通过强化学习编译扩散实现高效的属性对齐扇出检索

Pengcheng Jiang[1,2,*], Judith Yue Li[1], Moonkyung Ryu[1], R. LilyHu[1], Kun Su[1], Zhong YiWan[1], Liam Hebert[1], Hao Peng[1], Jiawei Han[2],DimaKuzmin[1] 和 Craig Boutilier[1]1Google Research, 2伊利诺伊大学厄巴纳-香槟分校

arXiv:2603.06397v1 [cs.IR] 6 Mar 2026

许多现代检索问题都具有集合值特性：给定宽泛的意图，系统必须返回一组结果，在基于固定数据库的同时优化高阶属性（例如，多样性、覆盖度、互补性、连贯性）。集合值目标通常不可分解，现有的监督式（查询、内容）数据集也无法体现这些目标，因为它们只关注 Top-1 检索。因此，人们常采用扇出检索 生成多样化的子查询，以检索物品集合。尽管强化学习（RL）可以通过交互优化集合级目标，但在推理时部署经过强化学习调优的大语言模型进行扇出检索，成本高得难以承受。另一方面，基于扩散的生成式检索能够在嵌入空间中高效实现单次扇出，但需要与目标对齐的训练目标。为解决这些问题，我们提出 R4T（检索用于训练），采用强化学习一次作为目标转换器，并分三步进行：（i）使用复合集合级奖励训练扇出大语言模型；（ii）合成与目标一致的训练对；（iii）训练轻量级扩散检索器，以建模集合值输出的条件分布。在由精选物品集合构成的大规模时尚与音乐基准测试中，我们证明，与强基线方法相比，R4T 能够提升检索质量，同时将查询时扇出延迟降低一个数量级。

1. 引言

检索系统越来越需要返回结果集，而非单个最佳匹配。在许多现实应用中，所需输出是一组能够共同满足高阶属性的结果：搜索界面可以将宽泛查询扩展为多个意图，以提高覆盖度；推荐系统可以生成兼具多样性与连贯性的推荐列表；捆绑系统则可以检索互补物品，共同满足复杂查询的需求。这些要求推动了生成式检索的相关方法，其中候选项由模型生成，而非直接选取（例如仅依据最近邻相似度）（Deffayet 等人，2023年；Tay 等人，2022；Tomasi 等人，2025年）。

一个核心挑战在于，许多此类问题都是集合值且不可分解的。不同于只有一个可标注正确物品的经典检索任务，扇出检索往往没有唯一的真实标注：对于同一宽泛意图，许多不同的物品集合都可能有效。因此，检索质量由集合级属性（例如多样性、意图覆盖、互补性和风格连贯性）来界定，同时还要求严格依据目标数据库。

强化学习（*RL*）是通过与数据库进行奖励驱动交互来优化此类行为的自然框架（Jiang 等人，2025b；Zhang等人，2025a），近期研究表明，强化学习能够塑造超越逐点相关性的检索策略（Jiang 等人，2025a、c）。然而，在推理时直接部署经过强化学习训练的语言模型往往不切实际：自回归扇出生成需要反复调用检索，导致延迟显著增加；而集合级奖励可能存在较高方差，也容易受到捷径利用的影响。这些问题使稳定部署更为困难，并促使人们将奖励驱动优化与推理时检索分离。

与此同时，基于扩散的生成式检索

*此项工作是在 Google Research 担任学生研究员期间完成的。

能够在嵌入空间中高效进行非自回归采样（Bao 等人，2025年；Guinot 等人，2025年；Tomasi 等人，2025年）。这使检索从繁重的“系统2”自回归生成转向并行的“系统1”采样，从而能够一次生成整个结果列表或结果包。通过将检索建模为去噪过程，这些系统可以一次生成整个结果列表或结果包。然而，扩散检索器严重依赖大量的属性对齐的训练目标，而这恰恰是在不可分解的集合值检索任务中稀缺或含义模糊的资源。

本文针对这一瓶颈展开研究：检索目标日益丰富，而在这些目标下训练高效扇出检索器所需的监督信号却不足。我们的核心思路是，不将强化学习用作部署后的推理机制，而是将其作为一次性的目标转换器，把复杂的集合级奖励规范转化为可扩展的监督信号，用于训练检索器。具体而言，我们提出了**R4T（检索用于训练）**，一种分三个阶段运行的方法：(1) 使用强化学习训练一个由语言模型实现的扇出策略，并通过复合奖励明确编码所需的集合级属性；(2) 利用优化后的策略，通过收集并筛选成功轨迹，合成目标一致的训练数据；(3) 使用这些合成数据训练一个轻量级、单次执行的生成式检索器。在我们的具体实现中，该检索器是一种连贯的基于嵌入的扩散模型，可在推理时执行高效、可控的扇出检索。

贡献。 我们有三项主要贡献。首先，我们提出一种通用框架，可将针对集合值、不可分解检索目标优化奖励所产生的行为，编译为用于监督训练的数据。其次，我们使用 Soft-GRPO 实例化这一框架，以实现扇出策略优化，并通过连贯的、基于嵌入的扩散，在推理时进行单次生成。第三，我们在两种不同场景中展示了 R4T 的有效性：开放式抽象检索（*OAR*），其中不存在真实标注，质量由奖励所指定的集合级属性来定义，这些属性作用于集合级输出；以及弱监督组合检索（*WSCR*），其中每个查询都对应多个有效物品集合，监督信号则通过弱参考集提供。在这两种场景中，R4T 均能提升检索质量，优于强大的扇出基线，同时保持生产级部署所需的高效推理。

2. 方法：R4T

我们引入**R4T**（Retrieve-for-Train，检索用于训练），这是一种用于在集合值不可分解目标下训练生成式检索器的框架。其核心思想是仅一次 将强化学习（RL）用作目标转换器；我们不在推理时部署经过强化学习调优的语言模型，而是利用强化学习发现与奖励对齐的扇出行为，并将其转换为可扩展的监督信号，以训练轻量级、单次处理的检索器。

问题设定：集合值扇出检索。我们考虑这样一种场景：给定一个宽泛查询 q，我们的检索系统从固定数据库 D 中返回一个集合 结果。为处理检索的集合性质，我们采用扇出形式：策略首先生成 k 子查询 $Q = \{q_1, \ldots, q_k\}$，然后由固定检索器 $R(\cdot)$ 执行每个子查询，以检索候选结果。令 $C_i = R(q_i, D)$ 表示子查询 q_i 所检索到的候选结果，并令 $R(Q) = \bigcup_{i=1}^{k} C_i$ 表示扇出结果的并集。关键在于，我们假设结果质量由集合级属性定义，例如多样性、覆盖度或互补性；这些属性通常不可分解，而且还允许存在多个（不唯一的）正确结果。后一个性质给真实标注监督带来了挑战。

2.1. R4T 概述

R4T 由三个主要阶段组成：第一阶段是强化学习策略优化，我们在此阶段训练一个扇出语言模型（*FOLM*） π_θ，使其生成子查询，让这些子查询与数据库的交互能够最大化某种特定任务的集合级奖励（下文将讨论各种奖励）。第二阶段是合成监督信号，我们运行优化后的 FOLM 策略，收集高奖励轨迹，并将其转换为集合-

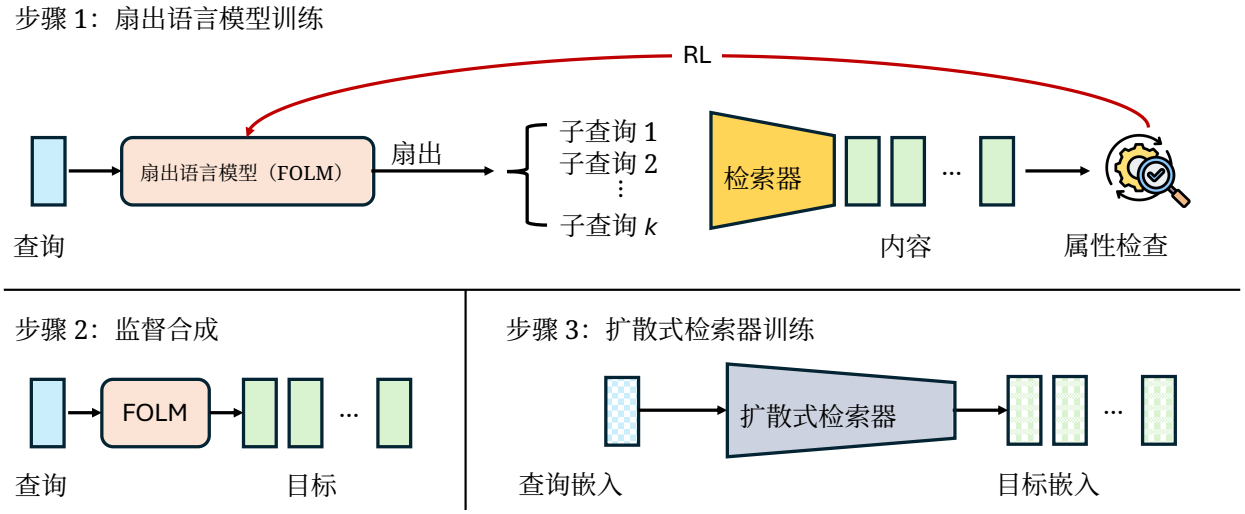


图 1 | **R4T 概览**。步骤 1（§2.3）使用强化学习训练扇出语言模型（FOLM），以生成属性对齐的子查询。步骤 2（§2.4）使用训练后的 FOLM 合成（q, c）监督数据。步骤 3（§2.5）训练基于扩散的扇出检索器，直接从查询嵌入中采样内容嵌入。

