其关联的候选段落以构建文档语料库。为构建知识图谱，我们采用基于开放信息抽取的抽取策略，使用 GPT-4o-mini 和 Gutiérrez 等人 [18] 所述的提示，从语料库中抽取实体、关系和三元组。随后，通过计算成对嵌入相似度 $s(e_i, e_j)$，使用 ColBERTv2 [72] 进行实体消歧：

$$s(e_i, e_j) = \textsc{TextEmb}(e_i)^\top \textsc{TextEmb}(e_j), \tag{E.1}$$

当它们的嵌入相似度超过阈值 $\tau = 0.8$ 时，会添加新的等价关系（$e_i, \equiv, e_j$）。为控制每个索引的规模，采样数据被划分为多个子集，每个子集约包含 1K 个问题和 10K 篇文档。由此得到 60 个知识图谱索引及其对应的问题-文档集，用于大规模预训练。

**表 E.1：超参数配置。**

| 阶段 | 超参数 | 值 |
|---|---|---|
| 知识图谱索引构建 | 开放信息抽取模型 | GPT-4o-mini |
| | 实体解析模型 | ColBERTv2 |
| | 文本嵌入维度 | 768 |
| | 相似度阈值 $\tau$ | 0.8 |
| GFM 架构 | 层数 | 6 |
| | 隐藏维度 | 512 |
| | 句子编码器 | all-mpnet-v2 |
| | 消息函数 Msg(·) | DistMult(·) |
| | 聚合函数 Agg(·) | 求和(·) |
| | 关系更新 $g(\cdot)$ | 两层多层感知机 |
| 预训练（第一阶段） | 损失权重 $\alpha_1$ | 0.3 |
| | 优化器 | AdamW |
| | 学习率 | 5e-4 |
| | 批量大小 | 4 |
| | 训练步数 | 30,000 |
| | 负样本数量 | 128 |
| 预训练（第二阶段） | 损失权重 $\alpha_2$ | 0.1 |
| | 损失权重 $\alpha_3$ | 0.1 |
| | 温度 $\tau$ | 0.1 |
| | 优化器 | AdamW |
| | 学习率 | 5e-4 |
| | 批量大小 | 4 |
| | 持续训练轮数 | 20 |
| 微调 | 文本嵌入维度 | 768 |
| | 隐藏维度 | 1024 |
| | 查询提示维度 | 512 |
| | 子图选择器维度 | 256 |
| | 正则化权重 $\beta_1$ | 0.01 |
| | 正则化权重 $\beta_2$ | 0.1 |
| | Gumbel-Sigmoid 温度 $\tau$ | 0.5 |
| | InfoNCE 温度 $\tau$ | 0.07 |
| | 优化器 | AdamW |
| | 学习率 | 5e-4 |
| | 批量大小 | 8 |
| | 微调轮数 | 20 |
| 推理（问答） | 前-$K$ 个文档的数量 | 5 |
| | 最大 DFS 深度 $l$ | 3 |
| | 随机种子 | 1024 |

*E.1.2* 模型配置。Gfm-Retriever 被实现为一个查询条件化图神经网络，包含 6 个消息传递层，隐藏维度为 512。实体和关系嵌入由冻结的 all-mpnet-v2 句子编码器 [74] 初始化，维度为 768。每层使用双线性交互函数执行关系感知消息传递，然后进行求和聚合；同时，通过轻量级的两层多层感知机对关系状态进行优化，以捕获依赖上下文的变换。预训练后的 Gfm-Retriever 包含约 8.1M 个可训练参数，可作为通用的跨域检索器，为子图选择生成查询条件化实体嵌入。对于文档排序，我们返回排名前 5 的文档语料库。

*E.1.3* 超参数。我们在表 E.1 中提供了预训练、微调和推理阶段的超参数配置。

## E.2 基线实现细节

我们提供的基线方法实现如下。

• GPT-4o-mini：https://platform.openai.com。• BM25：https://github.com/jxmorris12/bm25_pt。• Contriever：https://github.com/facebookresearch/contriever。• GTR：https://github.com/google-research/t5x_retrieval#dense-retrieval-models。

• ColBERTv2：https://github.com/stanford-futuredata/ColBERT。• RAPTOR：https://github.com/parthsarthi03/raptor。• Proposition：https://github.com/chentong0/factoid-wiki。• GraphRAG：https://github.com/microsoft/graphrag。• G-Retriever：https://github.com/XiaoxinHe/G-Retriever。• LightRAG：https://github.com/HKUDS/LightRAG。• HippoRAG：https://github.com/OSU-NLP-Group/HippoRAG。• HippoRAG 2：https://github.com/OSU-NLP-Group/HippoRAG。• SubgraphRAG：https://github.com/Graph-COM/SubgraphRAG。• PropRAG：https://github.com/ReLink-Inc/PropRAG。• GFM-RAG：https://github.com/RManLuo/gfm-rag。• IRCoT：https://github.com/stonybrooknlp/ircot。• FLARE：https://github.com/jzbjyb/FLARE。• Adaptive-RAG：https://github.com/starsuzi/Adaptive-RAG。

