LaSER：将显式推理内化至潜在空间以实现稠密检索

Jiajie Jin[1,2] **Yanzhao Zhang**[2] **MingxinLi**[2] **DingkunLong**[2]
PengjunXie[2] **Yutao Zhu**[1] **Zhicheng Dou**[1] *[1]中国人民大学高瓴人工智能学院[2]阿里巴巴集团通义实验室

arXiv:2603.01425v1 [cs.CL] 2 Mar 2026

摘要

大语言模型（LLM）从根本上改变了稠密检索，将骨干模型从判别式编码器升级为生成式架构。然而，当前仍存在一个关键脱节：尽管大语言模型具备强大的推理能力，现有检索器却主要将其用作静态编码器，其复杂推理潜力尚未得到发掘。为解决这一问题，现有方法通常采用“先改写后检索”流水线，在检索前生成显式思维链（CoT）推理依据。然而，这会造成难以承受的延迟。相反，利用潜在词元的隐式推理方法虽然效率更高，却常因缺乏显式监督而出现语义退化。本文提出**LaSER**，这是一种新颖的自蒸馏框架，可将显式推理内化到稠密检索器的潜在空间中。LaSER 基于共享大语言模型骨干网络运行，并引入双视图训练机制：显式视图对真实推理路径进行显式编码，潜在视图则执行隐式潜在思维。为弥合这两种视图之间的差距，我们设计了一种多粒度对齐策略。除了标准的输出对齐外，我们还引入轨迹对齐机制，使潜在路径的中间潜在状态与显式推理片段的语义进展保持同步。这样，检索器无需自回归文本生成，便能静默而有效地“思考”。在域内和域外的推理密集型基准上进行的大量实验表明，LaSER 的表现显著优于最先进基线。此外，对不同骨干网络和模型规模的分析验证了我们方法的稳健性，证实我们的统一学习框架对于激发有效的潜在思维至关重要。我们的方法成功地将显式思维链流水线的推理深度与标准稠密检索器的推理效率结合起来[1]。

1 引言

稠密检索已发展成为现代信息检索（IR）系统的基础组成部分，广泛应用于网络搜索、问答 [16, 42], 和检索增强生成（RAG）系统 [10, 20]。这些检索器通过共享嵌入空间中的语义匹配，有效弥合了词汇不匹配的鸿沟 [16, 35, 40]。随着大语言模型（LLM） [1, 47, 49], 的兴起，该领域经历了一次重大的范式转变：检索器骨干网络正从小规模判别式模型（如 BERT、RoBERTa）转向大规模生成式大语言模型 [21, 36]。基于大语言模型的最先进检索器 [19, 40, 46]展现出了更强的语义理解能力，优于其前代模型。尽管基于大语言模型的检索器展现出了更强的语义理解能力 [4, 36, 46],，但仍存在一个悖论：这些检索器虽然继承了底层大语言模型的海量知识和推理潜力，却主要通过对比学习进行训练，以优化表征区分性 [15]。这一训练目标在很大程度上使生成式架构固有的推理能力处于休眠状态，将大语言模型仅仅视为更强大的编码器，而非推理器 [30, 32, 34]。

处理复杂用户查询时，这一局限尤为明显。现实世界的搜索场景往往涉及隐含意图 [26, 48], 多跳逻辑 [13, 42], 或模糊描述 [25, 39] ，需要进行超越表层语义匹配的推理。为弥合这一差距，一种常见的解决方案采用“先改写后检索”流水线 [9, 41], ，利用外部大语言模型在检索前显式生成查询扩展或思维链（CoT）推理依据 [7, 33, 38, 43, 45]。虽然这种方法能够提升性能，但其依赖冗长的自回归生成，因此造成了难以承受的延迟瓶颈 [11]（如图1所示）。为缓解这些延迟问题，近期研究探索了隐式推理 方法，使检索器能够在生成最终表征 [34]之前，在嵌入空间中生成“潜在思维令牌”。尽管前景可观，这些方法仍面临一个关键挑战：**缺乏显式语义监督**。仅依赖最终的对比学习损失，生成的潜在令牌往往缺乏充分引导，难以准确捕捉检索所需的关键信息，尤其是在涉及推理的查询中。

我们认为，**解决这一困境的关键在于内化**：利用大语言模型的显式推理能力作为特权监督，引导检索器进行轻量级潜在思考。我们的核心洞见是，尽管推理时生成显式思维链效率低下，但这些推理路径所承载的语义极具价值。我们不将显式推理和隐式推理视为彼此独立的范式，而是提出在统一框架中对齐二者，使检索器能够在潜在空间中“默默思考”。基于这一洞见，我们提出一种双视图框架，使模型学会将显式推理能力内化到潜在空间中。显式视图学习将推理逻辑编码进嵌入，作为语义上界。**潜在视图**生成潜在思维令牌，在内部模拟推理过程。

为了有效地将推理能力从显式视图迁移到潜在视图，我们提出了一种多粒度对齐策略。我们不仅对齐最终的查询表征（输出对齐），还对中间潜在状态施加辅助监督（轨迹对齐），以促使潜在

*通讯作者。[1]代码、模型和训练数据可在 https://github.com/ignorejjj/LaSER 获取。

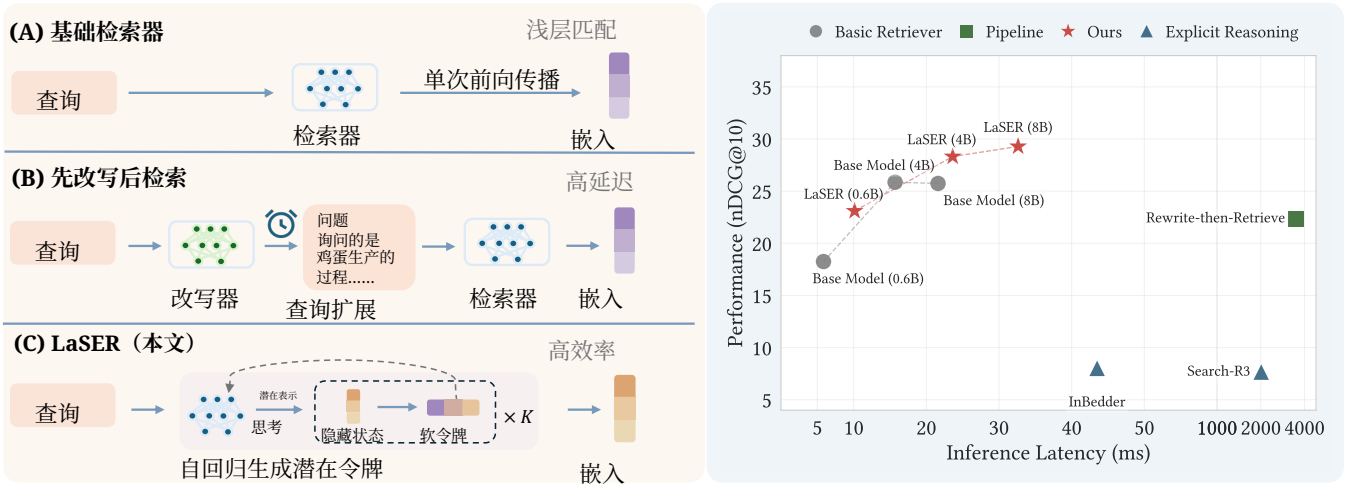


图1：左：我们提出的流水线与传统检索器和先改写后检索方法的比较。右：不同方法在 Bright 基准上的性能与延迟比较。延迟表示处理单个查询所需的时间。“先改写后检索”方法使用一个 0.6B 检索器和一个 3B 改写器。

词元捕捉显式思维链的语义进展。该框架具有很强的灵活性。在训练期间，显式视图充当语义教师；在推理期间，LaSER 以精简模式进行潜在推理，无需生成文本，同时保持与计算密集型流水线方法相当的性能。我们在一系列全面的嵌入基准上评估 LaSER，重点关注 Bright [32], BrowseComp-Plus [6],和 FollowIR [39] 等推理密集型任务。实验结果表明，仅在推理期间引入少量潜在词元，我们的方法便能显著提升检索性能，同时避免自回归解码带来的延迟。此外，LaSER 的表现始终优于现有的显式和隐式推理基线。值得注意的是，在某些推理密集型场景中，LaSER 甚至优于依赖外部大语言模型进行查询改写的高计算开销流水线，从而验证了我们所提出的蒸馏机制的有效性。

我们的主要贡献总结如下：

- 我们提出 LaSER，这是一种统一的检索框架，通过共享骨干模型中的自蒸馏，弥合显式思维链推理与隐式潜在思维之间的差距。
- 我们设计了一种双视图对齐机制，结合输出级和过程级轨迹对齐，有效地将推理语义蒸馏到潜在词元中，并防止隐式推理模型中常见的退化现象。
- 大量实验表明，LaSER 在不同骨干架构和模型规模下均持续优于基线，有力验证了我们提出的自蒸馏机制的有效性。
- 我们通过实证验证，显式推理路径的语义能够有效蒸馏至潜在表示空间。这使 LaSER 在达到计算开销巨大的“先改写后检索”流水线相当性能的同时，仍能保持标准稠密检索器的推理效率。

