arXiv:2603.01082v1 [cs.CV] 1 Mar 2026

# 超越全局相似性：迈向细粒度多条件多模态检索

Xuan Lu[1, 2, 3*] Kangle Li[1, 2*] Haohang Huang[3] Rui Meng[†] Wenjun Zeng[2,3] Xiaoyu Shen[2, 3‡]

[1]上海交通大学[2]东部理工大学数字孪生研究院，宁波[3]宁波市空间智能与数字衍生重点实验室lux1997@sjtu.edu.cn
xyshen@eitech.edu.cn

## 摘要

多模态大语言模型（*MLLMs*）的最新进展大幅拓展了多模态检索的能力，使系统能够在视觉和文本模态之间对齐并检索信息。然而，现有基准测试大多关注粗粒度或单条件对齐，忽视了现实场景中用户查询会跨模态指定多个相互依赖约束的情况。为弥合这一差距，我们推出***MCMR***（多条件多模态检索）：这是一个大规模基准测试，旨在评估自然语言查询下的细粒度、多条件跨模态检索。*MCMR*涵盖五个商品领域：上装和下装、珠宝、鞋类及家具。它还保留了丰富的长文本元数据，这对组合式匹配至关重要。每个查询都融合了互补的视觉与文本属性，要求模型联合满足所有指定条件，才能判定相关性。我们对一系列多模态大语言模型驱动的多模态检索器和视觉—语言重排序器进行了基准测试，以评估它们的条件感知推理能力。实验结果表明：（*i*）不同模型呈现出明显的模态不对称性；（*ii*）视觉线索主导早期排名精度，而文本元数据则能稳定长尾排序；（*iii*）基于多模态大语言模型的逐项重排序器通过明确验证查询与候选项的一致性，显著提升了细粒度匹配效果。总体而言，*MCMR*建立了一个具有挑战性且具诊断性的基准测试，旨在推动多模态检索迈向组合式、具备约束感知能力且可解释的理解。我们的代码和数据集可在 https://github.com/EIT-NLP/MCMR获取。

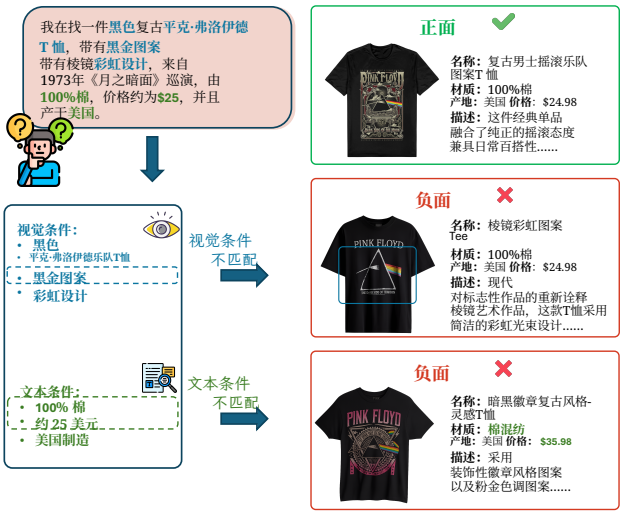



图 1. 由自然语言查询驱动、具有细粒度视觉与文本约束的多条件多模态检索。

## 1. 引言

多模态检索系统旨在对齐并检索文本、图像或其交错组合等异构复合模态中的信息 [1, 29, 38, 40, 41]。早期的模型（如 CLIP [37], 、ALIGN [14], 和 BLIP [19]）通过对比学习，基于整体的图像-描述对进行训练，并已广泛部署于网络搜索 [15, 21, 31], 、电子商务 [39, 44], 和推荐系统 [11, 25]中。然而，图像说明通常只能提供关于视觉内容整体且泛化的描述，因此，这些模型往往更重视全局语义一致性，而非细粒度跨模态理解。近年来，多模态大语言模型（多模态大语言模型）的快速发展，通过支持在开放式自然语言指令 [17, 31,下进行检索，开始重塑这一领域

*贡献相同
†在 Google Cloud AI Research 工作。
‡通讯作者。

33, 46, 48]。基于这一能力，VLM2Vec [17]，、MM-EMBED [21]，和 GME [46]等代表性方法通过生成能够捕捉指令条件语义的统一嵌入，进一步拓展了多模态检索的范围。这一演进标志着检索范式从静态全局对齐转向更具表现力和灵活性的方向。