对于所有基线，若使用相同的标准基准数据集和相同的基础大语言模型设置，则本论文报告的结果直接取自原始出版物。当此类设置不可用时，我们采用原始出版物中建议的或官方实现提供的超参数配置，以确保基线性能合理。为保持公平且受控的比较，我们将隐藏维度、层数和架构设置等其他因素与所提出框架中使用的设置保持一致。

## E.3 硬件与软件配置

我们在基于 Linux 的服务器环境中进行实验，硬件和软件配置如下。

• **操作系统：** Ubuntu 20.04 LTS。• **CPU：** 10 颗 Intel(R) Xeon(R) Platinum 8358 CPU @ 2.60GHz，配备 1TB DDR4 内存。• **GPU：** 8 块 NVIDIA Tesla A100 SMX4，配备 80GB 内存。• **软件：** CUDA 12.4 、 Python 3.8.12 、 PyTorch3 1.9.1、PyTorch Geometric4 2.0.1。

# F 其他结果

## F.1 领域专属微调的影响

为进一步考察和比较检索器的跨域适应能力，我们在最相关的基线模型 GFM-RAG 和 Gfm-Retriever 上，使用各自对应目标数据集的训练集进行微调，并在第 5.4 节（研究问题***3***）中使用的相同 **七个** 跨领域基准测试上进行评估。表 F.1 报告了使用 $R@5_D$ 衡量的文档检索性能，表 F.2 报告了三个支持自由形式答案的数据集上的生成式问答性能（采用 ROUGE-L 衡量）。标有**“*”**的结果表示领域特定微调后的性能，未标注的结果则对应与研究问题***3***一致的零样本设置。

[3] https://github.com/pytorch/pytorch. [4] https://github.com/pyg-team/pytorch_geometric.

**表 F.1：领域特定微调对跨领域文档检索的影响（R@$5_D$ %）。带“*”的结果表示微调后的性能。**

| 数据集 | PropRAG | HippoRAG 2 | GFM-RAG | | Gfm-Retriever | |
|---|---|---|---|---|---|---|
| PubMedQA | 55.2 | 60.1 | 58.5 | 63.6* | 61.9 | **64.2*** |
| DelucionQA | 65.9 | 68.2 | 70.8 | 82.7* | 72.3 | **84.3*** |
| TechQA | 44.1 | 47.5 | 46.6 | 49.5* | 50.3 | **51.6*** |
| ExpertQA | 60.9 | 63.5 | 62.7 | 60.8* | 65.3 | **65.9*** |
| EManual | 58.3 | 56.2 | 60.6 | **75.9*** | 61.9 | 71.7* |
| MS Marco | 65.2 | 68.4 | 71.0 | 77.5* | 71.6 | **78.0*** |
| HAGRID | 78.8 | 81.6 | 84.7 | 86.6* | 86.2 | **87.5*** |

**表 F.2：领域特定微调对跨领域生成式问答的影响（ROUGE-L %）。带 * 的结果表示微调性能。**

| 数据集 | PropRAG | HippoRAG 2 | GFM-RAG | | Gfm-Retriever | |
|---|---|---|---|---|---|---|
| ExpertQA | 27.1 | 29.0 | 27.6 | 27.3* | 31.9 | **32.3*** |
| Emanual | 43.3 | 46.1 | 47.6 | 50.9* | 49.8 | **51.6*** |
| MS Marco | 33.8 | 33.1 | 30.2 | 36.5* | 35.6 | **37.1*** |

结果表明，领域特定微调显著提升了 GFM-RAG 和 Gfm-Retriever 在检索和生成式问答方面的表现，说明两者均能从域内监督中获益。然而，在这两种设置下，Gfm-Retriever 始终优于 GFM-RAG。具体而言，在跨域文档检索方面，Gfm-Retriever 在微调前便优于 GFM-RAG，微调后仍保持这一优势。生成式问答也呈现相同趋势：无论是在零样本设置下还是经过微调后，Gfm-Retriever 的 ROUGE-L 均更高。总体而言，结果表明，Gfm-Retriever 不仅在跨域检索方面具有更强的泛化能力，而且在领域特定微调下也始终能提供更好的适应起点。