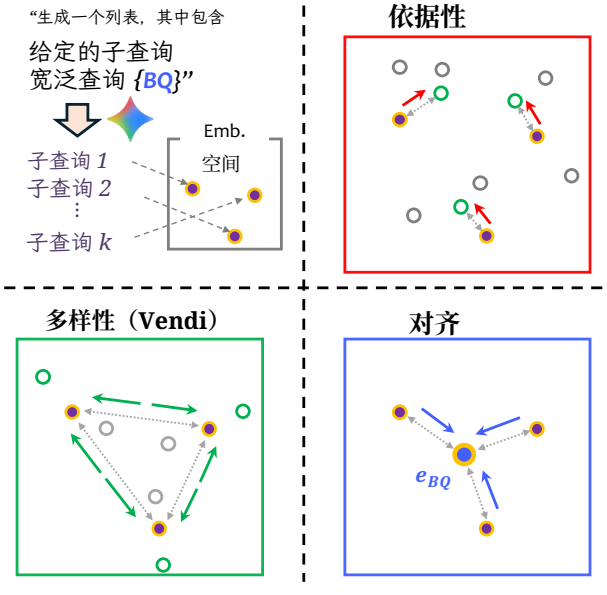


图 2 | OAR 中使用的奖励示意图。

值目标。在最后阶段，通过编译部署训练一个轻量级生成式检索器，以建模给定查询时集合值目标的条件分布，从而在推理时实现高效的单次扇出。下文将详细介绍各个阶段（图1 展示了示意图）。

两种部署变体：R4T-FOLM 与 R4T-Diffusion。R4T 将奖励驱动的扇出行为发现与高效的推理时扇出分离。因此，我们评估了两种部署变体：它们采用相同的强化学习优化阶段，但在测试时执行扇出的方式不同。

R4T-FOLM（强化学习调优的扇出语言模型）。我们直接部署经强化学习优化的 FOLM π_{θ^*}，在推理时生成 k 子查询 $Q = \{q_1, \ldots\ldots, q_k\} \sim \pi_{\theta^*}(\cdot \mid q)$，并将检索结果聚合为 $\mathrm{R}(Q) = \bigcup_{i=1}^{k} R(q_i, \mathrm{D})$。该变体体现了奖励优化扇出的质量，但需要承担自回归生成和重复检索调用的开销。

R4T-Diffusion（经强化学习蒸馏的扩散检索器）。我们将π_{θ^*} 的奖励对齐扇出行为蒸馏到轻量级扩散模型 D_ϕ 中，并使用合成监督信号进行训练（第2.4节）。在推理时，D_ϕ 通过单次非自回归遍历生成 L检索方向，$\mathbf{Z}^0 \sim p_\phi(\mathbf{Z} \mid z_q)$，再通过最近邻检索将其映射到数据库内容。该变体用于评估在高效的单次遍历推理下，能否保留经奖励优化的扇出行为。

比较这些变体，可以将奖励优化扇出的收益与

部署成本区分开来，直接检验强化学习发现的行为能否提炼出来，以实现高效推理。

2.2. 集合级目标与奖励

R4T 的一个关键组成部分，是能够将集合级目标指定为奖励，用于 FOLM 的策略优化（阶段 1）。我们研究了两种常见的集合值检索需求场景。

2.2.1. 开放式抽象检索

开放式抽象检索（*OAR*）面向开放式探索，不存在唯一的真实结果集。给定一个宽泛查询 q（例如某个主题或场景），目标是检索一组集合级或抽象结果，这些结果应当（i）具有多样性（即涵盖多种解读），（ii）与 q高度契合，并且（iii）以数据库为依据。

我们将生成的扇出集合 Q 上的复合奖励 R_{abs} 定义如下：

$$\mathcal{R}_{\text{abs}}(q,Q)=\lambda_g r_{\text{ground}}(Q)+\lambda_d r_{\text{div}}(Q)+\lambda_a r_{\text{align}}(q,Q). \tag{1}$$

该奖励采用子奖励 $r$ground、 $r$div 和$r$align，分别衡量依据性、多样性和对齐度，每项奖励按照其混合权重（$lambda$）进行贡献。

为了衡量结果集的多样性，我们使用 Vendi 分数（Friedman 和 Dieng，2022）来计算扇出的语义广度，计算对象是每个子查询的代表项（例如，每个子查询检索到的排名第一的项目），即

$$r_{\text{div}}(Q)=\text{Vendi}(\{e_{\text{content}}(c_i^\star)\}_{i=1}^k), \tag{2}$$

其中， $c_i^\star$ 是 q_i的代表性检索项目，$e$content(·) 是内容编码器，用于嵌入数据库条目。

为确保依据性，我们鼓励子查询停留在（由嵌入空间诱导的）数据库流形上，方法是惩罚每个子查询嵌入与其在数据库中的最近邻之间的距离，即，

$$r_{\text{ground}}(Q)=1-\frac{1}{k}\sum_{i=1}^{k}\min_{c\in\mathcal{D}}\|e_{\text{text}}(q_i)-e_{\text{content}}(c)\|_2, \tag{3}$$

其中，$e$text（q_i）是查询嵌入。

最后，我们通过以下奖励将每个子查询锚定到 q ，以防止其偏离原始查询 q 的意图，从而实现对齐：

$$r_{\text{align}}(q,Q)=\frac{1}{k}\sum_{i=1}^{k}\cos(e_{\text{text}}(q_i),e_{\text{text}}(q)). \tag{4}$$

我们默认设定为 $\lambda_g=0$。6 和 $\lambda_d=\lambda_a=0$。2 。图 2展示了 OAR 任务所用的奖励。

2.2.2. 弱监督组合检索

弱监督组合检索（*WSCR*）针对的是这样一种设定：每个查询可以对应许多有效的物品集合，但我们只有弱参考集，它们代表一种合理的实现。每个查询 q 与一个参考集合Y ={y_1，……，y_m}配对；重要的是，并不假定Y是唯一正确的查询响应。

我们为扇出集合 Q 定义基于参考集合的覆盖度奖励如下：

$$\mathcal{R}_{\text{set}}(q,Q;\mathcal{Y})=\frac{|\mathcal{Y}\cap\mathcal{R}(Q)|}{|\mathcal{Y}|}. \tag{5}$$

该奖励鼓励策略生成互补的子查询，使其并集覆盖Y中所反映的不同语义成分，而不是记忆某个固定目标。

2.3. 通过Soft-GRPO进行扇出语言模型训练

R4T 的第一阶段是对扇出语言模型（FOLM）进行 RK 训练。给定输入查询q，FOLM π_θ 会生成一组候选子查询 $Q=\{q_1,\ldots,q_k\}$。这些子查询由冻结的稠密检索器执行 $R(\cdot)$ 以获取候选文档 $C_i=R(q_i,\mathcal{D})$。我们使用相关任务目标（定义见第 2.2 节）为每个采样的扇出输出赋予奖励。

为了优化 π_θ，我们采用组相对策略优化（*GRPO*）。对于查询 q，我们采样一组 G 个输出 $\{o_1, \ldots, o_G\}$，并使用组统计量计算优势：

$$A_i = \frac{r_i - \mu_G}{\sigma_G + \epsilon}, \quad \mu_G = \frac{1}{G}\sum_{j=1}^{G} r_j. \tag{6}$$

软PPO正则化。为稳定开放式生成，我们采用 软*PPO* (Becker等人，2025年；Liang等人，2025年；Zhang等人，2025b)，它在当前策略 π_θ 与采样策略πold之间加入前向和反向KL惩罚，从而增强GRPO：

$$\begin{aligned}\mathcal{J}(\theta) = \mathbb{E}_{\pi_{\text{old}}}\big[\mathcal{L}_{\text{GRPO}} - \beta_1 \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{old}}) \\ - \beta_2 \mathbb{D}_{\text{KL}}(\pi_{\text{old}} \| \pi_\theta)\big].\end{aligned} \tag{7}$$

在实践中，我们将其实现为：

$$\begin{aligned}\mathcal{L}(\theta) = \mathbb{E}_t\Big[&- \min(\rho_t A_t, \text{clip}(\rho_t, 1-\epsilon, 1+\epsilon)A_t) \\ &+ \beta_1 \cdot \rho_t(\log \pi_\theta(o_t) - \log \pi_{\text{old}}(o_t)) \\ &+ \beta_2 \cdot (-\log \pi_\theta(o_t))\Big],\end{aligned} \tag{8}$$

其中，$\rho_t = \pi_\theta/\pi$old是重要性比率。

2.4. 合成监督

R4T的第二阶段涉及生成一个用于训练检索模型的合成数据集。FOLM π_{θ^*} 充当行为生成器，通过奖励塑造扇出检索模式的分布。对于每个查询 q，我们从 π_{θ^*} 采样扇出输出，并针对固定数据库执行检索。我们使用FOLM生成的内容

作为监督信号，使下游模型能够学习上述强化学习过程所诱导的完整行为分布。

为了训练一个能在单次前向传播中生成多个检索嵌入的下游模型，我们将每个扇出结果表示为一个连贯的目标张量$\mathbf{Z}_{\text{target}} \in \mathrm{R}^{L\times d}$。$\mathbf{Z}^{\text{target}}$ 的每一行对应 FOLM 发现的一个检索方向。$\mathbf{Z}$ target 的构造取决于优化目标：

对于OAR任务，目标强调产生多样且有依据的检索结果。因此，我们构建$\mathbf{Z}^{\text{target}}$ 从与扇出输出相对应的检索内容$\{z_{c_1}, \ldots\ldots, z_{c_L}\}$ 的嵌入出发。该表述直接蒸馏出由强化学习训练的策略所诱导的、数据库依据的集合上的分布。