尽管取得了这些进展，现有基准测试仍很少能同时满足复杂多模态检索所需的三项关键属性：**(i) 细粒度**属性推理，**(ii) 多条件**查询，以及 **(iii) 跨模态证**据，其中不同条件必须同时在图像和文本中得到定位。MS-COCO [22] 和 Flickr30K [36] 等经典图文数据集主要评估图像与图像说明之间粗粒度的全局对齐：每条图像说明都被视为一个完整的整体描述，相关性也不取决于是否满足多个相互独立的约束。FashionIQ [10]，和 CIRR [28]，等时尚与定位领域的细粒度基准测试引入了局部编辑或相对调整（例如，“袖子更短”“风格相同但颜色更亮”），并依赖参考图像来锚定这些修改。尽管这些数据集比全局图像说明基准测试提供了更细致的信息，但它们仍以单一视觉编辑为核心，实际上仍属于单模态，因为大多数属性仅凭图像即可验证。与此同时，研究者也在纯文本场景 [32, 45],中探索了多条件检索，查询会列出多个约束，但所有证据都来自文本，因此无需整合异质的视觉和文本线索。现有基准测试要么涵盖细粒度细节，要么体现多条件结构，却未能兼顾跨模态的两者。MERIT [7]等近期工作通过多语言、交错式的文本—图像查询扩展了检索方式，将参考图像与简短文本片段相结合。在 MERIT 中，许多查询条件是相对于示例图像来表述的（例如，“与产品 *1* 风格相同，且与产品 *2* 颜色相同”），这使任务侧重于与所提供示例进行视觉比较，而非独立指定属性约束。这种设定还预设用户能够在查询时提供有代表性的参考图像；然而，在许多现实搜索场景中，用户更常通过自然语言描述直接表达需求。此外，MERIT 并未明确区分需要视觉定位的属性与仅存在于文本元数据中的属性，因此难以分析模型如何平衡不同模态的证据。

为了解决这些局限，我们推出**MCMR**（多条件多模态检索），这是一个围绕上述三个维度专门设计的大规模基准测试：细粒度、多条件和跨模态。如图1所示，每个查询都是一段自然语言描述，它结合了针对视觉和文本属性的多项组合约束，只有所有条件均得到满足时，候选项才相关。关键在于，MCMR采用双重证据设计，其中某些属性只能从图像中推断（例如特定的图形布局或纹理），而另一些属性只能通过长篇文本元数据获得支持（例如材质、版型或制造细节）。因此，无法仅凭单一模态完成这项任务，并且可以直接检验模型整合不同模态互补证据的能力。

我们在统一协议下对一系列具有代表性的多模态检索器进行基准测试，以评估其细粒度条件感知推理能力；此外，还引入基于多模态大语言模型的逐点重排序器，考察大型视觉语言模型能否通过一对一的约束感知相关性估计，弥合细粒度语义差距。实验揭示了多条件多模态检索中持续存在的挑战。模态消融分析发现，不同模型对模态的依赖存在明显差异：GME [46] 和 LamRA [27] 等模型在移除文本元数据后仍相对稳健，表明它们更依赖视觉特征；而 LLaVE [18] 等模型在纯图像设置下则性能显著下降。进一步分析表明，视觉线索主导首位准确率，文本元数据有助于稳定长尾排序，而基于多模态大语言模型的重排序器通过显式核验每个查询-候选项对，显著提升了细粒度匹配效果。总体而言，这些发现揭示了当前系统在真正的多条件跨模态检索方面的局限，并凸显了开发能够整合组合推理且不牺牲可扩展性的架构的必要性。

我们的主要贡献总结如下：

- **基准测试**。 我们提出了**MCMR**，这是一个大规模基准测试，同时满足三项要求——细粒度属性、多条件查询和跨模态证据，用于评估自然语言查询下的多模态检索。
- **全面评估**。 我们对具有代表性的多模态检索器和基于多模态大语言模型的逐点重排序器开展了全面研究，揭示了在受控组合式设置下系统性的模态不对称性与细粒度推理差距。
- **研究发现**。 我们的分析表明，当前检索器难以同时满足多个异构跨模态约束，而基于多模态大语言模型的重排序显著提升了精确率，凸显了构建可扩展且具备约束感知能力的检索架构的必要性。