*F.1.1* 不同基础大语言模型的影响。为评估 Gfm-Retriever 相对于底层生成器的模块化能力，我们比较了它与不同基础大语言模型搭配时的问答性能。由于我们的检索器与大语言模型无关，因此固定检索到的子图，仅替换生成大语言模型。我们测试了**六种**具有不同能力和成本的代表性大语言模型：GPT-4o-mini、GPT-5-mini、Gemini-2.5-Flash、Qwen3-Max、Grok-4-Fast 和 DeepSeek-V3.2。我们报告了三**个**多跳问答基准测试上的精确匹配（EM %），并分别评估是否集成多步推理框架 IRCoT。此外，我们还比较了输入和输出令牌的单令牌货币成本。

图 F.1 显示，更强大的大语言模型能够持续提升答案生成效果，而检索组件保持不变。对于全部三个数据集，随着模型能力增强，精确匹配率稳步提高，这表明更优秀的生成器能够更有效地利用 Gfm-Retriever 提供的结构化证据。与 IRCoT 结合时，大语言模型之间的性能差距进一步扩大，说明多步推理从更强大的基础大语言模型中获益更多。值得注意的是，在结构化子图证据的引导下，即使是轻量级模型也能取得有竞争力的表现，这凸显了检索阶段的鲁棒性。成本分析揭示了明显的质量与成本权衡：虽然更大型的模型（例如 GPT-4o-mini）具有良好的效率与性能平衡，但 Qwen3-Max 等模型的准确率最高。总体而言，结果证实 Gfm-Retriever 可与各种大语言模型骨干网络灵活搭配，而且生成质量的提升能够直接转化为更出色的端到端问答效果，无需修改检索器。

*F.1.2* 预测推理跳数的分布。为了进一步检验检索到的结构是否反映了问题内在的多跳复杂度，我们分析了不同方法所隐含的推理跳数分布。对于 HotpotQA、MuSiQue 和 2WikiMultiHopQA 测试集中的每个问题，我们根据标注的支持事实估算真实推理链的跳数。随后，我们计算不同检索方法所导出的推理路径的跳数，并将其分布与真实值进行比较。

图 F.2 显示，与 HippoRAG 和 GFM-RAG 相比，Gfm-Retriever 的表现始终更接近真实值，这一点体现在它在全部三个数据集上的平均绝对误差（MAE）最低。相比之下，HippoRAG 明显倾向于生成更长的推理链，尤其是在 2WikiMultiHopQA 上，它会高估高跳数路径，并且与真实分布存在显著偏差。尽管 GFM-RAG 更好地追踪了总体趋势，但其跳数分布仍不如 Gfm-Retriever 与真实分布吻合。结果表明，Gfm-Retriever 不仅能检索相关证据，还能捕捉各个查询所需的充分结构深度。它并非依赖浅层或深层推理路径，而是倾向于恢复符合推理需求的子图，这进一步说明该子图选择器能够有效地建模最小但充分的多跳结构。

## G 讨论

在本节中，我们将解答这些**“关键问题”**，以深入探讨其动机、理论和设计选择。

***Q1.*** 为什么 ***GFM-Retriever*** 可以被视为基础模型，而非特定任务检索器？Gfm-Retriever 在从多领域语料库构建的大规模异构知识图谱上进行预训练，而非针对单一的下游任务进行训练。它通过知识图谱补全学习可泛化的结构推理模式，随后迁移到结构各异的任务中，包括子图选择和领域转移情况下的文档检索。命题 4.1 进一步表明，查询条件化 GFM 能够表达多领域知识图谱上的分级模态逻辑，从而在形式上证明，该模型捕捉的是领域感知的多跳推理规则，而非数据集特定启发式。

***Q2.*** 检索器跨领域的可迁移性源自何处？可迁移性主要源自原型驱动的连续预训练（第二阶段）。标准图神经网络通常会过拟合领域特定结构模式。相比之下，第二阶段训练会显式地使实体嵌入与领域语义原型对齐（$\mathbf{c}_d$，见公式（7）），从而对检索器进行正则化，使其学习通用领域知识，同时区分领域不变的语义差异。IGC 正则化器进一步抑制虚假相关性。这确保消息传递机制能够捕捉基本推理模式（例如传递性、组合性），无论具体领域如何，这些模式都成立。