2 相关工作

2.1 基于大语言模型的检索与查询扩展

稠密检索的范式已从小规模判别式骨干网络（例如 BERT [17]）显著转向大规模生成式大语言模型 [36, 46]。借助大语言模型蕴含的海量世界知识和语义理解能力，这些检索器在标准语义匹配任务中表现出色。然而，它们仍主要依赖对比学习目标，而这类目标往往难以应对需要深度推理或意图消歧的复杂查询。为弥合这一差距，“先改写后检索”流水线日益普及（如图1所示），利用外部大语言模型扩展查询 [9, 37] 或在检索前生成推理依据 [3, 29]。尽管这种方法有效，却将推理与检索割裂开来，并因自回归解码而带来显著延迟，同时也使系统部署更加复杂。与这些多阶段流水线不同，我们的工作旨在将改写和推理能力内化到检索器参数中，在保留生成式扩展所带来语义优势的同时，维持单阶段双编码器的推理效率。

2.2 检索中的推理

近期研究探索了将推理能力直接整合到检索系统中，通常分为以数据为中心的方法 [8, 23, 40] 和以机制为中心的方法。以数据为中心的方法通过合成推理密集型查询及其对应的困难负例来增强检索。相反，以机制为中心的方法会修改检索器的架构以支持推理，并进一步分为显式和隐式策略。显式方法利用检索器本身生成自然语言思维链推理依据，然后生成最终嵌入 [7, 11, 28, 33, 43]。虽然这种方法提高了可解释性，但要求模型同时学习两项任务，因而增加了训练难度 [27]。此外，它也面临相同的

计算瓶颈。另一种隐式推理则通过生成“潜在思维令牌” [34] 或多向量表示 [2]，在嵌入空间中模拟这一过程，从而在不进行可见文本生成的情况下捕捉查询的细微差异。尽管效率较高，现有隐式方法主要依赖最终的对比损失进行监督，这成为捕捉复杂查询语义的瓶颈，并阻碍了性能的进一步提升。我们的框架通过将显式思维链用作特权监督信号来解决这一局限，将显式推理路径的语义蒸馏到潜在词元中，以确保隐式过程具有语义意义。

2.3 信息检索中的知识蒸馏

知识蒸馏（KD）广泛用于将强大教师的能力迁移至高效学生 [24], ，通常分为信息检索中的输出级蒸馏和特征级蒸馏。输出级蒸馏通常利用大语言模型生成合成训练数据，例如查询扩展或伪相关性标签 [46], ，供学生模仿。特征级蒸馏则侧重于对齐连续表示，通常训练学生检索器，使其逼近强大教师模型的隐空间或相关性分数 [22, 44]。关键在于，这些方法以结果为导向：它们改善相关性信号，却没有从根本上改变学生的信息处理机制。与这一范式不同，我们受到大语言模型潜在推理的启发 [5, 12, 31]，将显式推理步骤内化为潜在状态 [34]。我们并非仅仅蒸馏相关性分数，而是利用显式思维链作为“真实依据”逻辑，来监督潜在思维。据我们所知，LaSER 是首个将显式推理蒸馏为潜在推理应用于稠密检索的工作。

3 方法

本节介绍 LaSER，这是一种统一框架，旨在将显式推理能力内化到稠密检索器的潜在空间中。我们首先对检索问题进行形式化，并概述该框架。随后，我们详细说明框架的训练过程与优化目标，重点介绍多粒度自蒸馏机制。

3.1 问题形式化

给定用户查询 q和候选文档语料库 $\mathbf{D} = \{d_1, d_2, \ldots\ldots, d_N\}$，稠密检索的目标是学习一个编码器 $f(\cdot)$，将文本输入映射到 m 维嵌入空间 $\mathbb{R}^m$。查询 q 与文档 d 之间的相关性分数通常通过其嵌入 v_q和 v_d) 的余弦相似度计算，即 $s(q, d) = \cos(v_q, v_d)$。

在标准的基于大语言模型的检索器中，查询表示通过单次前向传播获得。为增强指令遵循能力，通常会在 q前添加任务专属指令 I。随后，在输入序列末尾追加一个[序列结束标记]，并提取与该标记对应的最后一层隐藏状态作为嵌入：

$$v_q = f([I_{\text{task}}; q; [\text{EOS}]]). \tag{1}$$

然而，这种浅层处理往往无法捕捉推理密集型查询所需的复杂隐含意图。

为了解决这一问题，现有的“先改写后检索”方法引入外部大语言模型推理器 M 来生成显式推理路径，然后由检索器对这一丰富后的上下文进行编码：

$$v_q^* = f([I_{\text{task}}; q; r_q; [\text{EOS}]]), \text{where } r_q = \mathcal{M}(I_{q2r}, q). \tag{2}$$

虽然 v_q^* 能够捕捉更深层的语义，但文本的自回归生成r_q 会带来难以承受的延迟和部署成本。

在本研究中，我们旨在通过训练模型执行潜在推理来弥合这一差距。模型不再生成显式文本r_q，而是在嵌入空间中自主生成一系列 K连续的潜在思维令牌 $T = \{t_1, \ldots, t_K\}$。最终的查询表示v_q 由这些潜在状态推导而来，旨在近似v_q^* 的语义深度，同时保持单模型架构的高效性。

3.2 框架概述

如图2所示，LaSER建立在具有因果注意力机制的统一生成式大语言模型骨干网络之上。为实现推理能力的内化，我们采用双视图训练范式，让模型通过两个共享相同参数 θ且相互对齐的视角进行学习：

显式视图。 该视图在训练过程中充当语义上界。它接收经过增强的查询，其中包含由更强大的推理器生成的高质量显式思维链推理依据。通过在这一丰富上下文上执行单次前向传播，显式视图获得能够全面融合原始查询与显式推理路径的表征。

潜在视图。 该视图是推理时的目标模式。它仅以原始查询作为输入，并在生成最终查询表示之前，于潜在嵌入空间中生成一系列潜在思维令牌 。我们设计了一种自蒸馏框架，以促进这两个视图的协同训练，尤其是潜在视图，详见第 3.5 节。

推理。 推理时，LaSER 仅在潜在视图中运行。它利用学到的潜在思维令牌对查询进行编码，确保延迟与标准稠密检索器相当，同时受益于内化的推理能力。此外，如果可用，该模型也可以灵活地接收外部大语言模型改写后的查询，从而更好地适应多样化的部署场景。

3.3 潜在视图

潜在视图的主要目标是增强模型对复杂查询的推理能力，同时避免显式文本生成带来的延迟开销。受连续推理领域近期进展的启发，我们提出一种机制：由 θ 参数化的编码器模型在连续嵌入空间中进行“思考”。该方法以自回归的连续思维向量序列取代离散词元生成，使模型能够以完全可微分的方式逐步完善对查询意图的理解。

如图2所示，给定输入查询 q，模型首先将离散词元序列映射为稠密嵌入 $\mathbf{X}_0 = \mathbf{E}(q)$，其中 $\mathbf{E} \in \mathbb{R}^{|\mathcal{V}| \times m}$ 是嵌入矩阵，$|\mathbf{V}|$ 是词汇表大小。骨干大语言模型 f_θ 处理该序列，生成初始隐藏状态。

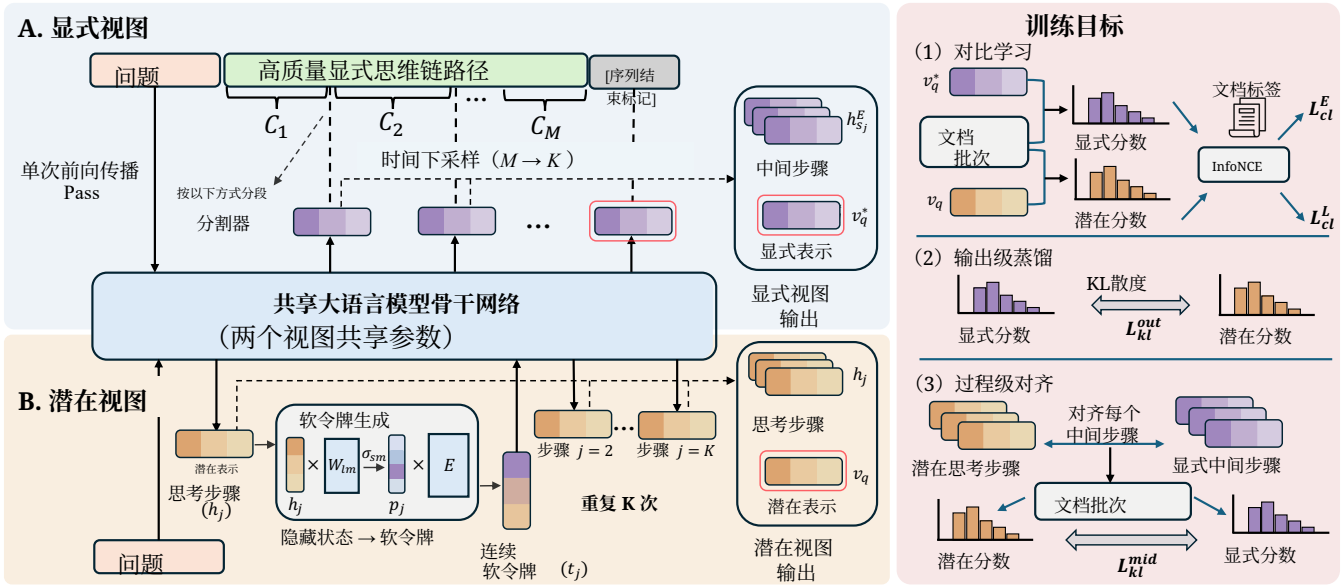


图2：LaSER 示意图。在训练过程中，每一步都涉及两条推理路径：显式视图和潜在视图（如左侧所示）。这两条路径的输入有所不同，显式视图包含额外信息。然而，在推理阶段，仅使用潜在视图模式。损失函数根据两条路径的结果计算（如右侧所示）。