## 2. 相关工作

**多模态检索模型**。 多模态检索旨在将异构模态——最常见的是图像和文本——对齐到统一的语义空间中，从而实现

跨模态搜索与推理。一种主流范式是使用 CLIP 等预训练视觉语言模型对各个模态进行编码 [37], BLIP [19], 和 ALIGN [14], , 并通过共享嵌入空间中的余弦距离来衡量相似度。在这些双编码器架构的基础上，任务无关检索器通常分别对图像和文本进行编码，随后通过后期融合协调这两种模态。UniVL-DR [30]和 UniIR [42] 等方法会对异构数据集进行归一化，并使用类似 CLIP 或 BLIP 的编码器将其映射到共享空间，然后在推理时应用校准融合，以平衡不同检索任务中的模态贡献。另一类研究则将轻量级视觉适配器接入强大的文本嵌入模型，从而无需重新训练整个模型，即可对图文对进行组合式编码。代表性方法包括 VISTA [47] 和 MARVEL [49], , 它们将视觉信息注入文本编码器，以获得统一的多模态表示。近来，研究者通过直接将最后几个 Transformer 层的隐藏状态池化为嵌入，将多模态大语言模型（MLLM）改造为通用检索器，E5-V [16] 和 VLM2Vec [17]便是例子。这种多模态大语言模型嵌入范式增强了迁移能力，统一了文本—文本检索与跨模态检索，并且更好地契合检索增强生成流程。在双编码器和基于多模态大语言模型的检索器这两类方法中，模型通常针对全局语义对齐进行优化，且往往呈现出不同的模态依赖性：有些模型更依赖视觉外观，另一些则高度依赖文本语义。这种不对称性是我们在研究中通过实证重点考察的核心方面之一。

**多模态检索基准**。 早期的多模态检索基准侧重于图像与图像说明之间的粗粒度或单条件对齐。MS-COCO [22] 和 Flickr30K [36] 等数据集定义了图像说明与图像之间的一对一对应关系，关注全局相似性，而非组合推理。包括用于新闻图像与标题对齐的 VisualNews [24]和用于商品检索的 Fash- ion200k [10] 在内的领域专用数据集，将这一范式扩展到特定场景，但仍以整体层面为主。除了自然图像，以文档为中心的基准套件（例如由 ViDoRe [9]处理的 DocVQA系列）将检索视为：给定一个自然语言问题，选出正确的富文本文档图像。后续的“组合式”检索基准将重点从全局相似性转向定向属性编辑 [48]。FashionIQ [43] 和 CIRR [28]提出了根据简短文本指令修改参考图像的思路——例如检索“同款衬衫，但要红色”——以测试模型能否纳入局部属性变化。与此同时，WebQA [3] 和 EDIS [26] 等融合模态基准整合文本与视觉证据，用于开放域检索；而 OVEN [13]和 INFOSEEK [4] 则将图像条件问题与维基百科段落或标题配对，以评估混合模态监督下的检索能力。ReMuQ [34], 、OKVQA [35], 和 LLaVA 风格的对话数据集 [23]等进一步扩展，拓宽了查询和候选项的模态，以支持问答和基于对话的检索。近期，MERIT [7] 通过引入多语言和交错多模态输入，突破了单参考查询的局限；这些输入可同时引用多张图像和多个文本片段，从而支持更复杂的条件指定。

| 基准测试 | Q | T | DE | MA | LM |
|---|---|---|---|---|---|
| MS-COCO [22] | T | I | ✗ | ✗ | ✗ |
| Flickr30K [36] | T | I | ✗ | ✗ | ✗ |
| FashionIQ [43] | T+I | I | ✗ | ✗ | ✗ |
| CIRR [28] | T+I | I | ✗ | ✗ | ✗ |
| WebQA [3] | T | T+I | ✗ | ✗ | ✗ |
| EDIS [26] | T | T+I | ✗ | ✗ | ✗ |
| MERIT [7] | T+I | T+I | ✓ | ✓ | ✗ |
| MultiConIR [32] | T | T | ✗ | ✓ | ✓ |
| **MCMR（本文方法）** | T | T+I | ✓ | ✓ | ✓ |

表1. 代表性多模态检索基准的比较。Q = 查询模态，T = 目标模态，DE = 双重证据，MA = 多属性，LM = 长文本元数据。Q/T中的条目使用T表示文本、I表示图像，同时具备两种模态时则使用T+I。MCMR独特地支持双重证据定位、多属性条件和长文本元数据。