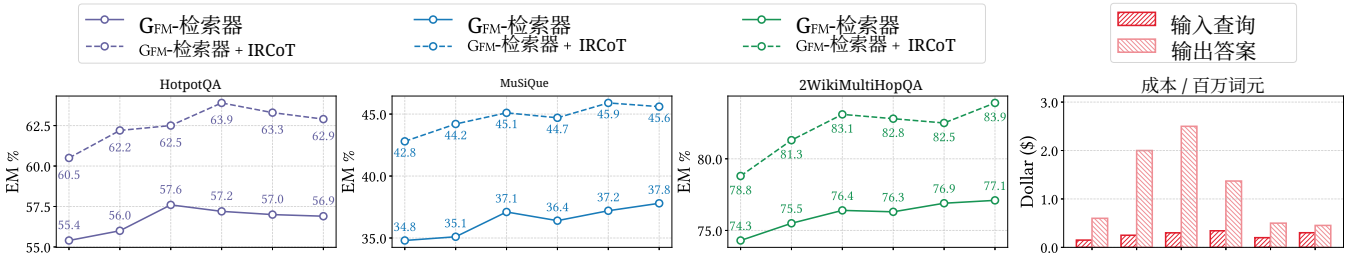


**图 F.1：Gfm-Retriever 在不同基础大语言模型上的问答性能。**

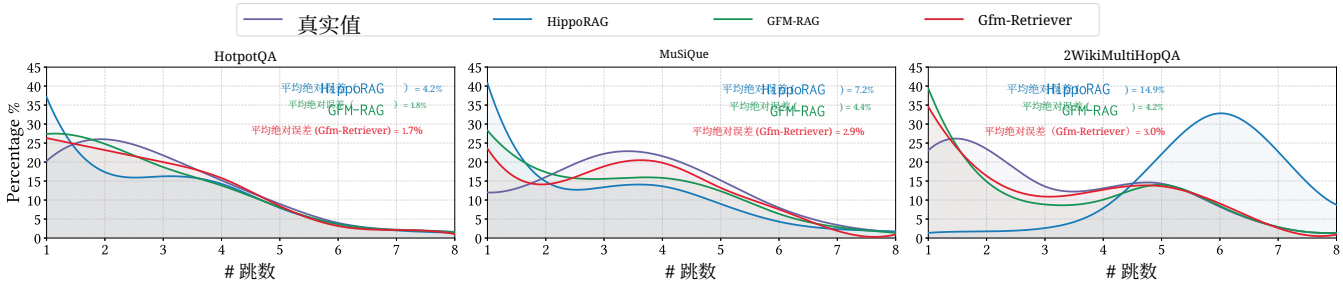


**图 F.2：预测推理跳数的分布。**

***Q3.*** 为什么信息瓶颈（***IB***）公式被称为“无标签”？“无标签”是指在检索优化阶段无法获得真实答案 $y$。因此，我们提出一种替代目标，用查询 $\mathbf{q}$ 替换 $y$，转而优化 $I(\mathbf{q}; G_{\mathbf{q}})$。命题 4.2 从理论上保证了这种近似引入的误差受条件熵 $H(\mathbf{q} \mid y)$ 所界定，因此我们只需使用查询-数据对即可优化子图选择器。

***Q4.*** 为什么采用参数高效微调（***PEFT***），而不是全量微调？ **(1)** 保留泛化能力：在特定数据集上进行全量微调存在灾难性遗忘的风险。通过冻结图基础模型骨干网络，仅训练轻量级投影矩阵和子图选择器，我们能够保留跨领域推理能力，同时适应检索任务的特定格式。**(2)** 任务差异管理：预训练任务（知识图谱补全）与下游任务（子图选择）并不相同。参数高效微调起到桥梁作用，在不扭曲用于编码实体的语义空间的前提下优化决策边界。

***Q5.*** 理论上的性能瓶颈是什么，又如何缓解？ 理论上，瓶颈在于推理，具体来说是路径提取过程。虽然子图选择效率很高（$O(|V^T|)$），但通过深度优先搜索枚举推理路径的最坏情况复杂度呈指数级，为 $O(\Delta^l)$，其中 $\Delta$ 是节点度数，$l$ 是其深度。我们通过实现带有严格跳数限制（$l \leqslant 3$）的截断深度优先搜索，并采用束搜索评分来解决这一问题。

***Q6.*** 性能提升是源于子图选择器还是上下文提示器？两个组件都至关重要。消融研究（图 6）区分了它们各自的贡献。移除子图选择器（恢复为标准检索）会导致检索性能大幅下降，这证明内容质量主要由检索器的选择能力决定。移除上下文提示器会导致问答性能明显下降，但对检索几乎没有影响。这表明，选择器负责寻找“原料”，而提示器则提供“配方”，通过明确组织结构来引导大语言模型的推理过程。

***Q7. GFM-Retriever*** 与相关工作有何系统性差异？在遵循标准评估协议（数据集和指标）的同时，Gfm-Retriever 在整体范式上存在根本差异。我们在表 D.1 中展示了二者的比较。

***Q8.*** 构建知识图谱索引是否会为现实应用引入高昂的成本？尽管知识图谱构建会产生初始开销，但这是一项一次性且离线的投入。完成索引后，该图能够以高效率支持无限次查询。由于知识图谱构建并非我们的主要贡献，且我们遵循 Gutiérrez 等人 [18], 、Li 等人 [48] 和 Luo 等人 [54], 所采用的相同实验协议，因此我们直接在他们发布的知识图谱上进行评估，从而确保比较的公平性。