对于 WSCR 任务，优化目标着重于发现互补的搜索方向，以共同覆盖一个参考集合。在这种情况下，我们根据 FOLM 生成的优化后子查询的嵌入构造$\mathbf{Z}^{\text{target}}$ ，即 $\{e\text{text}(q_1), \ldots\ldots, e\text{text}(q_L)\}$。这种选择使下游模型能够内化通过强化学习学到的搜索分解策略。

由于集合值检索与顺序无关，我们在训练 $\mathbf{Z}^{\text{target}}$ 时随机打乱其各行，以增强对排列的鲁棒性。由此得到的合成数据集为 $\mathrm{T}^{\text{syn}} = \{(z_q, \mathbf{Z}^{\text{target}})\}$，它捕获了奖励优化策略所诱导的完整扇出检索分布。

2.5. 单次扇出扩散

RT4 的最后阶段是训练一个生成式检索器 D_ϕ，以建模 $p(\mathbf{Z}_{\text{target}} \mid z_q)$。我们采用 方差爆炸（*VE*）扩散 表述，即 Song 等人 （2020）在 EDM 框架（Karras 等人，2022）中的表述。去噪器$D_\phi(\mathbf{Z}_t;\, \sigma, z_q)$ 经训练，使用该损失从噪声输入中恢复干净目标

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\sigma,\epsilon}\Big[\lambda(\sigma) \cdot \|D_\phi(\mathbf{Z}_{\text{target}} + \sigma\epsilon;\, \sigma, z_q) - \mathbf{Z}_{\text{target}}\|^2\Big], \tag{9}$$

其中 $\lambda(\sigma) = (\sigma^2 + \sigma^2\,\text{data})/(\sigma \cdot \sigma\text{data})^2$ 用于平衡不同噪声水平的贡献。

架构。去噪器采用 Transformer 架构（Peebles 和 Xie，2023年；Tomasi 等人，2025年），专为结构连贯性设计，并采用 Karras 等人（2022）提出的预条件方案。查询嵌入 z_q 通过交叉注意力注入。训练时，我们通过随机丢弃 无分类器引导（*CFG*）（Ho 和 Salimans，2022） z_q 。推理时，我们求解概率流随机微分方程以生成$\mathbf{Z}_0$, ，再将其切分为L 嵌入，并通过最近邻检索映射到数据库内容。

	Polyvore				音乐			
	依据性	多样性	对齐	平均值	依据性	多样性	对齐	平均值
无扇出	22.4	34.4	21.4	26.1	48.8	20.0	41.8	36.9
Gemini-2.5-Flash								
零样本	24.0	47.0	23.6	31.5	45.8	45.2	44.4	45.1
N选最佳	26.1±1.6	52.2±2.3	25.2±1.9	34.5±1.8	48.2±2.1	48.4±2.6	49.0±2.0	48.5±2.2
Gemma3-4B								
零样本	28.4	56.0	31.2	38.5	49.8	42.6	51.8	48.1
N选最佳	28.9±1.5	61.0±2.1	32.7±1.7	40.9±1.6	51.4±2.0	43.2±2.3	53.0±1.9	49.2±2.1
R4T-FOLM	**30.8**±1.9	**76.8**±2.7	**39.8**±2.3	49.1±2.1	**63.1**±2.2	**49.2**±2.4	**62.0**±2.0	58.1±2.2
R4T-Diffusion	\	74.3±3.4	37.6±2.8	\	\	46.7±3.1	59.6±2.6	\
Qwen3-4B								
零样本	23.8	37.0	23.4	28.1	42.0	38.8	41.2	40.7
N选最佳	27.0±1.8	40.3±2.5	24.0±1.9	30.4±1.7	44.0±2.2	40.3±2.7	43.7±2.1	42.7±2.3
R4T-FOLM	**37.0**±2.1	62.8±2.9	**28.0**±2.0	42.6±2.2	**48.2**±2.3	**44.8**±2.5	49.4±2.1	47.5±2.2
R4T-Diffusion	\	**65.0**±3.6	27.4±2.7	\	\	44.5±3.0	**52.0**±2.8	\

表 1 | **任务 1 的性能：开放式摘要检索（OAR）**。结果以均值 ± 标准差报告。**粗体** 表示各模型系列中的最佳性能。R4T 在不同数据集和指标上均持续优于扇出基线方法。请注意，由于不存在中间子查询，R4T-Diffusion 不适用依据性指标（以“\”表示）。

	Recall@5K	Hit@5K	VS
基于大语言模型的检索			
Gemini-2.5-Flash	15.7	52.1	33.4
Gemma3-4B	6.0	25.9	44.2
Qwen3-4B	10.1	33.9	46.4
R4T-FOLM（Gemma）	16.9	54.4	40.5
R4T-FOLM（Qwen）	**20.9**	**64.6**	**27.5**
基于扩散的检索			
R4T-Diffusion（Gemma）	15.0	54.1	46.2
R4T-Diffusion（Qwen）	**16.5**	**57.5**	**34.7**

表 2 | **任务2的表现：弱监督组合检索（WSCR）**。我们使用大语言模型（Gemini-2.5-Pro）为 Polyvore 中的每套服装搭配生成一个宽泛查询，并将每套搭配中的物品设为真实标注。为使用归一化 Vendi 分数评估多样性，我们将基于大语言模型的检索温度设为=0.9，并为所有方法设置 # samples=5 （即总计 5×10=50 个子查询）。

3. 实验

3.1. 实验设置

我们在 OAR（无真实标注；属性定义的质量）和 WSCR（弱参考集；面向覆盖率）任务上评估 R4T，目标函数定义见第2.2节。

数据集。我们在两个多样化的真实世界数据集上评估 R4T，这些数据集反映了不同的检索模态（文本到图像、文本到音乐）和领域。第一个是**Polyvore**，一个大规模时尚基准（Han等人，2017），由用户策划的穿搭组成。每套穿搭都作为一个真实标注物品集，包含不同类别中相互搭配的时尚产品（例如上装、下装、鞋类、配饰）。该数据集提供丰富的多模态特征，包括产品图像和文本元数据，我们用它们评估多模态落地性和集合连贯性。任务1和任务2的候选检索池规模分别为 21,888（合集/摘要）和 142,472（物品）。第二个数据集**音乐**是专有的工业数据集，由专家生成的音乐播放列表组成。每个播放列表都代表一组连贯的曲目序列，可作为检索的真实标注集。该数据集用于评估模型在音乐领域建模主题一致性和意图覆盖的能力。任务1的候选检索池规模为 8,522（播放列表嵌入）。宽泛查询生成的详情见附录C。

检索骨干网络。我们采用针对特定数据集的嵌入骨干网络，以在共享嵌入空间中实现高效扇出

Model	Sub Query	Retrieved Images
R4T	bohemian festival dress	
R4T	straw boots festival style	
R4T	lace bohemian festival	
Qwen3_4b	bohemian festival style	
Qwen3_4b	bohemian festival fashion	
Qwen3_4b	festival bohemian clothes	

图 3 | 针对开放式抽象检索任务，在给定宽泛查询 “波西米亚节日风格”时，于 Polyvore 数据集上的定性比较。R4T 生成语义各异且切题的子查询，从而检索到多样化的穿搭集合；而 Qwen3-4B 零样本基线生成的子查询大多只是改述，导致结果更加同质化。

。对于 Polyvore，我们使用基于 CLIP 的图文编码器，并通过套娃式表征学习进行训练（Kusupati 等人，2022），它支持多模态检索，同时允许灵活截断嵌入。除非另有说明，在主要实验中，我们使用 128 维嵌入，以平衡检索准确率与效率。对于音乐数据集，我们采用 MuLan（Huang 等人，2022），这是一种在大规模音频语言对上训练的音乐—文本联合嵌入模型，并直接在 MuLan 嵌入空间中进行检索。

基线方法 我们将 R4T 与三种基线方法进行比较。**无扇出** 基线直接使用原始查询检索$n \times k$ 内容项，不进行任何子查询扩展。这代表了传统的稠密检索方法，并作为扇出方法的性能下界。第二种基线方法，**零样本扇出**使用基础语言模型（强化学习训练前）生成 k子查询，且不进行任何任务特定的优化。每个子查询检索 n 个物品，从而得到 $n \times k$ 个物品总数。我们测试了几种零样本扇出基线的变体，采用不同的语言模型生成子查询：(a)**Gemini-2.5-Flash** （Comanici 等人，2025年），一种具备强大零样本能力的大规模专有模型；(b)**Gemma3-4B**（Team 等人，2025年），一种规模较小的开源模型（40亿参数）；以及 (c)**Qwen3-4B** （Yang 等人，2025年），另一种具有竞争力的开源模型（40亿参数）。**N选最佳**是我们的第三种基线方法，这是一种强基线，使用零样本模型执行扇出 N 次，并根据训练奖励选择最佳结果。我们设定 $N = 5$。

我们比较了所提方法的两种变体，**R4T-（FOLM/扩散）**。R4T 首先使用任务特定奖励通过强化学习训练扇出语言模型，合成训练数据，并训练基于扩散的检索器。在推理时，扩散检索器直接从查询嵌入中采样 k 内容嵌入。对于所有扇出基线方法和 R4T，我们报告了使用 $k = 10$ 个子查询时的结果。

评估指标。我们使用针对具体任务的指标评估集合值检索。对于不存在真实集合的 OAR任务，我们采用大语言模型即评审员的评估方式，使用 Gemini-2.5-Pro 评估以下三个维度：集合多样性、查询-集合对齐和依据性。每个维度均按 5 点李克特量表评分，遵循先前研究的做法；该研究表明，这种评估与人工判断高度相关（Zheng 等人，2023年）。

对于 WSCR任务，我们报告基于参考集的覆盖度指标（Recall@5K 和 Hit@5K），并结合 Vendi 分数衡量多样性和生成稳定性。由于参考集代表的是查询意图的一种合理实现，而非穷尽性的真实标注，因此召回率被视为语义覆盖率的代理指标，而非二元正确性指标。