## 3. 多模态多条件检索

我们提出MCMR，一个包含10,400件商品、涵盖五种检索场景的多模态商品检索数据集。其原子单元是带有图像和长篇文本描述的商品实例，两种模态提供互补信息。表1 总结了MCMR与先前多模态检索基准测试在查询—目标模态、证据结构和元数据丰富度方面的比较情况。

### 3.1. 数据收集与预处理

MCMR 基于亚马逊评论（2023）语料库 [12], , 涵盖五个商品领域——上装、下装、珠宝、鞋类和家具，同时保留对细粒度多模态检索至关重要的长篇商品元数据。为确保质量与可复现性，我们通过多阶段流程构建数据集，将自动筛选与人工核验相结合，从广泛收集候选项逐步推进至细粒度质量控制。我们遵循三项原则：广泛覆盖、

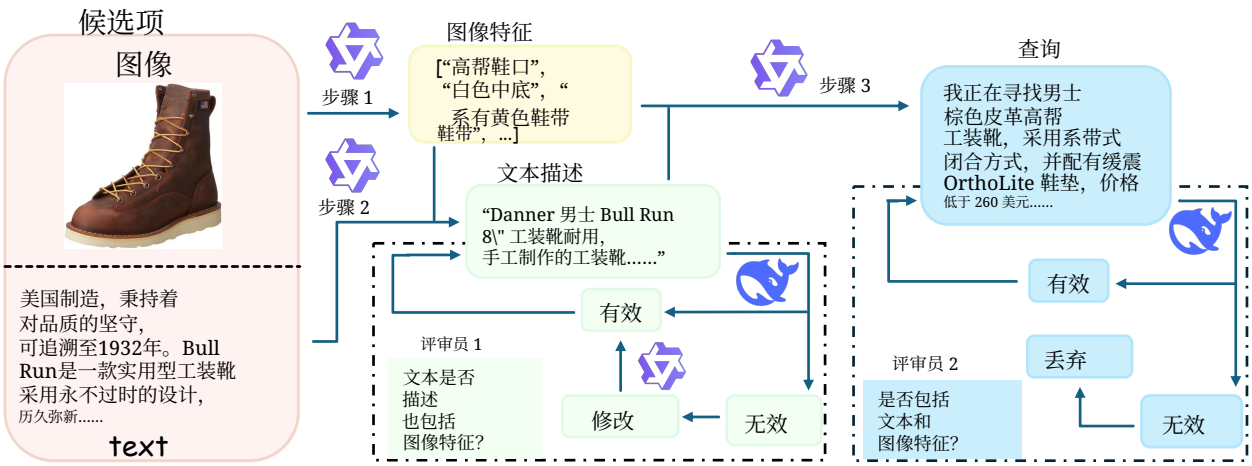



图2. MCMR 的制作流程。通过 Qwen-VL 和 Qwen-Instruct 提取视觉和文本特征，经两个评判阶段验证后，组合成用于细粒度检索评估的查询。

| | 领域分布（数量） | | | | | | 词元统计（词元数） | | |
|---|---|---|---|---|---|---|---|---|---|
| | 上装 | 下装 | 鞋履饰品 | | 家具 | **总计** | max | min | Avg. |
| 查询 | 991 | 803 | 847 | 602 | 754 | **3 997** | 57.00 | 25.00 | 35.86 |
| 候选项 | 29 986 | 29 514 | 24 997 | 5 491 | 14 993 | **104 981** | 269.00 | 54.00 | 190.94 |

表2. 按领域统计的数量以及查询和候选项的令牌长度统计（最大值/最小值/平均值）。

高质量和跨模态一致性。我们从亚马逊评论元数据出发，在四个目标领域中保留图像可解析且具有长篇文本描述的商品。数据整理分三个阶段进行：

**(1) 属性规范化**。我们通过统一单位和货币、规范日期，并统一材料和尺寸的受控词汇表，来标准化结构化属性，从而在不同领域创建统一表示。

**(2) 质量筛选与去重**。我们根据文本长度、图像分辨率、宽高比以及感知相似度或嵌入相似度进行筛选和近重复检测。我们会删除质量低、信息不完整或重复的记录。为防止信息泄漏，会屏蔽或排除 ASIN 和 URL 等直接标识符。

**(3) 互补性约束**。每个条目必须至少包含一个仅文本细粒度属性和一个仅图像细粒度属性，确保两种模态都能提供独特证据，而不是让任务仅凭单一模态即可完成。

### 3.2. 数据集构建流程

为平衡成本与保真度，我们采用协作式流程，将高效的中型模型用于大规模生成，并将更强的模型用于验证和优化。目标是生成基于互补视觉与文本证据的多条件、跨模态且可验证的自然语言查询，内容基于互补的视觉与文本证据。[1]整体构建流程如图2所示。