# H 其他相关工作

## H.1 检索增强生成

检索增强生成（RAG）是一种通用范式，通过在推理过程中显式访问外部知识来增强生成式大模型。RAG 系统不再仅仅依赖存储在模型参数中的信息，而是根据输入检索相关证据，并将其融入

生成过程。这种设计已在各种现实应用中展现出显著成效，包括金融分析 [89], 医疗保健系统 [95], 法律推理任务 [91], 和教育 [58]。随着大语言模型的快速发展，检索增强变得愈发重要，因为纯参数化模型往往会产生幻觉输出、包含过时知识，且适应能力有限。先前的综述指出，以检索到的证据为生成内容提供依据，能够显著提升忠实性、降低隐私风险，并增强模型在分布偏移下的鲁棒性 [28, 114]。因此，RAG 已成为一种广泛采用的策略，无需对模型进行全面重新训练，即可扩展大语言模型的知识容量。

大多数检索增强生成方法都采用“先检索、后生成”的范式。随后，通过拼接、基于注意力的融合或编码器—解码器条件控制 [13]，将检索到的文本单元作为额外上下文融入生成器。与基于关键词的检索相比，稠密检索器能够实现更高的语义召回率，并与下游生成更好地对齐。在此框架基础上，后续的 RAG变体通过在不同粒度上运行，探索更紧密的检索—生成耦合，包括句子级 [20, 38], 文本块级 [44, 107], 和词元级 [5, 96], , 从而在解码过程中实现动态证据整合。其他方法则进一步约束生成，以提升依据关联和可验证性。尽管架构各异，检索增强生成系统的共同原则是将检索视为模块化且可更新的组件，使外部知识访问与生成器的参数化能力相解耦。

然而，现有的大多数 RAG 都检索非结构化文本，并将证据视为无序的上下文集合。尽管这种形式对许多知识密集型任务都很有效，但它在很大程度上忽略了检索项之间的关系依赖和高阶结构，因此催生了 GraphRAG 等基于图的扩展。

## H.2 基于图的检索增强生成

图为结构化关系提供了富有表现力的表示方式，已广泛应用于知识表示、社交网络分析和生物医学建模 [55, 92]。受纯文本检索局限性的启发，基于图的检索增强生成（GraphRAG）应运而生，作为检索增强生成的一种扩展，将图结构化知识纳入检索过程 [109]。在 GraphRAG 中，概念通常表示为节点，而边则编码语义、逻辑或其他关系，使检索能够利用显式结构，而非仅依赖非结构化文本相似度 [12, 66, 115]。近期研究探索了这一范式在知识图谱和分子图等结构化图场景中的应用 [21, 66], , 展现了其在关系感知检索与推理方面的潜力。

除了从现有图中检索，越来越多的研究致力于从非结构化文本中构建图。通过将文本信息组织成图，GraphRAG 能够结构化地访问隐式关系知识，从而有益于文档摘要 [12], 规划 [49], 和多步推理 [22]等任务。在这一研究方向中，图充当中间表示，有助于抽象、聚合和基于遍历的证据选择，并对稠密文本检索形成补充。现有方法可归纳为若干反复出现的设计模式。层次化图构建方法通过树结构或社区级抽象来组织知识，从而支持在大型语料库上进行由粗到细的检索，RAPTOR [73] 和微软的 GraphRAG [12]便是其中的代表。神经图检索方法将图神经编码器与专门的目标函数相结合，以支持多跳推理和可扩展检索，其中包括 G-Retriever [24], 、LightRAG [17], 和 GFM-RAG [54]。其他方法则强调与大语言模型紧密结合的动态图构建和自适应遍历，使检索策略和图结构能够在推理过程中不断演变，例如 DALK [45], 、KGP [90], 和 ToG [77]。此外，HippoRAG [18] 采用基于个性化PageRank的受记忆启发的形式，实现高效的单步多跳检索，表明基于图的传播机制能够显著提升推理效率。

尽管近期取得了进展，许多 GraphRAG 方法仍在过于稠密的图上进行检索，往往引入冗余结构，并产生无法满足任务需求的上下文。此外，依赖固定的图构建方式限制了其在不同任务间的泛化能力，凸显出一个尚待解决的挑战：如何识别紧凑且与任务相关的子图，同时保留关键信息关系。