完整的评估流程、提示模板和指标定义见附录B、D和E。

3.2. 主要结果

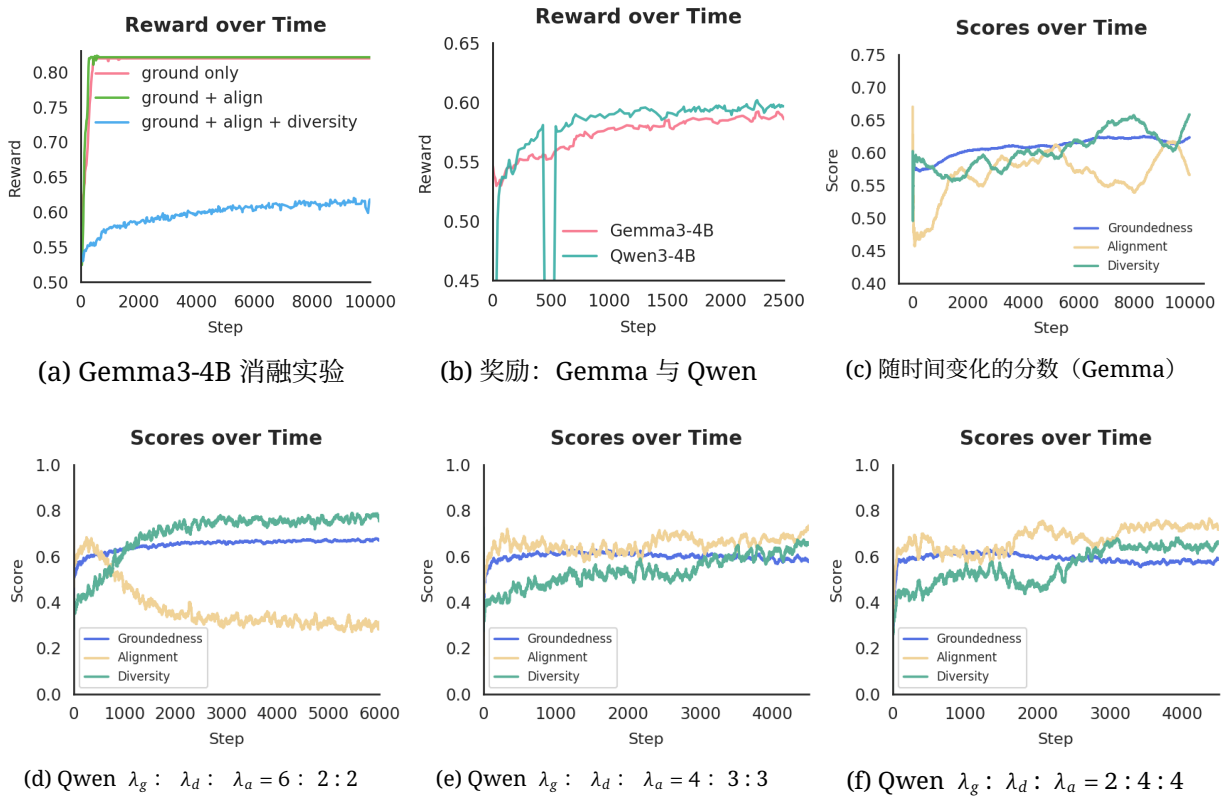
表1 和表2 总结了所有方法在 OAR 和 WSCR 任务上的表现。我们强调了若干一致的规律，这些规律阐明了现有扇出策略的优势与局限，以及 R4T 的优势。

扇出是必要条件，但并非充分条件。 在所有数据集和基础模型上，零样本扇出始终优于直接使用原始查询进行检索的无扇出基线。这证实了长期以来的假设：查询扩展能够提高对多样化语义维度的覆盖度（Carpineto 和 Romano，2012）。然而，零样本扇出也存在明显的弱点。生成的子查询往往会偏离数据库的语义空间，或退化为近乎重复的表达，导致依据性有限，检索结果冗余。这些影响在 OAR 任务中尤为明显，因为没有真实标注来约束探索。

奖励感知选择能够提升质量，但难以扩展。Best-of-N 基线通过选择高奖励的扇出实例，显著优于零样本扇出。这表明我们设计的奖励函数是塑造检索行为的有效信号。然而，Best-of-N 要求针对每个查询执行多次独立的扇出操作，导致推理成本增加一个数量级。因此，其性能提升是以延迟和可扩展性为代价的，限制了它在实际系统中的适用性。

R4T 始终在准确率—效率前沿占据优势。 R4T 在各数据集、任务和基础模型上均取得了最强的整体表现。值得注意的是，R4T 的表现优于 Best-of-N，同时部署时只需进行一次推理。此结果凸显了 R4T 的核心优势：通过强化学习一次性发现高质量的扇出行为，再将这些行为蒸馏到轻量级扩散检索器中。经过强化学习训练的 FOLM 学会生成兼顾多样性、对齐和依据性的子查询，而扩散模型则在嵌入空间中忠实地捕捉这一分布，从而实现高效且可控的扇出检索。

OAR 中的奖励黑客行为。精心设计奖励对于实现 R4T 的性能提升至关重要。我们在OAR任务上进行了消融实验（图4），以分析不同奖励组成部分如何影响扇出行为。缺少多样性项时，奖励很容易被利用。仅使用依据性奖励进行训练时，策略会收敛到退化且毫无意义的字符串（例如“行尾 行尾 行尾”），因为这类字符串恰好能最小化与某个特定数据库条目的嵌入距离。同时使用依据性和对齐进行训练时，策略崩溃得更快，因为它只需重复原始查询的释义，就能轻易最大化对齐，同时忽略语义离散度。相比之下，联合优化依据性、对齐和多样性可防止这种崩溃，并实现稳定的 GRPO



(a) Gemma3-4B 消融实验　(b) 奖励：Gemma 与 Qwen　(c) 随时间变化的分数（Gemma）

(d) Qwen $\lambda_g:\ \lambda_d:\ \lambda_a = 6:2:2$　(e) Qwen $\lambda_g:\ \lambda_d:\ \lambda_a = 4:3:3$　(f) Qwen $\lambda_g:\ \lambda_d:\ \lambda_a = 2:4:4$

图 e 4 | 开放式抽象检索的奖励和分数随时间的变化（FOLM 训练）.

训练。

图4 （d–f）进一步展示了奖励加权如何影响训练动态。当依据性占主导时，多样性会提高，但对齐程度会下降，表明过分强调数据库邻近性是以牺牲语义保真度为代价的。相反，强调对齐和多样性会抑制探索，限制对替代解释或意图的覆盖。均衡加权可使三个分数稳定收敛，凸显了探索与语义保真度之间固有的权衡。总之，这些结果表明，多样性和对齐互为制衡因素，促使策略进入奖励空间中的均衡区域；在该区域中，捷径解难以奏效，只有生成有意义且以数据库为依据的子查询变体，才能获得高奖励。

集合检索行为。表2揭示了不同检索范式在基于参考的覆盖率与多样性之间存在明显的权衡。在大语言模型方法中，较高的 Recall@5K 和 Hit@5K 往往伴随着较低的 Vendi 分数，这表明，当自回归扇出以参考重叠度为优化目标时，检索结果往往集中在少数几个占主导地位的语义模式上。这一现象与近期关于语言模型强化学习的研究结果一致：更强的奖励优化通常会降低输出熵，并促使模式坍塌（Cui 等人，2025年）。相比之下，零样本大语言模型基线和基于扩散的检索表现出更高的多样性，反映出它们在多次推理运行中对语义空间进行了更广泛的探索。

重要的是，在这些情况下，召回率较低并不一定意味着检索质量较差。参考集合并非“穷尽性的真实标注”，而只是代表查询意图可能存在的诸多合理实现中的一种。因此，多样性更高表明模型可能发现了参考项未涵盖的其他同样有效的解读。R4T 使这种权衡朝着更有利的方向转变。虽然 R4T-FOLM 提高了

参考覆盖度，但代价是多样性降低；R4T-Diffusion 则保留了基于扩散的生成所固有的大部分多样性，同时大幅提升了覆盖度。这表明，将与奖励对齐的行为蒸馏到扩散先验中，为弱监督下的集合值检索提供了一种切实可行的方法，能够平衡覆盖度与语义丰富性。

3.3. 定性示例

图3 和图 6展示了 OAR 任务在 Polyvore 上针对两个宽泛查询的定性结果：“波西米亚节日风格”和“劳动节野餐穿搭”。这些示例说明，联合使用依据性、对齐和多样性奖励，如何引导 R4T 生成有意义且多样的子查询分解。对于波西米亚节日风格查询，R4T 生成的子查询会分支到不同的主题方向，例如“波西米亚节日连衣裙”、“草编靴节日风格”和“蕾丝波西米亚节日风”。每个子查询检索到的物品在视觉上协调一致，但语义各不相同，表明其具有较强的多样性，同时仍与整体风格保持一致。相比之下，Qwen-4B 等基线模型往往会生成几乎同义的表达变体，例如“波西米亚节日风格”和“波西米亚节日时尚”，从而导致检索结果更加同质化。对于“劳动节野餐穿搭”查询，也观察到了类似趋势。R4T 将查询分解为互补的方面，例如波西米亚风格、极简风格和连身裤风格，检索到的服装搭配在廓形、色彩搭配和配饰选择上各不相同，同时又适合季节性野餐场景。Gemini-2.5-Flash 虽然能够生成句法形式多样的子查询，但其检索结果往往更加趋同，语义覆盖度也较为有限。