为便于进行潜在推理，我们为查询扩展一系列 K 潜在思维令牌 $T = \{t_1, \ldots, t_K\}$。与通过采样离散索引进行的标准解码不同，我们将这些令牌构造为连续向量，以保留语义丰富性和可微性。该过程以自回归方式进行：在第 j 个思考步骤（$1 \le j \le K$），令 $h_{j-1} \in \mathbb{R}^m$ 表示骨干模型输出的最后一个隐藏状态。我们通过语言建模头 $W_{\text{lm}} \in \mathbb{R}^{m \times |\mathcal{V}|}$ 将 h_{j-1} 投影到词汇空间，得到逻辑值 l_j：

$$l_j = W_{\text{lm}} h_{j-1}. \tag{3}$$

我们不进行硬选择（例如，arg max），而是计算词汇表上的概率分布 $p_j = \text{Softmax}(l_j)$。然后，根据该分布下的期望嵌入向量计算软潜在标记 t_j：

$$t_j = p_j^\top \mathbf{E} = \left(\text{Softmax}(W_{\text{lm}} h_{j-1})\right)^\top \mathbf{E}. \tag{4}$$

这个软令牌 t_j 会附加到下一步的输入序列中。模型以自回归方式运行，其中步骤 j 的输入包含原始查询和此前所有潜在词元：

$$\mathbf{H}_j = f_\theta\left([\mathbf{E}(q), t_1, \ldots, t_j]\right), \tag{5}$$

其中，$\mathbf{H}_j$ 表示隐藏状态序列。通过因果注意力机制，每个潜在令牌 t_j 都可以关注查询上下文和先前的思考步骤，从而模拟显式思维链推理的序列逻辑。

经过 K 步后，我们得到一系列推理状态 $\mathbf{H}_{\text{think}} = \{h_1, \ldots\ldots, h_K\}$。为了将这一推理过程的全局语义合成为统一的表征，我们对这些状态进行均值池化，得到最终的查询向量 v_q：

$$v_q = \frac{1}{K} \sum_{j=1}^{K} h_j. \tag{6}$$

文档的表征 v_d 也以相同方式获得。这一潜在思维过程保持严格的自回归特性，确保与 KV 缓存优化完全兼容。此外，通过限制推理范围 K（通常为 $K < 10$），该方法实现了与标准稠密检索器相当的推理延迟，显著优于需要解码较长文本链的生成式方法。

3.4 显式视图

显式视图充当语义教师，利用从先进外部推理器获得的特权知识来学习显式推理过程。具体而言，对于每个查询 q，我们提示外部推理器生成全面的思维链推理依据 r_q，其中包含中间推理逻辑以及对查询潜在意图的理解（详细提示见 1）。我们通过拼接原始查询 q、生成的推理依据 r_q 和 [序列结束标记]，构造增强输入序列 x_{aug}：

$$x_{\text{exp}} = [q; r_q; [\text{EOS}]]. \tag{7}$$

随后，模型按照公式 2 对这个增强后的输入进行编码。由于显式视图与潜在视图的输入不同，这两项任务之间不会相互干扰。由于推理逻辑已明确包含在文本输入中，编码器只需进行单次前向传播，并将最后一个隐藏状态作为最终表征 v_q^*，如公式 (2) 所示。鉴于

表1：用于标注给定查询的显式推理路径的提示语。

指令：
给定一个问题，你的任务是遵循以下指令：
1. 确定核心问题。
2. 请逐步思考，推理并描述哪些信息可能
与详细回答这些问题相关且有帮助。
3. 尽可能多地思考并拟定答案。
给定的问题：
[问题开始]
{Original Query}
[问题结束]

显式路径本身可以提供中间信号，因此我们还会提取序列中特定中间词元的隐藏状态，用以表征中间推理过程（如图2所示）。这些细节将在第3.5.3节讨论。

3.5 通过自蒸馏进行优化

为了将推理能力从显式视图迁移到潜在视图，我们提出了一种联合训练目标，结合了对比学习和多粒度蒸馏。

3.5.1 对比训练。两个视图都必须学习基本的判别信号。我们对两个视图均采用标准的 InfoNCE 损失 [15]，使用真实标注的正样本文档 d^+ 和负样本 d^-。对于潜在视图：

$$\mathcal{L}_{cl}^{L} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(v_q \cdot v_{d^+}/\tau)}{\sum_{d\in\{d^+\}\cup\{d^-\}}\exp(v_q \cdot v_d/\tau)}, \tag{8}$$

其中，N 表示批次大小，τ 表示温度超参数。显式视图也使用 v_q^* 计算类似的损失 L_{cl}^{E}。

3.5.2 输出级蒸馏。为了确保潜在视图有效地内化显式推理能力，我们将显式视图中的语义知识蒸馏到潜在视图中。由于原始输入与经过推理增强的输入之间存在固有的信息差距，严格对齐查询嵌入 v_q 和 v_q^* 可能会对潜在视图施加过度约束，因此我们采用基于分数的蒸馏策略。该方法着重迁移显式视图中蕴含的排序偏好。形式上，对于查询 q 和文档批次 $\mathrm{B} = \{d_1, \ldots, d_N\}$，我们对齐由两个视图得出的相关性分数概率分布。

令 $s_i^E = v_q^* \cdot v_{d_i}$ 和 $s_i^L = v_q \cdot v_{d_i}$ 分别表示教师和学生的相似度分数。我们应用带有蒸馏温度 τ_{kd} 的 softmax 函数，得到概率分布：

$$p_i^E = \frac{\exp(s_i^E/\tau_{kd})}{\sum_{k=1}^{N}\exp(s_k^E/\tau_{kd})}, \quad p_i^L = \frac{\exp(s_i^L/\tau_{kd})}{\sum_{k=1}^{N}\exp(s_k^L/\tau_{kd})}. \tag{9}$$

输出级蒸馏损失随后被定义为这两个分布之间的 Kullback-Leibler（KL）散度：

$$\mathcal{L}_{kl}^{out} = \frac{1}{N}\sum_{i=1}^{N} p_i^E \log\left(\frac{p_i^E}{p_i^L}\right). \tag{10}$$

通过最小化 L_{kl}^{out}，潜在视图学会模仿显式视图中文档相关性的细粒度信息，从而有效捕捉经推理增强的判断逻辑，而无需要求嵌入坐标完全相同。

3.5.3 过程级轨迹对齐。仅依赖输出对齐，往往会将推理过程视为黑箱，这可能导致潜在词元退化，无法编码有意义的中间语义。为避免这一问题，我们引入轨迹对齐 机制，对中间思考步骤进行显式监督。

我们的设计基于这样的假设：潜在词元序列应当作为冗长显式思维链的压缩语义轨迹。显式推理依据 r 由长度可变的逻辑片段 $\{s_1, ..., s_M\}$ 构成（其中 M 为推理步骤数），而潜在表示视图使用固定预算的 K 个词元。由于通常有 $M \geq K$，我们认为每个潜在词元 t_i 都应对应显式推理路径中的一个特定语义关键帧。为实现这一点，我们采用**时间下采样** 策略，以对齐两种视图的粒度。我们通过均匀采样，将第 i 个潜在步骤映射到对应的显式片段索引 j_i：

$$j_i = \lfloor i \times \frac{M}{K} \rfloor. \tag{11}$$

这种映射确保模型将其潜在思维与显式推理链的语义进展同步。随后，我们通过最小化文档批次上两者投影分布之间的 KL散度来对齐中间状态：

$$\mathcal{L}_{kl}^{mid} = \frac{1}{K}\sum_{i=1}^{K}\mathrm{KL}\left(\sigma(h_{j_i}^E \cdot v_{\mathcal{D}}), \sigma(h_i \cdot v_{\mathcal{D}})\right), \tag{12}$$

其中，σ 表示文档批次上的softmax函数。通过强制执行这种对齐，我们确保潜在词元不会偏离，而是充当语义检查点，严格近似显式推理逻辑。

3.5.4 总体目标。最终训练目标是各组成部分的加权和：

$$\mathcal{L} = \mathcal{L}_{cl}^{L} + \lambda_1\mathcal{L}_{cl}^{E} + \lambda_2\mathcal{L}_{kl}^{out} + \lambda_3\mathcal{L}_{kl}^{mid}, \tag{13}$$