## H.3 图基础模型

图基础模型旨在通过自监督预训练 [52, 93]从大规模图数据中学习可迁移的表示。通过利用属性扰动、结构重建和对比学习等目标，这些模型力求捕捉可泛化的结构与语义模式，以便在不同任务间复用。众多研究探索了这一范式，包括 GCC [69], GPT-GNN [27], GPPT [78], L2P-GNN [53], 以及后续变体 [7, 9, 34, 94, 99, 103]。预训练后，通常通过微调或提示来适配所学表示。然而，现有方法大多隐含地假设预训练图与下游图之间的分布差异有限，往往侧重于同质子图 [29]或范围狭窄的领域 [112]。因此，在领域偏移情况下，其可迁移性会下降。

为缓解这一问题，近期研究开始探索跨域和多领域 GFM，旨在缩小分布差异并提升泛化能力。一类研究致力于提取领域不变表示，以支持在差异显著的领域间迁移，例如 GCOPE [113], OMOG [50], PGPRec [102], UniAug [82], 和 UniGraph [25]。另一研究方向则借助语言模型，通过将图映射至文本空间来促进跨域对齐，例如 OFA [50] 和 HiGPT [81], ，但此类方法仅适用于文本属性图 [80, 87]。此外，还有一些策略引入结构或词元级机制来连接不同领域，包括 GCOPE [113]中的虚拟节点和 MDGPT [106], 中的领域令牌，但这些方法忽略了异质领域间的语义一致性。

已有一些尝试将图基础模型用作 GraphRAG 中的检索器，例如 GFM-RAG [54]。然而，GFM预训练如同一个语义黑盒，无法在不同领域之间显式对齐语义，导致检索依赖隐式的表示相似性，并限制了模型在领域偏移下的鲁棒性。

## H.4 关键子图挖掘

图结构蕴含丰富的关系语义，但其中真正对下游任务起决定作用的只占很小一部分。尽管图结构长期以来一直被视为图表征学习的基础 [2, 98], ，识别任务相关且数据依赖子图的系统性方法仍未得到充分探索。

早期研究主要依赖预定义的结构分解，其中最具代表性的是图核 [41],，它们提取特定模式，例如随机游走 [15], 、最短路径 [6], 或有根子树 [75]。这些方法具有较强的归纳偏置，在某些领域能够取得有竞争力的结果。然而，核函数编码的结构模式通常由人工设计且固定不变，因此当任务需求或图语义发生变化时，灵活性有限，泛化能力也较差。另一个重要方向则关注反复出现的小型局部模式，尤其是模体，以保留细粒度的结构属性 [42, 59, 67]。NEST [98] 进一步结合多个模体，以捕获更丰富的子图级交互。尽管模体方法效果显著，但其根本上依赖于对小型子结构进行穷举，这既需要领域专业知识，也限制了其扩展到更大规模或不规则模式的能力。

为突破僵化的邻域限制，多项研究探索了更灵活的子图建模策略 [43, 46, 47, 79]。SGN [97]通过人工指定的子图选择规则扩展结构特征空间，但随着子图规模增大，这类启发式方法所提供的信息会越来越有限。NCAT [116] 为组合邻居模式分配可学习权重，但仍受固定规模约束。从信息论视角出发，SIB [104, 105] 学习边级分配以识别瓶颈子图，但并未显式保留子图语义。CAL [76]则进一步探索因果子图发现，通过基于注意力的估计将因果信号与捷径相关性区分开来。

子图很少用于检索。SubgraphRAG [48] 在规模预算下通过三元组级评分检索以查询为中心的子图，但其关注的是覆盖度，而非特定任务所需的充分性，因此检索到的结构并非最小结构，关键子图识别问题也仍未得到解决。

## I 局限性

尽管所提出的Gfm-Retriever在跨域检索和多跳推理方面表现优异，我们仍承认若干局限性，这些局限性界定了当前工作的范围，并指明了未来的研究方向。

**对知识图谱构建的依赖**。我们的框架主要聚焦于检索与推理阶段，其前提是已预先构建知识图谱。尽管我们利用由大语言模型驱动的先进开放信息抽取技术进行索引，检索性能的上界仍自然受限于这一上游知识图谱构建的完整性。在原始语料库噪声极大，或抽取的三元组存在严重稀疏性的场景中，GFM的语义表达能力在理论上可能受底层结构质量的限制。

**路径提取深度的权衡**。 为确保实时推理效率，我们采用截断式深度优先搜索（DFS）策略，并预先设定跳数限制（$l \leqslant 3$），以从所选子图中提取推理路径。尽管我们的实证分析（附录 B.3）证实，这一设置涵盖了大多数标准基准测试中的推理复杂度，但在处理需要极长距离依赖关系的超复杂查询时，可能会受到限制。如何扩展路径提取机制以支持更深层的遍历，同时避免计算开销呈指数级增长，仍是未来优化面临的一项挑战。