总体而言，这些定性结果凸显了三种奖励组成部分在塑造检索行为方面的作用。依据性促使每个子查询都对应数据库中的真实物品，对齐则确保其与原始查询意图保持一致，而多样性则推动对多种有效解读进行探索。这些目标共同使 R4T 能够生成丰富的查询分解，从而实现多样且相关的检索。

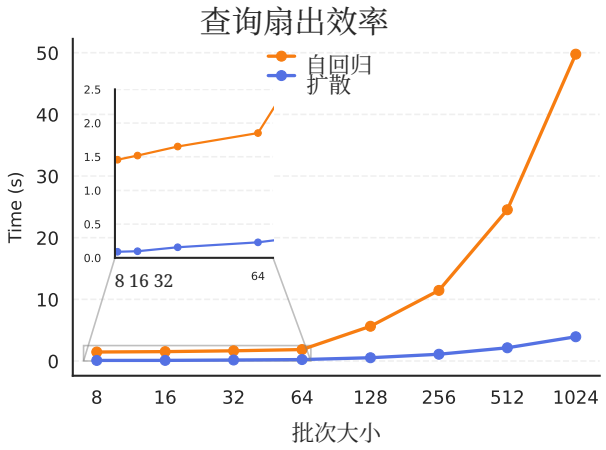


图 5 | **自回归大语言模型与扩散模型在查询扇出（$k = 10$）生成方面的效率比较。**请注意，x 轴采用对数刻度（每一步翻倍）。扩散模型展现出更强的可扩展性和更低的延迟：小批次时性能保持在亚秒级，批次大小较大时则实现了数量级的加速。

3.4. 效率分析

为评估推理时效率，我们对 R4T 的查询扇出延迟进行基准测试，并将其与基于自回归大语言模型的扇出基线进行比较。我们比较了一种基于扩散的检索器（它能通过一次前向传播生成所有 $k = 10$ 检索方向）与自回归模型，后者会依次生成子查询并调用检索。我们在不同批次大小下测量实际运行时间延迟。

图5 展示了扇出生成所需的实际耗时。自回归大语言模型在较小批次大小下受固定开销影响严重：即使批次大小为 8，也需要约 1.46 秒；随后耗时随批次大小线性增长，在批次大小为 1024 时接近 50 秒。相比之下，我们的扩散模型仅包含 53.9M 个参数，利用嵌入空间中的非自回归生成，仅需 0.07 秒即可处理小批次，处理最大批次也只需 4.21 秒。这种紧凑架构

大幅降低了部署内存开销，同时实现了稳定的 12×-20× 倍加速，这证实了将扇出负担从沉重的自回归系统2转变为轻量级的系统1扩散先验，对于构建实用的实时检索应用至关重要。

4. 相关工作

生成式检索。 近期研究探索了将检索建模为生成任务。DSI（Tay 等人，2022）训练序列到序列模型，使其直接根据查询生成文档标识符（docids），并将整个语料库视为模型记忆。NCI（Wang 等人，2022）通过改进索引策略对此进行了扩展，而 SEAL（Bevilacqua 等人，2022）则生成 n 元语法以进行段落检索。近期研究还探索了语义标识符（Li 等人，2023年）、生成式检索中的学习排序（Li 等人，2024年）以及面向大型语料库的可扩展方法（Pradeep 等人，2023年；Zeng 等人，2024年）。我们的工作不同之处在于，它在嵌入空间而非离散令牌空间中进行生成，从而实现更平滑的插值和多模态检索。与以往需要大量人工标注监督的生成式检索系统不同，我们通过强化学习合成训练数据。

查询扩展与改写。查询扩展在信息检索领域有着悠久的历史（Azad 和 Deepak，2019年；Carpineto 和 Romano，2012），经典技术包括伪相关反馈（Rocchio，1971）和词项共现分析。现代方法利用神经语言模型进行查询改写（Nogueira 等人，2019），近期研究则探索基于大语言模型的查询生成（Ma 等人，2023年）以及针对复杂信息需求的多查询分解。然而，这些方法通常采用先检索后重排序流程，且不会明确地以目标数据库为依据来扩展查询。我们的扇出方法有所不同：它通过强化学习训练查询扩展，以优化数据库特定的检索属性。

用于检索的强化学习。 近期研究已开始将强化学习应用于信息检索任务。DeepRetrieval（Jiang等人，2025a年）以及 s3（Jiang 等人，2025c年）利用强化学习训练与真实搜索引擎交互的搜索智能体，证明奖励驱动优化能够编码细粒度的检索目标。Zhou 等人（2023年）探索了利用相关性反馈进行生成式检索的强化学习。然而，这些方法将强化学习用于直接推理，导致查询时计算成本高昂。相比之下，我们将强化学习用作数据生成引擎，通过一次性合成训练数据，将成本分摊至各个推理查询，随后部署高效的基于扩散的检索器。

用于检索的扩散模型。扩散概率模型（Ho 等人，2020；Song 等人，2020）已在多个领域的生成式建模中展现出优异性能。近期研究已将扩散模型应用于检索任务。Diff4Steer（Bao 等人，2025年）将扩散先验与语义引导用于音乐检索，而 GDRetriever（Guinot 等人，2025年）开发了可控生成式文本-音乐检索。这些研究表明，在嵌入空间中采样能够实现灵活的多假设生成。我们的工作通过解决训练数据瓶颈扩展了这一范式：我们不依赖人工监督，而是通过强化学习引导的生成来合成属性对齐训练数据。

合成数据生成。 合成数据生成已成为解决机器学习中数据稀缺和隐私问题的有力技术（Goyal 和 Mahmoud，2024年；Lu 等人，2023年）。相关技术涵盖基于规则的系统以及深度生成模型，包括生成对抗网络（Goodfellow 等人，2014）和变分自编码器（Kingma 和 Welling，2013）。近期研究探索了利用大语言模型为各类任务生成训练数据（Chen 等人，2023年），包括指令微调和特定任务微调。在检索领域，研究者已生成合成查询来训练稠密检索器（Ma 等人，2021年），但通常缺乏特定属性控制。我们的工作则独特地结合了基于强化学习的数据合成与基于扩散的检索，从而能够以可控方式生成属性对齐训练对。

多模态检索。跨模态检索系统（Wang 等人，2025年）旨在弥合语义

差距，以连接文本、图像和音频等不同模态。CLIP（Radford 等人，2021年）和 CLAP（Wu 等人，2023年）等视觉语言模型会学习用于检索的联合嵌入空间。近期研究探索了多模态检索的生成式方法（Wu 等人，2024年），包括用于多模态任务的检索增强生成。我们的框架基于冻结的多模态编码器运行，证明强化学习引导的数据合成能够在不修改底层嵌入空间的情况下提升跨模态检索效果。

5. 结论

我们提出了 R4T，这是一种基于强化学习的框架，用于合成生成式检索的训练数据。通过使用属性特定奖励训练扇出语言模型，并利用该模型为基于扩散的检索器生成监督信号，我们无需人工标注数据即可实现可控扇出检索。在时尚产品基准上的实验表明，R4T 的表现优于强大的基线方法，同时保持了高效的推理能力。我们的工作为在监督信号稀缺的专业领域训练生成式检索系统开辟了新的方向。局限性详见附录 A。

影响声明

本研究致力于解决日益丰富的检索目标与训练高效检索系统所能获得的有限监督信号之间不断扩大的差距。通过将强化学习用作目标转换器，并将由此产生的行为编译到轻量级扩散检索器中，R4T 无需人工标注数据即可实现可控的集合值检索。这一设计具有若干潜在的积极影响。

从系统视角来看，R4T 为部署检索模型提供了一条切实可行的路径，使其能够在保持较低推理延迟的同时，优化多样性、覆盖度和互补性等高阶属性。这对于现实应用尤为重要，例如推荐系统、创意搜索和探索式信息获取等场景中，扇出检索颇具价值，但自回归生成的成本却高得难以承受。通过将奖励驱动的发现与推理时部署分离，我们的框架无需反复进行在线优化，即可支持可扩展且可定制的检索。

从研究视角来看，这项工作提出了一种通用范式，可将不可分解的集合级目标转化为可训练的监督信号。利用强化学习合成目标对齐训练数据的想法，或许可以从检索扩展至其他真实标注含糊或带有主观性的结构化生成任务，例如规划、设计和创意生成。我们希望这能推动人们进一步探索将交互式学习与高效生成模型相结合的编译式方法。

伦理考量与社会影响。由于检索行为受到明确定义的奖励函数影响，若奖励设定不慎，可能会将训练数据中已有的偏见编码其中或加以放大，进而导致推荐系统或内容发现等应用中的呈现不公平。基于强化学习的合成方法可能会通过合成训练数据大规模传播此类偏见。尽管我们的实验聚焦于时尚、音乐等无害领域，但相同技术也可能应用于检索偏见会造成重大后果的敏感情境。负责任的部署需要领域特定偏见审计、包容性设计实践以及适当的监督机制。我们认为，R4T 是一种用于受控检索设计的工具，必须配合保障措施使用，而不能取代人类判断和伦理监督。

参考文献

H. K. Azad 和 A. Deepak。用于信息检索的查询扩展技术：综述。《信息处理与管理》，56(5)：1698–1735，2019 年。

X. Bao、J. Y. Li、Z. Y. Wan、K. Su、T. Denk、J. Lee、D. Kuzmin 和 F. Sha。Diff4steer：可操控的

用于生成式音乐检索并带有语义引导的扩散先验。在 *ICASSP 2025- 2025*年*IEEE*国际声学、语音与信号处理会议 *(ICASSP)*，第 1–5 页。IEEE，2025年。