**图像侧结构化扩展**。对于每张产品图像，Qwen2.5-VL-32B-Instruct 会生成一份以证据为依据的结构化摘要，其中包含类别标签和仅图像属性（例如颜色、纹理、结构细节、形状、结构）。生成内容严格排除功能性或推测性信息，从而形成视觉证据层，用于 (i) 互补性验证和 (ii) 查询构建。

**文本侧结构化扩展**。产品标题、描述和特征列表通过 JSON提取模板转换为结构化文本档案。必填字段包括仅文本属性与类别文本；可选字段（例如价格、发行年份、材质、风格、品牌）在存在时予以添加。标识符经过净化，且仅当品牌提及与仅图像属性同时出现时才允许保留。模式感知合并在保留规范元数据的同时追加结构化字段，以实现多模态对齐。

[1]用于构建 MCMR 的提示见补充材料。

**文本描述生成**。 我们采用Qwen3-32B-Instruct，仅基于文本元数据生成简洁（80–120词）的目录式文本摘要，并明确排除任何视觉描述符。使用DeepSeek-R1-Distill-Qwen-32B的验证器-编辑器循环检测跨模态泄漏并强制一致性，同时通过人工抽查确保稳健性。

**查询生成**。 基于图像属性和文本摘要，Qwen3-32B-Instruct 会生成第一人称、多条件查询，将来自两种模态的互补约束结合起来。数值和时间信息会经过规范化处理，标识符会被隐藏，并保持中性的“购物者”语气。针对不同领域的提示变体（服装、珠宝、鞋类和家具）可引导属性选择，例如服装的面料/版型/护理 或珠宝的宝石/切工/镶嵌 ，以保持语义多样性。

**查询验证**。DeepSeek-R1-Distill-Qwen-32B 作为独立验证器，用于评估跨模态覆盖和数值/时间一致性。对于未通过验证或处于边界情形的样本，会重新生成；此外，还会对一小部分通过验证的样本进行人工检查。为进一步验证质量，我们开展了一项包含 100 个样本的人工研究：一位标注者根据每件商品的图像和元数据撰写自然语言查询，另一位标注者则在双盲条件下，从属性正确性、跨模态覆盖和自然度三个方面评估人工撰写的查询和生成的查询（采用 5 点李克特量表）。两类查询的平均分相近（4.33 对 4.41），偏好率也相似（47% 对 49%），这证实生成的查询与人工撰写的查询十分接近，且验证开销极低[2]。表 2展示了 MCMR数据集的关键统计数据，包括查询和候选项的数量，以及它们的文本词元长度。

## 4. 实验

本节首先在 § 4.1 介绍实验设置，随后在 § 4.2 展示主要结果，包括对检索模型的整体比较。接着，§ 4.3 展示多模态大语言模型的重排序性能。所有实验均使用两块 NVIDIA A100 GPU（80 GB）进行。

### 4.1. 基线方法与指标

我们在相同的预处理和推理设置下，评估五种具有代表性的多模态检索器和五种多模态大语言模型重排序器：

**•多模态检索器：** GME-Qwen2-VL-7B [46],LLaVE-7B [18], VLM2Vec [17], LamRA-Ret-Qwen2.5-VL-70亿参数 [27], 和 CORAL [6]。

[2]人工验证的详细信息见补充材料。

**•MLLM重排序器：** Qwen2.5-VL-32B，Qwen2.5-VL-70亿参数, Qwen3-VL-4B，Qwen3-VL-8B [2], InternVL3-80亿参数-Instruct [5]， Qwen3-VL-Reranker-8B [20]， 和 lychee-reranker-mm [8]。

对于检索，我们使用模型融合的图文接口对候选项进行嵌入，而查询则仅根据文本编码。所有系统均使用公开发布的检查点，以零样本模式运行。对于重排序，我们取最强第一阶段检索器返回的前 50 个候选项，并使用视觉语言多模态大语言模型作为逐点重排序器。每个查询-候选项对均独立评估：重排序器接收文本查询以及候选项的图像和元数据，并通过统一提示生成二元相关性判断（真/假）。模型对“真”标记的归一化逻辑值分数用作相关性分数，并据此对候选项排序，得到最终顺序。这种成对评估通过提供细粒度的跨模态相关性估计，补充了第一阶段检索器。