**对知识更新的适应性**。Gfm-Retriever采用“先预训练后微调”范式，该范式假定推理阶段的知识快照相对静态。尽管图基础模型在跨领域上展现出较强的泛化能力，但在不进行完整重新索引的情况下，使已索引的知识图谱适应实时数据流或高度时序事实仍构成实际挑战。我们展望未来工作引入持续学习，以在时间敏感应用中解决这一局限性。

## J 未来方向

基于已建立的结构化检索范式，我们展望了若干有望进一步推进图基础模型与大语言模型协同发展的方向：

**动态与持续式图检索增强生成**。现实世界的知识本质上具有时间性且不断演变。将持续图学习机制集成到图基础模型骨干网络中，是本工作的自然延伸。未来研究可探索高效的归纳式更新，使检索器能够从流式数据中动态整合新的实体和关系，而无需全面重新索引。这样，系统便能处理对时效性敏感的查询，适用于突发新闻分析或金融监测等场景。

**通过图指令微调实现对齐**。 尽管目前提出的Gfm-Retriever 通过上下文提示连接图与大语言模型，但更紧密的集成有望进一步提升性能。我们计划探索参数高效的联合训练策略，使图基础模型的表示空间与大语言模型直接对齐。通过将子图检索视为统一指令微调流程中的一个潜在步骤，我们旨在构建能够原生理解复杂结构化指令的集成架构。

**图基础模型驱动的自主智能体**。除了问答任务之外，检索最小且充分的子图，也是自主智能体在复杂环境中导航的一项关键能力。我们计划将 Gfm-Retriever 改造为智能体的“认知地图”模块，其中检索子图表示状态空间或动作约束。这有望使智能体能够在知识密集型环境中进行多步规划和推理，例如代码库导航或科学发现。