其中 $\lambda_1, \lambda_2, \lambda_3$ 是用于平衡各项任务的超参数。

4 实验设计

4.1 数据集与指标

为提升模型的推理能力，我们使用来自 ReasonEmb [3] 的合成数据集进行训练，其中包含涵盖 12 个领域的 81k 个训练样例。每个训练查询都附有一条由 GPT-4o-mini [14] 作为推理器生成的推理路径，作为显式视图中的显式输入。

在评估中，我们选择了三个需要深度推理的基准：Bright [32],FollowIR [39]（域外）和 BrowseComp- Plus [6]（域外）。遵循标准做法，我们报告 BRIGHT 的 nDCG@10，以及 BC-Plus 的 Recall@5、Recall@100 和 Recall@1000。对于 FollowIR，每个子集的评估均包含一个标准指标，以及一种旨在

评估指令遵循能力的成对指标 p-MRR。遵循官方协议，除 p-MRR 外，我们还报告 Robust'04 和 Core'17 子集的 MAP@5，以及 News'21 子集的 nDCG@5。

训练数据集和评估数据集的统计信息如表2所示。

4.2 基线

我们将LaSER与四类基线方法进行对比。为确保比较严谨，除非另有说明，所有可训练基线均使用与我们的方法相同的骨干模型和训练数据。

表2：我们实验中所用数据集的统计信息。"Q"和"C"分别表示查询和语料库。"Len."表示平均词元数。"Reas. Len."表示由外部大语言模型标注的平均推理长度。

数据集划分		# Q	# C	问题长度	推理长度
训练集	ReasonEmb	81,659	–	222.1	984.8
Test	Bright（分布内）	1,384	1,145,164	240.8	2,412.8
	FollowIR（分布外）	104	98,312	76.6	–
	BC-Plus（分布外）	830	100,195	123.2	–

标准稠密检索器： 我们纳入了当前最先进的基于编码器的模型，例如 BGE-M3 [4] 和 E5-Large-Instruct [35], , 以及基于大语言模型的检索器，包括 E5-Mistral [36] 和 Qwen3-Embedding [46]。此外，我们在复合数据集上使用标准对比学习训练了一个"公平基线"，以单独评估我们架构所带来的收益。

流水线方法（先改写后检索）： 该类别使用外部大语言模型，在检索前对查询进行显式改写。我们利用 BGE-Reasoner-Rewriter-7B [3] 生成改写后的查询，并在推理过程中将其作为检索器（使用公平基线模型）的输入。

显式推理方法： 这些方法将生成与检索整合在单一模型中，在输出表征之前生成显式思维链。我们将其与 Search-R3 [11] 和 GRACE [33] 进行比较。由于部分基线模型的训练代码未公开，我们使用其官方检查点进行推理。

隐式推理方法： 我们将其与 GIRCSE 这一代表性方法进行比较。该方法同样采用潜在词元对推理步骤进行建模，然后生成最终嵌入。

4.3 实现细节

我们使用 Qwen3 系列（0.6B、4B、8B）和 LLaMA 3.2 系列（1B、3B、8B）的基础版本进行实验。由于 LLaMA 3.2 没有提供 8B 版本，我们采用 LLaMA 3.1-8B 来评估该规模下的性能。所有模型均使用 LoRA ($r = 64$, $\alpha = 32$) 在 4 张 A100 GPU 上微调 1 个训练轮次，全局批次大小为 8（每个设备的批次大小设为 2，并进行 8 步梯度累积）。我们使用 AdamW 优化器，学习率为 1e-4，预热比例为 0.1。训练时，每个查询都配有一个困难负样本和批次内负样本。查询和文档的最大序列长度训练期间设为 512，测试期间则扩展至 8192，以适应更长的上下文。为适应显式视图处理的显式推理路径，我们将该视图在训练期间的最大序列长度设为 1024。对比损失（τ）和蒸馏损失（τ_{kd}）的温度超参数均设为 0.02。为平衡多任务目标，我们将损失权重设为：$\lambda_1 = 1$、$\lambda_2 = 10$ 和 $\lambda_3 = 0.1$。训练和推理阶段的潜在思考步骤数 K 均设为 3。

5 结果

本节将介绍 LaSER 的实验结果，并详细分析消融研究和效率。我们的评估旨在回答以下三个研究问题：

研究问题1：与采用标准对比学习及其他训练方法的公平基线相比，我们提出的框架表现如何？

RQ2：所提出的多粒度自蒸馏策略有何影响？

RQ3：与其他基线方法相比，LaSER 能否在检索性能和延迟之间实现更优的权衡？

5.1 主要结果

本节全面分析 LaSER 在域内推理基准（BRIGHT）和域外基准（FollowIR 和 BrowseComp-Plus）上的表现。为确保比较公平，我们使用所有方法一致的训练数据复现了公平基线，并基于两种不同的架构，采用三种模型规模进行了实验。主要结果见表3和表4（涵盖全部参数规模的实验详见第5.3小节）。我们的主要观察总结如下。

LaSER 显著优于标准稠密检索器。 首先，LaSER 的表现显著优于采用标准对比学习训练的公平基线。如表3所示，LaSER（Qwen3-8B）的平均 nDCG@10 达到 29.3。这一表现不仅超过了原始的 Qwen3-Embedding-8B（14.0），还比微调后的公平基线（25.7）高出约 15%。这一显著的性能差距表明，采用传统检索器架构进行标准对比训练，不足以应对复杂推理任务，因为它无法充分发挥模型内在的推理潜力。通过将推理能力内化到潜在空间中，LaSER 有效激发了大语言模型骨干网络的推理能力，弥合了浅层语义匹配与复杂推理之间的差距。

LaSER 达到显式流程的性能上限。 我们工作的一个关键动机，是在不承担高延迟代价的情况下，达到计算密集型"先改写后检索"流水线的性能。实验结果表明，在 Qwen3-0.6B 和 Qwen3-8B 骨干模型上，我们的方法均优于基于查询改写的显式推理方法（分别领先 0.7 和 1.2 分）。此外，LaSER 的表现显著优于需要生成多个思考令牌的显式推理模型，例如 Search-R3 和 InBedder。至关重要的是，LaSER 仅凭少量潜在词元便达到了这一竞争力水平，

表3：Bright 基准上的主要检索性能。我们报告所有子集的 nDCG@10。同一骨干网络设置下的最佳结果以粗体显示，次佳结果加下划线。† 表示一种流水线方法，该方法在推理过程中使用外部大语言模型重写查询。

模型	Size	StackExchange							编程		基于定理			Avg.
		Bio.	地球科学	经济学	Psy.	Rob.	Stack。	Sus.	Leet。	Pony	AoPS	TheoQ。	TheoT。	
标准稠密检索器														
BGE-M3	0.6B	7.2	13.3	12.5	12.8	12.6	10.0	10.1	15.6	30.7	1.5	5.6	5.0	11.4
Multilingual-E5-Large-Instruct	0.6B	15.0	24.6	14.0	14.6	16.1	10.3	13.8	15.1	2.1	3.0	11.6	6.9	12.3
E5-Mistral-7B-Instruct	7B	17.9	28.6	18.5	16.5	18.6	13.4	20.0	19.7	6.3	2.0	14.2	22.1	16.5
Qwen3-Embedding-0.6B	0.6B	12.7	26.3	17.9	16.5	12.5	12.4	12.2	14.3	0.7	3.1	17.2	26.5	14.4
Qwen3-Embedding-4B	4B	15.9	33.6	16.7	23.2	13.2	15.0	16.8	21.2	1.7	4.0	18.4	35.4	17.9
Qwen3-Embedding-8B	8B	14.7	17.9	15.5	19.9	9.1	12.9	16.5	17.4	0.8	2.5	16.8	24.5	14.0
基础对比学习														
公平基线（Qwen3-0.6B）	0.6B	28.8	31.6	25.2	28.1	15.1	22.3	24.2	8.8	**3.5**	**2.1**	14.1	15.4	18.3
公平基线（Qwen3-8B）	8B	49.7	51.2	26.9	37.4	23.4	28.0	34.1	3.7	3.2	**2.8**	16.8	31.8	25.7
公平基线（LLaMA3.1-8B）	8B	57.7	40.9	24.6	29.8	**20.1**	29.2	25.8	4.3	5.7	**1.6**	12.0	17.7	22.5
显式推理														
先改写后检索 (Qwen3-0.6B) †	0.6B	57.8	51.6	14.1	39.4	15.0	19.8	23.7	0.6	14.6	1.2	14.9	15.5	22.4
先改写后检索 (Qwen3-8B) †	8B	53.1	54.3	32.1	34.8	20.5	31.1	32.2	3.2	15.2	4.1	17.4	38.8	28.1
先改写后检索 (LLaMA3.1-8B) †	8B	63.1	54.5	27.7	40.6	17.1	28.1	28.2	1.6	11.4	3.8	16.1	30.3	26.9
Search-R3（Qwen2.5-1.5B）	1.5B	13.8	8.3	4.2	3.5	4.5	12.4	4.6	16.7	3.0	0.6	8.5	11.7	7.7
潜在推理														
GIRCSE（Qwen3-0.6B）	0.6B	29.0	32.8	24.7	30.6	13.6	24.0	**26.6**	**11.1**	1.1	1.3	13.5	20.6	19.1
GIRCSE（Qwen3-8B）	8B	**59.0**	**56.5**	27.2	40.3	19.0	28.5	31.4	3.2	3.6	1.7	14.0	27.2	26.0
GIRCSE（LLaMA3.1-8B）	8B	50.6	43.4	**28.3**	35.7	14.3	26.9	26.4	**6.0**	5.2	0.5	11.6	15.1	22.0
LaSER（Qwen3-0.6B）	0.6B	**50.0**	**45.9**	**25.7**	**32.4**	**18.4**	**27.1**	26.5	9.1	2.7	1.2	**16.0**	**22.4**	**23.1**
LaSER（Qwen3-8B）	8B	58.0	51.8	**29.0**	**44.1**	**26.0**	**32.9**	**34.5**	**9.2**	**11.7**	1.6	**18.7**	**34.0**	**29.3**
LaSER（LLaMA3.1-8B）	8B	**58.4**	**48.1**	28.0	**40.9**	17.0	**29.9**	**28.3**	1.7	**5.9**	1.5	**14.6**	**19.2**	**24.4**