P. Becker、N. Freymuth、S. Thilges、F. Otto 和 G.Neumann。Troll：信赖域改进了大语言模型的强化学习。*arXiv* 预印本 *arXiv:2510.03817*，2025年。

M. Bevilacqua、G. Ottaviano、P. Lewis、S. Yih、S. Riedel 和 F. Petroni。自回归搜索引擎：将子字符串生成为文档标识符。神经信息处理系统进展，35:31668–31683，2022。C. Carpineto 和 G. Romano。信息检索中自动查询扩展综述。*ACM*计算调查，44(1):1–50，2012。J. Chen、L. Chen、H. Huang 和 T. Zhou。ChatGPT 何时需要思维链提示？*arXiv* 预印本 *arXiv:2304.03262*，2023。G. Comanici、E. Bieber、M. Schaekermann、I. Pasupat、N. Sachdeva、I. Dhillon、M. Blistein、O. Ram、D. Zhang、E. Rosen 等。Gemini 2.5：凭借先进推理、多模态、长上下文和新一代智能体能力，推动前沿发展。*arXiv* 预印本 *arXiv:2507.06261*，2025。G. Cui、Y. Zhang、J. Chen、L. Yuan、Z. Wang、Y. Zuo、H. Li、Y. Fan、H. Chen、W. Chen 等。推理语言模型中强化学习的熵机制。*arXiv* 预印本 *arXiv:2505.22617*，2025。R. Deffayet、T. Thonet、J.-M. Renders 和 M. de Rijke。基于强化学习的生成式推荐列表推荐。载于第十六届 *ACM* 国际网络搜索与数据挖掘会议论文集，WSDM '23，第 580–588 页。ACM，2023 年 2 月。doi: 10.1145/3539597.3570412。URL http://dx.doi.org/10.1145/3539597.3570412。D. Friedman 和 A. B. Dieng。Vendi 分数：一种用于机器学习的多样性评估指标。*arXiv* 预印本 *arXiv:2210.02410*，2022 年。

I. Goodfellow、J. Pouget-Abadie、M. Mirza、B. Xu、D. Warde-Farley、S. Ozair、A. Courville 和 Y. Bengio。生成对抗网络。神经信息处理系统进展，27，2014 年。

M. Goyal 和 Q. H. Mahmoud。使用生成式人工智能的合成数据生成技术系统综述。*Electronics*，13(17)：3509，2024年。

J. Guinot、E. Quinton 和 G. Fazekas。Gd-retriever：基于扩散模型的可控生成式文本-音乐检索。*arXiv* 预印本 *arXiv:2506.17886*，2025。X. Han、Z. Wu、Y.-G. Jiang 和 L. S. Davis。利用双向长短期记忆网络学习时尚搭配。载于第 *25* 届 *ACM* 国际多媒体会议论文集，第 1078–1086 页，2017。J. Ho 和 T. Salimans。无分类器扩散引导。*arXiv* 预印本 *arXiv:2207.12598*，2022。J. Ho、A. Jain 和 P. Abbeel。去噪扩散概率模型。神经信息处理系统进展，33：6840–6851，2020。Q. Huang、A. Jansen、J. Lee、R. Ganti、J. Y. Li 和 D. P. Ellis。MuLan：音乐音频与自然语言的联合嵌入。*arXiv* 预印本 *arXiv:2208.12415*，2022。P. Jiang、J. Lin、L. Cao、R. Tian、S. Kang、Z. Wang、J. Sun 和 J. Han。DeepRetrieval：通过强化学习，利用大语言模型攻克真实搜索引擎和检索器。*arXiv* 预印本 *arXiv:2503.00223*，2025a。P. Jiang、J. Lin、Z. Shi、Z. Wang、L. He、Y. Wu、M. Zhong、P. Song、Q. Zhang、H. Wang 等。智能体人工智能的适应。*arXiv* 预印本 *arXiv:2512.16301*，2025b。P. Jiang、X. Xu、J. Lin、J. Xiao、Z. Wang、J. Sun 和 J. Han。s3：通过强化学习训练搜索智能体，无需那么多数据。*arXiv* 预印本 *arXiv:2505.14146*，2025c。

T. Karras、M. Aittala、T. Aila 和 S. Laine。阐明基于扩散的生成模型的设计空间。神经信息处理系统进展，35：26565–26577，2022。

D. P. Kingma 和 M. Welling。自动编码变分贝叶斯。*arXiv* 预印本 *arXiv:1312.6114*，2013。

A. Kusupati、G. Bhatt、A. Rege、M. Wallingford、A. Sinha、V. Ramanujan、W. Howard-Snyder、K. Chen、S. Kakade、P. Jain 等。套娃式表征学习。神经信息处理系统进展，35:30233–30249，2022。

Y. Li、N. Yang、L. Wang、F. Wei 和 W. Li。多视图标识符增强的生成式检索。*arXiv* 预印本 *arXiv:2305.16675*，2023年。

Y. Li、N. Yang、L. Wang、F. Wei 和 W. Li。生成式检索中的学习排序。载于人工智能促进协会会议论文集，第 38 卷，第 8716–8723 页，2024年。

C.Liang, D. Huang, C. Yang, X. Yang, A. Li, X. Yan 和 Simply 贡献者。Simply：一项加速并自动化人工智能研究的实验。GitHub 仓库，2025年。URL https://github.com/google-deepmind/simply。

Y. Lu, M. Shen, H. Wang, X. Wang, C. van Rechem, T. Fu 和 W. Wei。机器学习在合成数据生成中的应用：综述。*arXiv* 预印本 *arXiv:2302.04062*，2023年。

J. Ma, I. Korotkov, Y. Yang, K. Hall 和 R. Mc- Donald。通过面向特定领域的合成问题生成实现零样本神经段落检索。载于第 *16* 届欧洲计算语言学协会会议论文集：主会场，第 1075–1088 页，2021年。

X. Ma、Y. Gong、P. He、H. Zhao 和 N. Duan。面向检索增强大语言模型的查询改写。*arXiv* 预印本 *arXiv:2305.14283*，2023年。R. Nogueira、W. Yang、J. Lin 和 K. Cho。通过查询预测扩展文档。在*arXiv* 预印本 *arXiv:1904.08375*，2019年。

W. Peebles 和 S. Xie。使用 Transformer 的可扩展扩散模型，2023年。网址 https://arxiv.org/abs/2212.09748。R. Pradeep、K. Hui、J. Gupta、A. Lelkes、H. Zhuang、J. Lin、D. Metzler 和 V. Tran。生成式检索如何扩展到数百万个段落？载于*2023*年自然语言处理实证方法会议论文集，第 1305–1321 页，2023年。

A. Radford、J. W. Kim、C. Hallacy、A. Ramesh、G. Goh、S. Agarwal、G. Sastry、A. Askell、P. Mishkin、J. Clark 等。通过自然语言监督学习可迁移视觉模型。载于国际机器学习会议，第8748–8763页。机器学习研究会论文集，2021年。

J. J. Rocchio。信息检索中的相关性反馈。*SMART*检索系统：自动文档处理实验，第313–323页，1971年。

Y. Song、J. Sohl-Dickstein、D. P. Kingma、A. Kumar、S. Ermon 和 B. Poole。通过随机微分方程进行基于分数的生成建模。*arXiv*预印本 *arXiv:2011.13456*，2020。

Y. Tay、V. Tran、M. Dehghani、J. Ni、D. Bahri、H. Mehta、Z. Qin、K. Hui、Z. Zhao、J. Gupta 等。将变换器记忆作为可微分搜索索引。神经信息处理系统进展，35:21831–21843，2022。

G. Team、A. Kamath、J. Ferret、S. Pathak、N. Vieillard、R. Merhej、S. Perrin、T. Matejovicova、A. Ramé、M. Rivière 等人。Gemma 3 技术报告。*arXiv* 预印本 *arXiv:2503.19786*，2025年。

F. Tomasi、F. Fabbri、J. Carter、E. Kalomiris、M. Lalmas 和 Z. Dai。提示词到候选列表：用于提示词条件化候选列表生成的扩散模型，2025年。URL https://arxiv.org/abs/2408.06883。T. Wang、F. Li、L. Zhu、J. Li、Z. Zhang 和 H. T. Shen。跨模态检索：方法与未来方向的系统综述。*IEEE* 会议论文集，2025年。

Y. Wang、Y. Hou、H. Wang、Z. Miao、S. Wu、Q. Chen、Y. Xia、C. Chi、G. Zhao、Z. Liu 等。用于文档检索的神经语料库索引器。神经信息处理系统进展，35：25600–25614，2022。

S. Wu、W. Wei、M. Zhang、Z. Chen、J. Ma、Z. Ren、M. de Rijke 和 P. Ren。生成式检索即多向量稠密检索。载于第*47*届 *ACM* 信息检索研究与发展国际会议论文集，第1828–1838页，2024年。

Y. Wu、K. Chen、T. Zhang、Y. Hui、T. Berg- Kirkpatrick 和 S. Dubnov。通过特征融合和关键词到标题的增强，进行大规模对比式语言-音频预训练。载于*ICASSP 2023-2023 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第1–5页。IEEE，2023年。

A. Yang、A. Li、B. Yang、B. Zhang、B. Hui、B. Zheng、B. Yu、C. Gao、C. Huang、C. Lv 等。通义千问3技术报告。*arXiv* 预印本 *arXiv:2505.09388*，2025年。

H. Zeng、C. Luo、B. Jin、S. M. Sarwar、T. Wei和H. Zamani。可扩展且有效的生成式信息检索。见*2024*年*ACM*万维网大会论文集，第1441–1452页，2024年。

K. Zhang、Y. Zuo、B. He、Y. Sun、R. Liu、C. Jiang、Y. Fan、K. Tian、G. Jia、P. Li等。面向大型推理说明模型的强化学习综述。*arXiv*预印本 *arXiv:2509.08827*，2025年a。