**元数据（Python 字典）**

```
{
  "_id": 5a8c7595554299585d9e36b6 ,
  "答案":     "礼宾司司长"          ,
  "问题":       “在电影《Kiss
  ↪  与“讲述”？ ,
  “支持_事实”: [
    [《亲吻与告白》（1945年电影）”?],
    [“秀兰·邓波儿”,      0],
    [“秀兰      寺庙",    1],
  ],
  \"context\": [
    [
      \"《科莉丝之吻》\"    ,
      [
        \"《科莉丝之吻》是一部1949年美国喜剧电影，由理查德·华莱士执导，霍华德·
        ↪      丁斯代尔编剧。\"  ,
        \"该片由秀兰·邓波儿主演，这是她最后一次担任主演，也是她最后一次出演电影。\",
        这是1945年的电影的续集                《Kiss and Tell》。,
        《科莉丝之吻》              在上映前更名为《几乎成为新娘》，这个片名也出现在
        ↪  片头字幕中。       ,
        这部电影于11月25日上映,                    1949年，由联美影业发行。”,
      ],
    ],
    [
      “财政大臣”,
      [
        “财政大臣一职曾是英格兰政府的职位，后来成为
        ↪  自1707年联合法案以来的英国政府职位。”,
        “该职位的任职者将是排名第三高的国家重臣,                                    低于大
        ↪  总管和大法官。",
      ],
    ],
    [
      "《遇见科莉丝·阿彻》（电视剧）",
      [
        "《遇见科莉丝·阿彻》是一部在哥伦比亚广播公司播出的美国电视情景喜剧（7月13日,          1951年 - 8月10日,
        ↪  1951年），并通过Ziv公司以辛迪加形式于1954年4月至12月播出。",
        “该节目改编自同名广播系列剧，该系列剧则改编自一系列
        ↪  F. Hugh Herbert 创作的短篇小说。”,
      ],
    ],
    [
      “乡村会计”,
      [
        “乡村会计（也称为                    “Patwari”，“Talati”，"Patel"，"Karnam"，"Adhikari",
        ↪  "Shanbogaru","Patnaik"  等）是印度次大陆农村地区的一种政府行政职位
        ↪  印度次大陆。'       ,
        ' 该职位及其任职者称为                    "特伦甘纳邦的 Patwari，孟加拉、印度北部和
        ↪  巴基斯坦；而在信德省则称为“土地登记员”。'       ,
        ' 该职位称为                  安得拉邦的“karnam”       ，奥里萨邦的“patnaik”或泰米尔纳德邦的“adhikari”
        ↪  ，而通常称为                    卡纳塔克邦的“talati”，古吉拉特邦和马哈拉施特拉邦。”
        “该职位被称为                  “北卡纳塔克邦和马哈拉施特拉邦的‘库尔卡尼’。”       ,
        “该职位被称为                  “南卡纳塔克邦的‘shanbogaru’。”
      ],
    ],
    [
      “Joseph Kalite”  ,
      [
        “Joseph Kalite（2014年1月24日去世）是一名中非政治家。”,
        “作为政府部长，他负责住房事务或卫生事务。”
        “据报道，穆斯林 Kalite 在首都中央清真寺外被反巴拉卡杀害
        ↪  班吉，当时正值中非共和国冲突。”
        “临时总统 Catherine Samba-Panza 上台当天，他在班吉被砍刀杀害,
        ↪  掌权 .",
        “袭击发生时，Kalite 没有担任任何政府职位，在塞雷卡统治时期也没有。”
        “据报道，他支持塞雷卡领导人 Michel Djotodia 的统治。”,
      ],
    ],
```

```
    [
      “Charles Craft” ,
      [
        “Charles Craft（5月9日， 1902年－9月19日， 1968）是一位英裔美国人电影和电视
        ↪ 剪辑师。”,
        “1902年5月9日出生于英格兰汉普郡, 1902年, Craft 于 进入电影业
        ↪ 1927 年进军好莱坞。”,
        他剪辑的第一部电影是环球影业的无声电影 《Painting the Town》。”,
        “在接下来的25年里, Craft 将剪辑90部长篇电影。”,
        “20世纪50年代初，他将工作重心转向小荧屏, 他的首部剧集是 《Racket
        ↪ Squad》, 从1951—1953年, 他担任主要剪辑师, 剪辑了98集中的93集。',
        '他在20世纪50年代还参与了其他几部电视剧, 包括 《遇见科莉丝·阿彻》 (1954年),
        ↪ 《科幻剧场》 (1955—1956年), 和《公路巡警》 (1955—1957年)。',
        ' 20世纪50年代末和20世纪60年代初，他是……的主要剪辑师之一 《海底追踪》，由劳埃德
        ↪ ·布里奇斯主演，剪辑了半数以上的剧集。',
        ' 他最后一部电影作品是剪辑 《海豚弗利珀》的新冒险》 (1964年, 1963年电影的续集
        ↪ 《海豚弗利珀》。' ,
        “当这部电影被拍成电视剧时, Craft 开始负责该剧的剪辑工作,
        ↪ 在 1966年退休前，他剪辑了前 28集，”,
        “Craft 于 1968年9月19日去世, 地点在洛杉矶, 加利福尼亚州。",
      ],
    ],
    [
      《遇见科莉丝·阿彻》 ,
      [
        《遇见科莉丝·阿彻》，这是一档广播黄金时代的节目, 于1月7日开播, 播出至9月30日,
        ↪ 1956年。" ,
        虽然这是哥伦比亚广播公司为应对全国广播公司热门节目《与朱迪约会》而推出的节目, 它也曾由全国广播公司播出,
        ↪ 于1948年作为《鲍勃·霍普秀》的暑期替代节目播出。 ,
        “从10月3日 1952年到6月26日，1953年, 该节目曾在美国广播公司播出, 最终回归哥伦比亚广播公司。”,
        “尽管该节目播出了很长时间, 已知现存的剧集还不到24集。”,
      ],
    ],
    [
      “珍妮特·沃尔多”,
      [
        “珍妮特·玛丽·沃尔多（2月4日, 1920年 – 2016年6月12日, 2016年）是一位美国广播节目主持人和配音女演员。”,
        她在动画领域最为人熟知的角色是朱迪·杰森, 南希在 《沙赞》, 佩内洛普·皮特斯托普，以及
        ↪ 乔西在 《乔西与小猫咪》中，以及在广播剧《 遇见科莉丝
        ↪ ·阿彻》中担任 ,
        ↪ 标题角色。' ,
      ],
    ],
    [
      《亲吻与告白》（1945年电影）,
      [
        《Kiss and Tell》是一部1945年的美国喜剧电影，由当时17岁的秀兰·邓波儿饰演科莉丝
        ↪ 阿彻。” ,
        “在这部电影中，两个十几岁的女孩开始对男孩产生兴趣，这让她们各自的父母忧心忡忡
        ↪ 。”,
        “父母们争论哪个女孩的影响更坏，结果带来的麻烦比解决的问题还多。”,
      ],
    ],
    [
      “宪法事务大臣”,
      [
        宪法事务大臣这一职位是英国政府的一个职位,
        ↪ 于2003年设立。" ,
        "大法官的一些与大法官部相关的职能
        ↪ 被移交给国务大臣。",
        "后来，更多职能也从第一国务大臣移交给宪法事务大臣,
        ↪ 第一国务大臣是政府中的一个职位，由副
        ↪ 首相担任。" ,
      ],
    ],
  ],
  "type": "bridge",
  "level": "hard" ,
}
```


**图 D.1：来自 HotpotQA 案例研究的元数据。**