从而避免了解码冗长显式思维链依据所带来的高昂计算成本。这验证了我们的核心观点：推理路径的语义可以通过自蒸馏有效压缩到潜在状态中。采用合适的骨干模型后，我们的方法能更有效地激发模型固有的推理能力，从而取得更优的表现。

LaSER 优于现有隐式推理方法。我们进一步将 LaSER 与最先进的隐式推理方法进行比较，具体而言是 GIRCSE。在涉及不同模型规模和数据集的 9 种实验设置中，LaSER 在 8 种情况下优于 GIRCSE。鉴于两种方法使用相同的输入上下文和计算预算，这些结果凸显了我们提出的自蒸馏框架的有效性。这表明，融入来自教师模型的监督有助于显著提升潜在思维的学习。与仅依赖输出对比信号的 GIRCSE 不同，LaSER 采用双重对齐策略，同时结合过程级和输出级蒸馏，确保中间潜在令牌能够捕获更丰富的语义信息。

5.2 消融研究

为回答 RQ2，我们以 Qwen3-0.6B 作为骨干模型，开展了全面的消融研究。鉴于 LaSER 是一个旨在内化显式推理的自蒸馏框架，我们的分析聚焦于两个关键维度：

（1）架构有效性。我们考察双视图结构和潜在思考机制的贡献。**无显式视图**会完全移除显式视图分支，仅依靠对比损失训练学生进行潜在推理，不使用特权监督。**无潜在视图**则以标准的单次前向传播取代潜在推理机制，以获得最终嵌入，从而评估即使没有潜在词元，共享骨干的学习是否也能使标准架构受益。

（2）学习策略有效性。我们分别考察协同学习范式和对齐目标的影响。**无过程知识蒸馏**移除轨迹对齐损失（L_{kl}^{mid}），以单独考察监督中间潜在状态的影响。**无输出级蒸馏**移除最终输出对齐损失（L_{kl}^{out}），仅保留过程级监督。**无协同学习**将在线自蒸馏替换为两阶段过程：由固定的显式视图教师（离线微调）生成静态监督标签，以验证动态共享骨干训练的有效性。

架构有效性。 我们首先验证所提出的潜在推理机制的必要性。如表5所示，移除潜在视图会导致最显著的性能下降（例如，Bright上的得分从23.10降至19.93）。这表明，标准稠密检索器固有的单向量输出形式造成了表示瓶颈，限制了推理能力，并妨碍模型有效吸收教师输入提供的额外信息。相比之下，

表 4：检索性能。Robust04、News21 和 Core17 下的列属于 FollowIR 基准。

模型	Size	Robust04		News21		Core17		FollowIR-Avg		BrowseComp-Plus		
		MAP@5	p-MRR	nDCG@5	p-MRR	MAP@5	p-MRR	得分 p-MRR		R@5	R@100	R@1000
标准稠密检索器												
BGE-M3	0.6B	1.5	-1.6	21.4	-1.3	7.5	-8.8	10.1	-3.9	4.1	21.7	47.2
Multilingual-E5-Large-Instruct	0.6B	2.4	2.7	25.0	1.2	8.5	-6.0	12.0	-0.7	10.1	36.8	68.6
E5-Mistral-7B-Instruct	7B	2.7	2.5	28.8	0.8	13.7	-7.8	15.1	-1.5	9.3	37.1	70.2
Qwen3-Embedding-0.6B	0.6B	2.8	8.9	27.3	3.6	11.4	2.7	13.8	5.1	8.0	29.1	64.8
Qwen3-Embedding-4B	4B	5.0	11.0	26.7	6.9	11.2	11.1	14.3	9.7	5.9	18.1	39.4
Qwen3-Embedding-8B	8B	3.1	6.2	25.5	8.8	10.1	6.6	12.9	7.2	7.7	31.6	61.3
基础对比学习												
公平基线（Qwen3-0.6B）	0.6B	2.1	3.5	13.4	-0.6	6.7	**-3.4**	7.4	-0.2	3.5	21.3	46.9
公平基线（Qwen3-8B）	8B	2.8	4.4	18.9	**2.0**	11.2	**-1.3**	11.0	1.7	11.3	37.4	63.2
公平基线（LLaMA3.1-8B）	8B	2.5	**2.8**	18.9	0.1	8.1	-2.6	9.8	0.1	6.1	25.4	50.8
显式推理												
Search-R3（Qwen2.5-1.5B）	1.5B	3.2	7.1	26.2	-0.1	10.1	2.7	13.2	3.2	0.0	0.3	1.1
InBedder（LLaMA2-7B）	7B	2.6	3.2	8.5	3.1	1.6	-6.4	4.2	-0.1	1.4	8.1	27.4
潜在推理												
GIRCSE（Qwen3-0.6B）	0.6B	**3.0**	**3.9**	12.8	-0.9	5.7	-7.7	7.2	-1.6	**6.9**	25.2	52.8
GIRCSE（Qwen3-8B）	8B	3.0	4.2	**22.6**	0.7	8.5	1.0	11.4	2.0	**13.0**	**40.8**	**68.1**
GIRCSE（LLaMA3.1-8B）	8B	2.9	1.0	19.3	-1.3	**12.2**	**-1.2**	11.5	-0.5	6.7	23.5	51.0
LaSER（Qwen3-0.6B）	0.6B	2.0	3.4	**25.0**	**1.0**	**7.4**	-4.5	**11.5**	**0.0**	6.8	**26.8**	**54.9**
LaSER（Qwen3-8B）	8B	**4.1**	**5.8**	21.8	0.2	**11.4**	1.3	**12.5**	**2.4**	11.7	38.4	66.9
LaSER（LLaMA3.1-8B）	8B	**3.6**	2.7	**22.0**	**1.6**	11.1	-1.9	**12.2**	**0.8**	**6.8**	**25.7**	**52.9**

表5：我们提出的方法的消融研究。对于 Fol- lowIR，我们报告所有子集的平均分数。

方法	Bright	FollowIR		BrowseComp-Plus	
	nDCG@10	分数	p-MRR	R@5	R@1000
LaSER（Qwen3-0.6B）	**23.10**	11.49	-0.02	6.76	**54.88**
无显式视图	20.59	9.78	-1.71	6.48	50.14
无潜在视图	19.93	8.62	-0.11	5.93	52.25
无过程对齐	22.33	**11.63**	-1.58	6.18	52.43
无输出蒸馏	19.97	7.55	**0.43**	**7.03**	53.02
无协同学习	20.98	10.87	-0.54	5.56	52.57
基础对比学习	18.25	7.38	-0.17	3.54	46.9

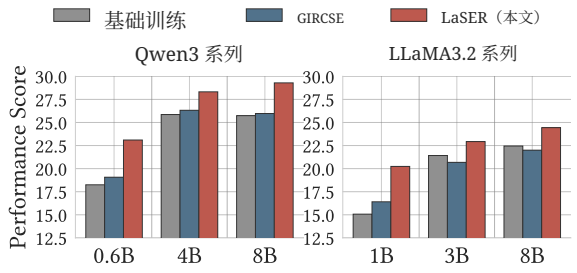


图 3：LaSER 与基线在 Bright 上使用不同骨干网络和模型大小时的性能比较。

我们的潜在思维令牌成功扩展了语义容量。我们还考察了显式教师的作用。移除显式视图也会对性能产生类似的负面影响。这表明，若没有显式思维链监督，中间潜在词元的学习效果会减弱，这可能是因为缺少明确的语义指导。

多粒度对齐的影响。我们进一步分析了各项学习策略的有效性。实验结果表明，输出级蒸馏和过程级蒸馏都至关重要；移除其中任一部分都会显著削弱模型的推理能力和跨任务泛化性能。值得注意的是，去除输出蒸馏会导致更明显的性能下降，这证实了将教师的细粒度排序偏好迁移给学生，是有效开展学生学习的前提。我们还观察到，在线学习具有显著影响。在分布内和域外数据集上，允许教师动态学习都能带来更好的结果。这验证了我们的共享骨干设计有助于促进不同视图之间更好的相互适应（关于这一主张的进一步分析见第 5.6 节）。