Y. Zhang、Y. Liu、H. Yuan、Y. Yuan、Q. Gu和A. C.-C. Yao。关于面向大语言模型推理说明的KL正则化策略梯度算法的设计。*arXiv*预印本 *arXiv:2505.17508*，2025年b。

L. Zheng、W.-L. Chiang、Y. Sheng、S. Zhuang、Z. Wu、Y. Zhuang、Z. Lin、Z. Li、D. Li、E. Xing 等。使用 MT-Bench 和聊天机器人竞技场评判大语言模型作为评审。神经信息处理系统进展，36:46595–46623，2023年。

Y. Zhou、Z. Dou 和 J.-R. Wen。通过相关性反馈强化学习增强生成式检索。载于会议论文集*2023*年自然语言处理实证方法会议，第12481–12490 页，2023年。

A. 局限性

尽管 R4T 效果显著，但它仍存在若干局限，这些局限也指出了未来的研究方向。

首先，强化学习阶段需要与冻结检索器反复交互，并显式计算奖励。虽然这部分成本可在部署时摊销，但对于规模极大或频繁变化的数据库，前期训练开销可能相当可观。未来的工作可以探索样本效率更高的强化学习方法、部分检索器更新或离线近似方法，以降低这部分成本。

其次，R4T 假设所需的检索属性能够通过显式奖励函数得到合理表达。尽管我们证明了结合依据性、对齐和多样性能够带来稳定的行为，但某些用户偏好（例如对创造力、新颖性或文化敏感性的主观理解）可能难以编码为标量奖励。通过用户反馈或偏好数据学习奖励函数，是解决这一局限的一个有前景的方向。

第三，我们的评估部分依赖大语言模型作为评审来评估开放式检索质量。尽管先前的工作表明，这类评估结果与人工判断高度相关，但仍可能引入评审模型继承的偏见，也未必能充分捕捉检索实用性的所有定性方面。纳入人工评估或采用混合评估协议，将有助于加强未来的研究。

最后，我们对 R4T 的具体实现聚焦于一种特定的扇出架构和基于扩散的检索器。尽管该框架具有通用性，实证结果仍可能取决于基础语言模型、嵌入空间和扩散架构的选择。将 R4T 扩展到其他检索骨干网络、模态和任务设置，仍是有待探索的开放领域。

总体而言，我们认为 R4T 是在复杂目标下对集合值检索系统进行可扩展训练的初步尝试，而非解决所有形式的可控检索问题的完整方案。

B. 评估指标

B.1. LLM 作为评审

对于开放式摘要检索（OAR），检索质量由集合级属性定义，因此不存在唯一的真实目标。为此，我们采用基于 Gemini-2.5-Flash 的 LLM 作为评审进行评估，该模型支持多模态输入（文本、图像和音频）。研究表明，这种评估范式与人类判断在检索质量评估方面具有良好的相关性（Zheng 等人，2023年）。

我们评估三个维度：

1. **集合多样性：**检索到的合集在多大程度上涵盖了宽泛查询的不同语义解释。2. **查询—集合对齐：**检索到的合集整体上反映原始查询意图的程度。3.**依据性：**每个检索到的合集与其生成该合集的子查询对应得有多准确。

对于每个维度，大语言模型评审器使用 5 点李克特量表打分（1=较差， 2=一般，3=良好， 4=很好， 5=优秀），并给出简短的理由。为减少评估偏差，我们随机排列项目顺序，采用明确的评估标准，要求每项评分都附带解释，并避免在提示中提及检索方法。完整的提示模板见附录E。

B.2. WSCR 指标

对于弱监督组合检索（WSCR），我们使用基于参考的覆盖率指标和多样性度量来评估检索质量。每个查询都关联一个参考项集合，该集合代表预期语义的一种合理实现，而非穷尽性或唯一正确的目标。因此，基于召回率的指标被视为语义覆盖率的代理指标，而非严格的正确性标准。

Recall@5K. Recall@5K衡量在候选范围内检索到的参考项所占比例

池。我们为每个k = 10 子查询（或生成的嵌入）检索500个候选项，从而在每次推理轮次中得到一个包含5,000个物品的池。

Hit@5K。Hit@5K 衡量至少有一个参考项出现在检索候选池中的查询所占百分比，从而反映参考实现的任一语义侧面是否得到恢复。

Vendi 分数。为评估多样性和生成稳定性，我们对每个查询执行 N = 5次独立推理运行。对于每次运行，我们通过对各个子查询嵌入取平均值来计算一个代表性嵌入。Vendi 分数（Friedman 和 Dieng，2022）基于这些嵌入计算，用于量化不同运行之间的语义差异。

C. 宽泛查询生成

为评估**R4T**在多种检索场景下的表现，我们构建了两类合成宽泛查询，分别对应本文研究的两项任务：(1)**开放式摘要检索（OAR）**和 (2) **弱监督组合检索（WSCR）**。本附录介绍各项任务所用的查询构建流程。

C.1. OAR：宽泛查询生成

OAR 的宽泛查询是开放式、探索性的文本提示，用于描述总体主题、风格或场景，而不指定具体物品。这些查询旨在引出多样化的集合级检索输出，并允许多种有效解读，这与 OAR 中缺乏真实标注监督相一致。

我们采用多模板生成策略，以确保对语义意图和查询风格的广泛覆盖度。查询由大语言模型生成，并使用多样化的提示模板，针对抽象的不同方面，包括：美学主题（例如，“波西米亚节日风格”）、基于活动的场景（例如，“周末早午餐穿搭”）、季节性语境（例如，“夏日度假氛围”）、生活方式身份（例如，“可持续生活”）、情绪表达（例如，“自信而大胆”）、文化趋势（例如，“田园风美学”）、功能需求（例如，“旅行必需品”）以及混合组合（例如，“复古与现代交融”）。

OAR 的宽泛查询生成提示

系统：你是一个用于检索系统的创意查询生成器。

用户：为时尚／音乐检索系统生成 100 条多样化、有创意的宽泛查询。每条查询应为 2–6 个词，描述一个总体主题、美学或场景。

要求：

- 查询应具有开放性和探索性
- 应聚焦于高层次主题，而非具体物品
- 使用自然、对话式的语言
- 避免使用品牌名称或具体产品引用
- 每个查询都应允许多种有效解读

仅返回查询字符串组成的 JSON 数组。

生成并去重后，我们为 OAR任务获得了**43,874 个唯一的宽泛查询**。这些查询按照 8:1:1 的比例随机划分为训练集、验证集和测试集。对于每个查询，我们使用温度为 0.9、样本量为 128 的扇出语言模型（FOLM）生成训练监督信号，为扩散模型训练合成目标。

C.2. WSCR 的宽泛查询生成

WSCR 的宽泛查询构建方式与 OAR 有根本不同。每个 WSCR 查询并非独立生成，而是源自一个已有的项目集合，作为对用户意图某种合理组合实现的弱参考式描述。

具体而言，我们从精选物品集合（例如 Polyvore 中的服装搭配）开始，每个集合都由多个互补物品组成。给定集合中的物品名称后，我们提示大语言模型

生成简洁的、穿搭级查询，以概括该集合的整体风格、美学或场合，而不列举单个物品。所得查询充当物品组成的高层语义抽象，而非唯一正确的标签。

此过程会生成形式为（q，Y）的配对，其中 q 是一个宽泛查询，Y 是一个合理的参考物品集。重要的是，Y不被视为真实标注：许多其他物品集也可能同样符合 q的语义。因此，这些查询集对构成了适用于 WSCR 的弱监督数据。

WSCR：从物品集生成宽泛查询

系统： 你是一个用于服装搭配级检索的查询生成器。

用户： 根据以下服装搭配中的时尚单品，生成一个宽泛的服装搭配级搜索查询，用来描述整套搭配的整体风格或主题。

查询应满足以下要求：

- 2–8 个词
- 宽泛且通用（不要列出具体物品详情）
- 聚焦于整体风格、美学或场合
- 适合用于搜索相似的服装搭配

物品：

```
- [item name 1]
- [item name 2]
...
```

仅返回 查询文本。

训练集和测试集增强。

WSCR：物品集增强与查询生成

系统： 你是穿搭组合器和查询生成器。

用户： 给定一份时尚单品列表，从中选出多个集合，每个集合恰好包含 10 件物品。每个集合应构成一套风格鲜明、协调一致的服装搭配。针对每个集合，生成一条宽泛的服装搭配层面查询（2–8 个词），描述整体风格或场合。

约束：

- 每个集合必须恰好包含 10 件物品
- 各集合应在风格或场合上有所不同
- 查询应较为宽泛，且不得逐一列举物品细节

以结构化格式返回所选的物品索引及相应的查询。

为提高覆盖度并减少对特定参考实现的过拟合，我们进一步利用语言模型引导的结构化物品集重组来增强 WSCR 数据。

- **训练集增强**。 对于每个原始物品集，我们构建一个包含 50 个物品的候选池，其中包括原始集合中的物品以及从全局物品池中随机采样的物品。我们提示语言模型通过一次调用选择多个（8 个）彼此不同的 10 件物品子集，每个子集都构成具有不同风格的连贯服装搭配。对于每个子集，生成一个相应的宽泛查询。

- **测试集增强**。 对于每个测试物品集，我们采样多个种子子集，其中包含原始物品的 3–60%。对于每个种子子集，语言模型从一个包含 30 个物品的候选池中选择其他物品，以构成完整的 10 件物品服装搭配，并为每套搭配配以一个宽泛查询。该过程能够针对单个查询生成多个合理的组合实现，同时保留与原始集合的部分重叠。

在所有数据划分中，该流程为 WSCR 生成了**84,704 个不重复的宽泛查询**。随后，按 8:1:1 的比例将所得数据集划分为训练集/验证集/测试集。与 OAR 中一样，通过在训练集上以温度