5.3 鲁棒性分析

我们验证了 LaSER 在不同骨干模型架构下的鲁棒性，主要有两点观察：

持续优于基线。我们在 Bright 上使用各种骨干模型进行了实验。如图3所示，在所有 6 种设置中，LaSER 在不同模型系列上始终优于所有基线。值得注意的是，在轻量级骨干网络（例如 0.6B 和 1B）上，LaSER 展现出更强的有效性（23.10），甚至超过了计算密集型重写流程（22.35）。这表明，我们的自蒸馏

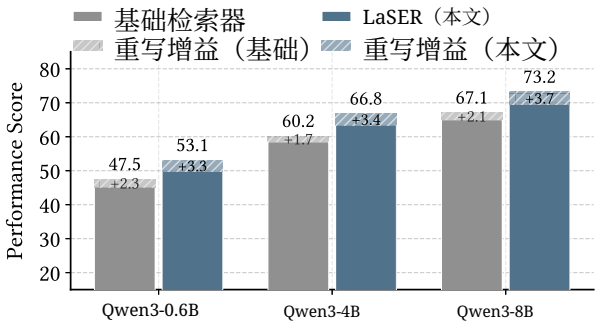


图 4：本文方法与基础训练在 Bright 上使用重写路径时的性能比较。

框架使小型模型无需承担外部推理器带来的开销，也能有效地"思考"。

仅对比式潜在思维的不稳定性。我们观察到，仅依赖对比学习目标的 GIRCSE 在不同架构上的表现不稳定。在某些设置下（例如 LLaMA3.2-3B），其性能低于标准检索器基线；而在其他设置下，提升也仅为边际水平。相比之下，通过引入蒸馏信号，我们的方法在类似架构上取得了显著提升。这表明，潜在思考机制可能需要特权监督，才能有效学习。

5.4 延迟分析

为回答 RQ3，我们在 BRIGHT 基准的 80 个查询子集上评估推理延迟，使用一块 NVIDIA A100（80GB）图形处理器。为确保比较严谨，我们使用 vLLM [18] 对显式推理流水线中的外部改写模型进行推理，而所有检索器骨干网络（包括我们的方法）均使用原生 Hugging Face Transformers 实现。所有基线和我们的方法均采用 8 的批次大小。如图1所示，LaSER 大幅降低了计算成本，其延迟仅为"先改写后检索"流水线的 0.3 %，同时保持了相当的检索效果。尽管由于潜在思维词元的自回归生成，LaSER 相较于标准基础检索器会产生少量延迟开销（约为 1.7×），但随着模型规模增大，这一开销会逐渐降低。对于更大的骨干网络，与模型前向传播时间相比，向量运算的相对成本微不足道，因此系统能够更充分地受益于 KV 缓存机制。

5.5 潜在推理步骤的影响

我们研究了训练和推理过程中潜在思考步数（K）对 Qwen3-8B 的影响，结果如图 5 所示。

训练效率。我们观察到，将训练步数从 $K = 3$ 增加到 $K = 6$，性能提升微乎其微，甚至略有下降。我们将这一现象归因于较高的序列——

我们的方法框架所促进的潜在空间语义密度。

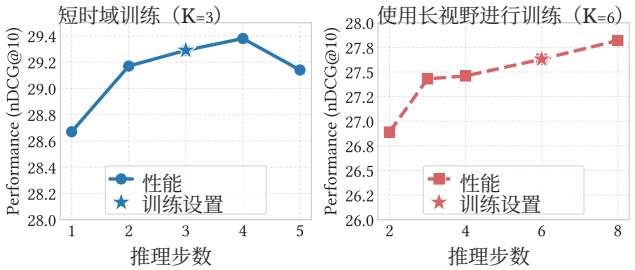


图5：训练和推理过程中潜在思考步骤的影响。

不同于标准的隐式推理，显式视图的引入提供了特权监督，成为明确的优化目标。这使 LaSER 能够有效压缩推理过程，只需少量潜在步骤即可捕捉推理路径的核心逻辑。因此，在训练过程中强制使用更大的 K，可能会迫使模型与显式文本中细碎且无信息量的片段对齐，从而引入噪声。

推理缩放。 相反，在推理过程中增加 K，性能会持续提升，这表明 LaSER 已习得稳健的嵌入迭代精炼机制。模型并非只是记忆固定的步骤轨迹或复制先前状态，而是会主动利用额外的计算步骤，深化语义理解并优化查询表示。

5.6 双视图协同学习分析

由于我们的框架涉及对共享同一骨干模型的两个不同视图进行联合训练，我们进一步探究了二者之间的协同作用。具体而言，我们通过比较推理时向基础检索器和 LaSER 提供显式改写查询后所获得的性能提升，评估骨干模型处理显式推理的能力。如图 4 所示，与基础检索器相比，在所有模型大小下，LaSER 从显式改写中获得的性能提升都显著更大（例如，Qwen3-4B 上分别为 3.4 和 1.7）。这一显著差异表明，我们的显式视图训练目标有效优化了共享骨干模型，使其能够更好地捕捉并编码思维链依据中的语义信息，以用于检索任务。因此，这种增强的显式处理能力构成了更优的语义上界，从而有助于更有效地蒸馏到潜在视图中。

6 结论与未来工作

本文提出 LaSER，这是一种新颖的自蒸馏框架，可将显式思维链推理内化到稠密检索器的潜在空间中。LaSER 采用多粒度对齐策略，使潜在思维令牌与显式推理轨迹同步，有效弥合了深层语义推理与检索效率之间的差距。大量实验表明，LaSER 的表现显著优于最先进基线，性能可媲美计算开销高昂的"先改写后检索"流水线，同时保持较低的

推理延迟。我们的工作验证了显式推理的语义可以有效压缩到潜在状态中，为复杂查询理解提供了一种高效方案。尽管结果令人鼓舞，我们的方法仍只是迈向完全自主潜在推理的初步一步。未来，我们计划探索强化学习技术，通过直接根据检索效用优化潜在推理轨迹，进一步优化中间推理过程。

参考文献

[1] Qingyao Ai、Ting Bai、Zhao Cao、Yi Chang、Jiawei Chen、Zhumin Chen、Zhiyong Cheng、Shoubin Dong、Zhicheng Dou、Fuli Feng、Shen Gao、Jiafeng Guo、Xiangnan He、Yanyan Lan、Chenliang Li、Yiqun Liu、Ziyu Lyu、Weizhi Ma、Jun Ma、Zhaochun Ren、Pengjie Ren、Zhiqiang Wang、Mingwen Wang、Ji-Rong Wen、Le Wu、Xin Xin、Jun Xu、Dawei Yin、Peng Zhang、Fan Zhang、Weinan Zhang、Min Zhang 和 Xiaofei Zhu。2023。《信息检索与大语言模型相遇：中国信息检索学界的战略报告》。*AI Open* 4 (2023), 80–90。doi:10. 1016 2023 08 001 /J.AIOPEN. . . [2] Hung-Ting Chen、Xiang Liu、Shauli Ravfogel 和 Eunsol Choi。2025。《超越单一嵌入：通过多查询检索捕获多样化目标》。*CoRR* abs/ 2511 02770 2025 10 48550 2511 02770 2511 02770 . ( ). doi: . /ARXIV. . [3] 2025 arXiv: . Jianlyu Chen、Junwei Lan、Chaofan Li、Defu Lian 和 Zheng Liu。 .《ReasonEm- bed：面向推理密集型文档检索的增强文本嵌入》。*CoRR* abs/ 2510 08252 2025 10 48550 2510 08252 2510 08252 . ( ). doi: . / [4] ARXIV. . arXiv: . Jianlv Chen、Shitao Xiao、Peitian Zhang、Kun Luo、Defu Lian 和 Zheng Liu。2023。《BGE M3-Embedding：通过自知识蒸馏实现多语言、多功能、多粒度文本嵌入》。arXiv:2309.07597 [cs.CL][5] Xinghao Chen、Anhao Zhao、Heming Xia、Xuan Lu、Hanlin Wang、Yanjun Chen、Wei Zhang、Jian Wang、Wenjie Li 和 Xiaoyu Shen。2025。《超越语言的推理：潜在思维链推理综合综述》。*CoRR* abs/2505.16782 (2025)。doi:10.48550/ARXIV.2505.16782 arXiv:2505.16782[6] Zijian Chen、Xueguang Ma、Shengyao Zhuang、Ping Nie、Kai Zou、Andrew Liu、Joshua Green、Kshama Patel、Ruoxi Meng、Mingyi Su、Sahel Sharifymoghaddam、Yanxi Li、Haoran Hong、Xinyu Shi、Xuye Liu、Nandan Thakur、Crystina Zhang、Luyu Gao、Wenhu Chen 和 Jimmy Lin。2025。《BrowseComp-Plus：更公平、透明的深度研究智能体评测基准》。*CoRR* abs/ 2508 06600 2025 10 48550 2508 06600 2508 06600 . ( ). doi: . /ARXIV. . [7] arXiv: . Xuanming Cui、Jianpeng Cheng、Hong-you Chen、Satya Narayan Shukla、Ab- hijeet Awasthi、Xichen Pan、Chaitanya Ahuja、Shlok Kumar Mishra、Yonghuan Yang、Jun Xiao、Qi Guo、Ser-Nam Lim、Aashu Singh 和 Xiangjun Fan。2025。《先思考，再嵌入：生成式上下文提升多模态嵌入》。*CoRR* abs/ 2510 05014 2025 10 48550 2510 05014 2510 05014 . ( ). doi: . /ARXIV. . [8] 2025 arXiv: . Debrup Das、Seán Ó Nualláin 和 Razieh Rahimi。 .《RaDeR：具备推理能力的稠密检索模型》。*CoRR* abs/2505.18405 (2025)。doi:10.48550/ARXIV.2505. 18405 arXiv:2505.18405[9] Luyu Gao、Xueguang Ma、Jimmy Lin 和 Jamie Callan。2023。《无需相关性标签的精准零样本密集检索》。载于计算语言学协会第*61*届年会论文集（第*1*卷：长篇论文），*ACL 2023*，加拿大多伦多，*2023*年7月9日至*14*日，Anna Rogers、Jordan L. Boyd-Graber 和 Naoaki Okazaki（编）。计算语言学协会，1762–1777。doi:10.18653/V1/2023.ACL-LONG.99[10] Yunfan Gao、Yun Xiong、Xinyu Gao、Kangxiang Jia、Jinliu Pan、Yuxi Bi、Yi Dai、Jiawei Sun、Qianyu Guo、Meng Wang 和 Haofen Wang。2024。《大语言模型的检索增强生成：综述》。arXiv: 2312 10997 [ ] . cs.CL Yuntao Gui 和 James Cheng。2025。《Search-R3：统一大语言模型中的推理与嵌入生成》。*CoRR* abs/2510.07048 (2025)。[11] doi:10.48550/ARXIV.2510.07048 arXiv:2510.07048[12] Shibo Hao、Sainbayar Sukhbaatar、DiJia Su、Xian Li、Zhiting Hu、Jason Weston 和 Yuandong Tian。2024。《训练大语言模型在连续潜在空间中进行推理》。*CoRR* abs/2412.06769 (2024)。doi:10.48550/ARXIV.2412. 06769 arXiv:2412.06769[13] Xanh Ho、Anh-Khoa Duong Nguyen、Saku Sugawara 和 Akiko Aizawa。2020。《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，*COLING 2020*，西班牙巴塞罗那（线上），*2020*年*12*月*8*日至*13*日，Do- nia Scott、Núria Bel 和 Chengqing Zong（编）。国际计算语言学委员会，6609–6625。doi:10.18653/V1/2020.COLING-MAIN.580[14] Aaron Hurst、Adam Lerer、Adam P Goucher、Adam Perelman、Aditya Ramesh、Aidan Clark、AJ Ostrow、Akila Welihinda、Alan Hayes、Alec Radford 等。2024。《Gpt-4o 系统卡》。*arXiv* 预印本 *arXiv:2410.21276* (2024)。[15] Gautier Izacard、Mathilde Caron、Lucas Hosseini、Sebastian Riedel、Piotr Bo- janowski、Armand Joulin 和 Edouard Grave。2022。《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 (2022)。https://openreview.net/forum?id=jKN1pXi7b0[16] Vladimir Karpukhin、Barlas Oguz、Sewon Min、Patrick Lewis、Ledell Wu、Sergey Edunov、Danqi Chen 和 Wen-tau Yih。2020。用于开放域问答的密集段落检索。载于*EMNLP*。6769–6781。[17] Jacob Devlin、Ming-Wei Chang、Kenton 和 Lee Kristina Toutanova。2019。BERT：用于语言理解的深度双向变换器预训练。载于*NAACL-HLT*。4171–4186。[18] Woosuk Kwon、Zhuohan Li、Siyuan Zhuang、Ying Sheng、Lianmin Zheng、Cody Hao Yu、Joseph Gonzalez、Hao Zhang 和 Ion Stoica。2023。利用PagedAttention高效管理大语言模型服务的内存。载于第*29*届操作系统原理研讨会论文集，*SOSP 2023*，德国科布伦茨，*2023*年*10*月*23*日至*26*日，Jason Flinn、Margo I. Seltzer、Pe-ter Druschel、Antoine Kaufmann 和 Jonathan Mace（编）。ACM，611–626。doi:10.1145/3600006.3613165[19] Chankyu Lee、Rajarshi Roy、Mengyao Xu、Jonathan Raiman、Mohammad Shoeybi、Bryan Catanzaro 和 Wei Ping。2025。NV-Embed：将大语言模型训练为通用嵌入模型的改进技术。载于第十三届国际学习表征会议，*ICLR 2025*，新加坡，*2025*年*4*月*24*日至*28*日。OpenReview.net。https://openreview.net/forum?id=lgsyLSsDRe [20] Patrick S. H. Lewis、Ethan Perez、Aleksandra Piktus、Fabio Petroni、Vladimir Karpukhin、Naman Goyal、Heinrich Küttler、Mike Lewis、Wen tau Yih、Tim Rock-täschel、Sebastian Riedel 和 Douwe Kiela。2020。面向知识密集型自然语言处理任务的检索增强生成。载于神经信息处理系统进展*33*：神经信息处理系统*2020*年会，*NeurIPS 2020*，*2020*年*12*月*6*日至*12*日，线上。https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html [21] Zehan Li、Xin Zhang、Yanzhao Zhang、Dingkun Long、Pengjun Xie 和 Meishan Zhang。2023。通过多阶段对比学习实现通用文本嵌入。*CoRR* abs/2308.03281（2023）。doi:10.48550/ARXIV.2308.03281 arXiv:2308.03281[22] Zhengyang Liang、Meiyu Liang、Wei Huang、Yawen Li 和 Zhe Xue。2024。面向高效跨模态表征学习的预训练多模态大模型动态自适应多尺度蒸馏。*CoRR* abs/2404.10838（2024）。doi:10.48550/ARXIV.2404.10838 arXiv:2404.10838[23] Wenhan Liu、Xinyu Ma、Weiwei Sun、Yutao Zhu、Yuchen Li、Dawei Yin 和 Zhicheng Dou。2025。ReasonRank：以强大推理能力赋能段落排序。*CoRR* abs/2508.07050（2025）。doi:10.48550/ARXIV.2508.07050 arXiv:2508.07050[24] Amir M. Mansourian、Rozhan Ahmadi、Masoud Ghafouri、Amir Mohammad Babaei、Elaheh Badali Golezani、Zeynab Yasamani Ghamchi、Vida Ramezanian、Alireza Taherian、Kimia Dinashi、Amirali Miri 和 Shohreh Kasaei。2025。知识蒸馏综合综述。机器学习研究汇刊 2025（2025）。https://openreview.net/forum?id=3cbJzdR78B [25] Sewon Min、Julian Michael、Hannaneh Hajishirzi 和 Luke Zettlemoyer。2020。AmbigQA：回答开放域歧义问题。载于*2020*年自然语言处理实证方法会议论文集，*EMNLP 2020*，线上，*2020*年*11*月*16*日至*20*日，Bonnie Webber、Trevor Cohn、Yulan He 和 Yang Liu（编）。计算语言学协会，5783–5797。doi:10. 18653/V1/2020.EMNLP-MAIN.466[26] Fengran Mo、Kelong Mao、Ziliang Zhao、Hongjin Qian、Haonan Chen、Yiruo Cheng、Xiaoxi Li、Yutao Zhu、Zhicheng Dou 和 Jian-Yun Nie。2025。对话式搜索综述。*ACM*信息系统汇刊43，6（2025），167:1–167:50。doi:10. 1145/3759453[27] Niklas Muennighoff、Hongjin Su、Liang Wang、Nan Yang、Furu Wei、Tao Yu、Amanpreet Singh 和 Douwe Kiela。2024。生成式表征指令微调。*CoRR* abs/2402.09906（2024）。doi:10.48550/ARXIV.2402.09906 arXiv:2402.09906[28] Letian Peng、Yuwei Zhang、Zilong Wang、Jayanth Srinivasa、Gaowen Liu、Zihan Wang 和 Jingbo Shang。2024。答案即所需：通过回答问题实现指令遵循文本嵌入。载于计算语言学协会第*62*届年会论文集（第*1*卷：长篇论文），*ACL 2024*，泰国曼谷，*2024*年*8*月*11*日至*16*日，Lun-Wei Ku、Andre Martins 和 Vivek Srikumar（编）。计算语言学协会，459–477。doi:10.18653/V1/2024.ACL-LONG.27[29] Xubo Qin、Jun Bai、Jiaqi Li、Zixia Jia 和 Zilong Zheng。2025。TongSearch-QR：用于检索的强化查询推理。*CoRR* abs/2506.11603（2025）。doi:10.48550/ARXIV.2506.11603 arXiv:2506.11603[30] Rulin Shao、Rui Qiao、Varsha Kishore、Niklas Muennighoff、Xi Victoria Lin、Daniela Rus、Bryan Kian Hsiang Low、Sewon Min、Wen-tau Yih、Pang Wei Koh 和 Luke Zettlemoyer。2025。ReasonIR：面向推理任务训练检索器。*CoRR* abs/2504.20595（2025）。doi:10.48550/ARXIV.2504.20595 arXiv:2504.20595[31] Zhenyi Shen、Hanqi Yan、Linhai Zhang、Zhanghao Hu、Yali Du 和 Yulan He。2025。CODI：通过自蒸馏将思维链压缩至连续空间。*CoRR* abs/2502.21074（2025）。doi:10.48550/ARXIV.2502.21074 arXiv:2502.21074[32] Hongjin Su、Howard Yen、Mengzhou Xia、Weijia Shi、Niklas Muennighoff、Han-yu Wang、Haisu Liu、Quan Shi、Zachary S. Siegel、Michael Tang、Ruoxi Sun、Jinsung Yoon、Serca n Ö. Arik、Danqi Chen 和 Tao Yu。2025。BRIGHT：一个现实且