0。9 运行扇出语言模型（FOLM），并为每个查询采样 128 个样本，从而生成扩散模型的训练目标。

总体而言，这种构造确保 WSCR 在弱参考式监督下评估检索效果，其成功标准是组合覆盖度，而非精确的集合匹配。

D. 查询扇出提示

面向开放式抽象检索（OAR）的查询扇出

你是一名查询编写器。给定一个宽泛查询，你的任务是创建一组查询，这些查询既要彼此多样，又要涵盖与宽泛查询相同的主题（最好增添或更改一到三个词）。这些查询将用于搜索时尚数据集。

你应该在<think> </think> 标签中展示你的思考过程。你必须在<queries> </queries> 标签中以 JSON 格式返回最终的查询列表。

例如：

```
<think>
[thinking process]
</think>
<queries>
[
    "query1",
    "query2",
    "query3",
    ... (up to 10 queries)
]
</queries>
```

宽泛查询是：

```
<broad_query>
{broad_query}
</broad_query>
```

弱监督组合检索（WSCR）的查询扇出

你是一名查询编写器。你的任务是针对我给出的宽泛查询创建一份查询列表，以便从时尚数据集中检索一组物品来组成服装搭配。

你应在<think> </think> 标签中展示你的思考过程。你必须在<queries> </queries>标签中以 JSON 格式返回最终查询列表。

例如：

```
<think>
[thinking process]
</think>
<queries>
[
    "query1",
    "query2",
    "query3",
    ... (up to 10 queries)
]
</queries>.
```

宽泛查询为：

```
<broad_query>
{broad_query}
</broad_query>
```

E. LLM 评审评估提示词

本节详细介绍了用于 LLM 作为评审评估的提示词。这些提示词经过精心设计，通过多种策略尽可能减少偏见并最大限度地提高公平性：

1. **随机化：** 物品以随机顺序呈现，以避免位置偏差 2. **客观标准：** 我们提供清晰、可衡量的评估维度 3. **解释要求：** 评审者必须说明分数的推理依据 4. **方法无关：** 不提及检索方法或系统 5. **校准示例：** 我们提供锚点示例以进行分数校准 6. **结构化输出：** JSON 格式可确保分数提取一致

OAR - 集合多样性评估

系统提示词：
您是一位专家评估者，负责评估检索到的内容项的多样性。您将看到一组物品（产品图像或文本描述），并需要评估它们在语义内容、主题和类别方面的多样性。

用户提示词：
`Original Query:{query}Retrieved Items:` [图像/文本以随机顺序呈现]

请按 1–5 的等级评估此检索结果集的**多样性**：

评估标准：

- **5 (优秀)：** 多样性极佳。 [图像/音频] 内容全面且富有创意地探索了查询，同时冗余极少。
- **4 (非常好)：** 多样性很强。 [图像/音频] 大多各不相同，并充分涵盖了查询的不同方面。
- **3 (良好)：** 多样性尚可。 [图像/音频] 探索了查询的多个不同想法或方面。虽然存在一些重叠，但可以接受。（这是预期的“良好”结果。）
- **2 (一般)：** 有一些重复。 [图像/音频]展现了 2-3 个不同的想法，但许多内容存在冗余。
- **1 (较差)：** 冗余度很高。 [图像/音频]都非常相似，只探索了一个想法。

响应格式：

```
{
  "score": <1-5>,
  "reasoning": "<brief explanation of your score>",
  "categories_observed": ["<category 1>", "<category 2>",  ...]
}
```

重要提示： 请关注语义多样性，而非视觉或风格上的相似性。考虑这些物品是否探索了原始查询的不同方面或解读。

F. 实现细节

扇出语言模型（FOLM）。 我们使用标准的指令微调检查点初始化扇出语言模型。训练采用带软PPO正则化的GRPO算法，并将KL惩罚直接应用于每个 token 的损失。我们优化策略，使其生成 k 个子查询，以最大化任务特定的集合级奖励，包括多样性、依据性和对齐。我们使用 1×10^{-7}的学习率和 512 的全局批量大小。训练在 TPUv6e-16 Ghostlite Pod 上进行（由 16 个加速器核心组成）。

监督合成。 经过强化学习优化后，我们使用训练好的 FOLM 合成与目标一致的训练对。对于每个查询，我们以 0.9 的温度生成 128 个样本以捕捉扇出行为的奖励塑造分布。此数据生成阶段在 TPUv6e-4 Ghostlite Pod（4 个加速器核心）上完成。

扩散训练。 扩散模型采用适用于连续输入的Transformer骨干网络，其输入为目标嵌入的拼接序列 $Z_{target} \in \mathrm{R}^{L\times d}$。该模型在EDM 框架内使用方差爆炸（VE）公式直接预测 Z_0。我们采用隐藏维度为1024、MLP维度为1024的扩散Transformer，并以预热余弦衰减计划在 10×10^6 步上优化。在推理阶段，模型使用SDE求解器以256步单次生成 L嵌入，再通过最近-

OAR - 查询-集合对齐评估

系统提示词：
你是一名专家评估者，负责评估一组检索到的 [图像/音频] 在*整体上*与原始查询的契合程度。用户提示词：你将看到子查询及其代表性[图像/音频]。请评估这**整组 [图像/音频]** 对**原始查询**的代表性。

用户提示词：
`Original Query: {query}Retrieved Items: [图像/音频以随机顺序呈现]`

请评估此 [图像/音频] 集合与**原始查询** 的**整体对齐**，评分范围为1-5：

评估标准：

- **5（优秀）：** [图像/音频] 与查询完全相关，完美地捕捉了原始查询的全部含义和意图。
- **4（非常好）：** 高度相关。所有 [图像/音频] 都与查询明确相关，并很好地捕捉了其意图。
- **3（良好）：** 相关。这组 [图像/音频]与原始查询明确相关。大多数图像都切合主题。（这是预期的“良好”结果）。
- **2（一般）：** 相关性较弱。少数 [图像/音频] 与查询相关，但整体偏离主题或未抓住主要意思。
- **1（差）：** 大多不相关。整体来看，这组 [图像/音频], 似乎与原始查询无关。

响应格式：

```
{"score": <1-5>,
"reasoning": "<brief explanation of your score>",
"relevant_items_count": <number>,
"total items count": <number>_     _}
```

重要 t: 请整体评估这组内容。这些 [图像/音频] 是否有助于你充分理解原始查询?

邻居检索。本次训练在 TPUv6e-16 Ghostlite Pod（16 个加速器核心）上进行。

我们实验所用的具体参数详见表 3。

表 3 | R4T 的实现超参数。

参数	值
扇出语言模型（强化学习训练）	
优化器	AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$)
学习率	1×10^{-7}
全局批量大小	512
微批量大小	64
梯度累积步数	8
最大序列长度	1024
组大小（G）	8
裁剪 epsilon（ϵ）	0.2
正向 KL 系数（β_1）	0.05
反向 KL 系数（β_2）	0.05
奖励归一化	组标准化
优势估计	组相对
扩散训练	
模型类型	连贯 Transformer
序列长度（L）	12
嵌入维度 (d)	128
隐藏维度	1024
MLP维度	1024
注意力头	16
层	6
随机失活 / 注意力随机失活	0.1
优化器	Adam
调度策略	预热余弦衰减
峰值学习率	3×10^{-4}
预热步数	20,000
总步数	10×10^{6}
批次大小	512
EMA衰减	0.9999
方案	方差爆炸
加权	EDM
噪声调度	切线
范围 [$\sigma_{\text{最小}}$, $\sigma_{\text{最大}}$]	$[10^{-4}, 80.0]$
数据标准差（σdata）	0.088
条件丢弃（$p$drop）	0.1
CFG强度	0.1
步数	256

OAR - 依据性评估

系统提示词：
你是一位专家评估员，负责评估检索到的 [图像/音频] 与其**中间子查询**的匹配程度。

用户提示词：
你将看到用于检索物品的子查询，以及每个子查询实际检索到的物品。`Sub-query 1: {subquery_1} Retrieved Item 1:` [图像/音频]`Sub-query 2: {subquery_2} Retrieved Item 2:` [图像/音频]..... [对所有 k 子查询]重复此过程

对于每个子查询及其检索到的项目，请按 1-5 分评估其**依据性**：

评估标准：

- **5（优秀）：** 完全匹配。该 [图像/音频]是其子查询的完美范例，堪称教科书式示例。
- **4（非常好）：** 高度匹配。该 [图像/音频]清晰且具体地诠释了子查询的概念。
- **3（良好）：** 匹配良好。该 [图像/音频] 是其子查询的一个清晰且合理的示例。（这是预期的“良好”结果。）
- **2（一般）：** 关联较弱。该 [图像/音频]属于这一总体主题（例如“校园风格”），但未能展现子查询的具体概念（例如“天鹅绒”）。
- **1（差）：** 完全不匹配。该 [图像/音频]与其子查询完全无关。

响应格式：

```
{"per_subquery_scores": [{
"subquery": "<subquery_text>",
"score": <1-5>,
"reasoning": "<brief explanation>"}, ...],
"average_score": <1-5>}
```

重要： 你并非在判断与原始查询的对齐。你只判断针对‘子查询1’的[图像/音频] 是否与‘子查询 1’匹配。

Model	Sub Query	Retrieved Images
R4T	bohemian labor day picnic outfit	
R4T	minimalist labor day picnic outfit	
R4T	labor day picnic jumpsuit	
Gemini 2.5 Flash	casual labor day picnic outfit	
Gemini 2.5 Flash	stylish labor day picnic outfit	
Gemini 2.5 Flash	comfortable labor day picnic outfit	

图 6 | : 针对开放式抽象检索任务，在给定宽泛查询“劳动节野餐服装搭配”的情况下，来自 Polyvore 数据集的生成子查询与检索图像示例。