面向推理密集型检索的挑战性基准。载于第十三届国际学习表征会议，*ICLR 2025*，新加坡，*2025*年*4*月*24*日至*28*日。OpenReview.net。https://openreview.net/forum?id=ykuc5q381b[33] Jiashuo Sun、Shixuan Liu、Zhaochen Su、Xianrui Zhong、Pengcheng Jiang、Bowen Jin、Peiran Li、Weijia Shi 和 Jiawei Han。2025。GRACE：通过对比策略优化进行生成式表征学习。*CoRR* abs/2510.04506（2025）。doi:10.48550/ARXIV.2510.04506 arXiv:2510.04506[34] Yu-Che Tsai、Kuan-Yu Chen、Yuan-Chi Li、Yuan-Hao Chen、Ching-Yu Tsai 和 Shou-De Lin。2025。让大语言模型说嵌入语言：通过迭代式对比优化生成文本嵌入。*CoRR* abs/2509.24291（2025）。doi:10.48550/ARXIV.2509.24291 arXiv:2509.24291[35] Liang Wang、Nan Yang、Xiaolong Huang、Binxing Jiao、Linjun Yang、Daxin Jiang、Rangan Majumder 和 Furu Wei。2022。通过弱监督对比式预训练生成文本嵌入。*CoRR* abs/2212.03533（2022）。doi:10.48550/ARXIV.2212. 03533 arXiv:2212.03533[36] Liang Wang、Nan Yang、Xiaolong Huang、Linjun Yang、Rangan Majumder 和 Furu Wei。2024。使用大语言模型改进文本嵌入。载于计算语言学协会第*62*届年会论文集（第*1*卷，*2024*，*2024*，*11*：长篇论文），*ACL* 曼谷，泰国，*8*月-Lun-Wei Ku、Andre Martins 和 Vivek Srikumar（编）。计算语言学协会，11897 11916 10 18653 1 2024 642 –。doi: . /V / .[37] 2023 2 ACL-LONG。Liang Wang、Nan Yang 和 Furu Wei。. Query doc：利用大语言模型扩展查询。载于*2023*年自然语言处理实证方法会议论文集，*EMNLP 2023*，新加坡，*2023*年*12*月*6*日至*10*日，Houda Bouamor、Juan Pino 和 Kalika Bali（编）。计算语言学协会，9414 9423 10 18653 1 2023 585 –。doi: . /V / .[38] EMNLP-MAIN。Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Ed H. Chi、Quoc Le 和 Denny Zhou。2022。思维链提示可激发大语言模型的推理能力。*CoRR* abs/2201.11903（2022）。arXiv:2201.11903 https: //arxiv.org/abs/2201 11903 。Orion Weller、Benjamin Chang、Sean MacAvaney、Kyle Lo、Arman Cohan、Ben-jamin Van Durme、Dawn J. Lawrie 和 Luca Soldaini。[39] 2025。FollowIR：评估并训练信息检索模型遵循指令。载于*2025*年美洲计算语言学协会分会会议：人类语言技术论文集，*NAACL 2025*——第*1*卷：长篇论文，美国新墨西哥州阿尔伯克基，*2025*年*4*月*29*日至*5*月*4*日，Luis Chiruzzo、Alan Ritter 和 Lu Wang（编）。计算语言学协会，11926–11942。doi:10.18653/V1/2025.NAACL-LONG.597[40] Shitao Xiao、Zheng Liu、Peitian Zhang、Niklas Muennighoff、Defu Lian 和 Jian-Yun Nie。2024。C-Pack：通用中文嵌入资源包。载于第*47*届*ACM SIGIR*信息检索研究与发展国际会议论文集，*SIGIR 2024*，美国华盛顿特区，*2024*年*7*月*14*日至*18*日，Grace Hui Yang、Hongning Wang、Sam Han、Claudia Hauff、Guido Zuccon 和 Yi Zhang（编）。ACM，641–649。doi:10.1145/3626772.3657878[41] Seunghan Yang、Juntae Lee、Jihwan Bang、Kyuhong Shim、Minsoo Kim 和 Simyung Chang。2025。学习上下文检索以实现稳健的对话式搜索。载于*2025*年自然语言处理实证方法会议论文集，Christos Christodoulopoulos、Tanmoy Chakraborty、Car-olyn Rose 和 Violet Peng（编）。计算语言学协会，中国苏州，11991–12003。doi:10.18653/v1/2025.emnlp-main.602[42] Zhilin Yang、Peng Qi、Saizheng Zhang、Yoshua Bengio、William Cohen、Ruslan Salakhutdinov 和 Christopher D. Manning。2018。HotpotQA：用于多样化、可解释多跳问答的数据集。载于*EMNLP*。计算语言学协会，比利时布鲁塞尔，2369–2380。doi:10.18653/v1/D18-1259[43] Sijia Yao、Pengcheng Huang、Zhenghao Liu、Yu Gu、Yukun Yan、Shi Yu 和 Ge Yu。2025。ExpandR：在大语言模型指导下训练稠密检索器处理查询以外内容。载于*2025*年自然语言处理实证方法会议论文集，Christos Christodoulopoulos、Tanmoy Chakraborty、Carolyn Rose 和 Violet Peng（编）。计算语言学协会，中国苏州，19036 19054 10 18653 1 2025 963 –。doi: . /v / .[44] 2024 emnlp-main。Dun Zhang 和 FulongWang。. Jasper和Stella：SOTA嵌入模型的蒸馏。*CoRR* abs/2412.19048（2024）。doi:10.48550/ARXIV.2412.19048 arXiv:2412.19048[45] Yichi Zhang、Jun Bai、Zhixin Cai、Shuhan Qin、Zhuofan Chen、Jinghua Guan 和 Wenge Rong。2025。你的稠密检索器其实是高效推理器。*CoRR* abs/2510.21727（2025）。doi:10.48550/ARXIV.2510.21727 arXiv:2510.21727[46] Yanzhao Zhang、Mingxin Li、Dingkun Long、Xin Zhang、Huan Lin、Baosong Yang、Pengjun Xie、An Yang、Dayiheng Liu、Junyang Lin、Fei Huang 和 Jingren Zhou。2025。Qwen3 E mbedding：通过基础模型推进文本嵌入与重排序。*CoRR* abs/2506.05176（2025）。doi:10.48550/ARXIV.2506.05176 arXiv:2506.05176[47] Wayne Xin Zhao、Kun Zhou、Junyi Li、Tianyi Tang、Xiaolei Wang、Yupeng Hou、Yingqian Min、Beichen Zhang、Junjie Zhang、Zican Dong、Yifan Du、Chen Yang、Yushuo Chen、Zhipeng Chen、Jinhao Jiang、Ruiyang Ren、Yifan Li、Xinyu Tang、Zikang Liu、Peiyu Liu、Jian-Yun Nie 和 Ji-Rong Wen。2023。大语言模型综述。*CoRR* abs/2303.18223（2023）。doi:10.48550/ARXIV.2303.18223 arXiv:2303.18223

[48] Ziliang Zhao、Changle Qu、Zhicheng Dou、Haonan Chen 和 Jiajie Jin。2025。为对话式搜索系统中的澄清生成检索覆盖意图的示例。载于第*31*届 *ACM SIGKDD* 知识发现与数据挖掘大会论文集，第*2*卷，*KDD 2025*，加拿大安大略省多伦多，*2025*年*8*月*3*日至7日，由 Luiza Antonie、Jian Pei、Xiaohui Yu、Flavio Chierichetti、Hady W. Lauw、Yizhou Sun 和 Srinivasan Parthasarathy（编）。ACM，3992–4001。doi:10.1145/3711896.3737104[49] Yutao Zhu、Huaying Yuan、Shuting Wang、Jiongnan Liu、Wenhan Liu、Chenlong Deng、Zhicheng Dou 和 Ji-Rong Wen。2023。大语言模型与信息检索：综述。*CoRR* abs/2308.07107（2023）。doi:10.48550/ARXIV. 2308.07107 arXiv:2308.07107