我们报告三种广泛使用的信息检索指标：Recall@K、NDCG@K 和 MRR@10。Recall@K 衡量在前-$K(K \in \{1, 5, 10, 50, 100\})$ 中检索到的相关项所占比例，反映系统在多条件约束下呈现正确结果的能力。NDCG@K 根据相关文档的位置加权评估排序质量，着重考察前列结果中的细粒度排序。MRR@10 则通过计算前 10 个位置中首个相关项的平均倒数排名，概括早期排序性能。

### 4.2. 主要结果

我们在三种候选可见性设定下评估 MCMR 上的多模态检索器：融合（图像+文本）、仅图像和仅文本。完整结果见表 3。

**融合候选项（图像 + 文本）**。如表3所示，当视觉和文本元数据同时可用时，整体准确率仍处于中等水平：主流检索器的 Recall@1 仅为 18–27%，而*VLM2Vec* 的表现明显更差，仅为 1.83%。尽管如此，大多数模型最终都能在较大的截断点找到正确项。最佳 Recall@10 达到 53.34%（CORAL），而最强的长程表现——78.64% Recall@100——由 LLaVE 实现，这表明相关项往往能够被找到，但排名并不靠前。早期排序与长截断检索之间不断扩大的差距，揭示了各系统中一致的模式：模型具备尚可的粗粒度检索能力，但在多条件约束下难以进行细粒度排序。这种差异表明，下游重排序器仍有很大的提升空间，可在检索到正确候选项后进一步优化排序。

**移除文本的影响（纯图像候选项）**。移除文本元数据而保留图像，会显著影响早期排序准确率。对于若干模型，R@1/5/10 大幅下降，某些情况下 Top-1 准确率甚至低于 1%。相比之下，少数视觉能力较强的编码器（如 GME）仍能保持有竞争力的分数，有时甚至在 R@1 上有所提升，这表明某些系统仅凭视觉线索就能满足多个条件。在较大的截断点处，性能变得更加稳定：R@50/100 仍相对较高，且 nDCG@100 的降幅远小于 nDCG@10。这一模式表明，移除文本主要会削弱细粒度辨别能力，但不会严重妨碍在候选集中检索到正确项。

**移除图像的影响（纯文本候选项）**。移除图像并仅依赖文本元数据，会导致所有指标的准确率相较于融合候选项明显下降。早期排名的检索表现尤其薄弱：最佳 Recall@1 仅为 12.98%（MM-EMBED），大多数模型的表现处于 7–12% 区间。中、长截断位置的表现也有所下降——Recall@50/100 远低于融合结果，最高约为 53–62%。仅文本检索的表现也始终弱于纯图像检索。视觉能力较强的模型（如 GME-Qwen2VL）表现大幅下降：Recall@10 从 51.10% 降至 29.60%，Recall@100 则从 78.86% 降至 57.50%。这些结果凸显了 MCMR 的判别线索主要来自视觉。文本主要作为补充信号，与图像结合时可增强对齐；而当两种模态都可用时，模型始终能取得最佳表现。

**逐点重排序带来的提升**我们对 LLaVE-7B 返回的固定前 50 个候选项应用第二阶段逐点重排序器。LLaVE-7B 在融合镶嵌方式下取得最高的 Recall@50（72.01）。每个重排序器都是一个视觉语言多模态大语言模型，会独立评估文本查询与融合图文候选项的相关性。模型会输出答案首个位置上 token`True` 的归一化概率，我们将其用作相关性分数 $p(\mathrm{True})$；随后按照该分数对候选项排序，若分数相同，则按其原始检索排名决定先后。如图4所示，在固定的前 50 个候选项中，逐点重排序器显著改善了前列排序，nDCG@1 分数稳定在 70–80 之间。lychee-reranker-mm 在各个截断位置均表现最佳，其 nDCG@1/5/10/50 分数分别为 92.35 / 93.41 / 94.42 / 94.86；其次是 InternVL-8B 和 Qwen3-VL-Reranker-8B，而基于 Qwen3-VL 的模型则略显落后。参数量本身无法预测排序能力；架构设计和视觉文本 grounding 对成对相关性估计似乎更为关键。在成对相关性估计中，参数量本身无法预测排序能力；架构设计和视觉文本 grounding 似乎更为关键。nDCG@1 上的性能差距最大，随着指标趋近 nDCG@50，差距逐渐缩小，这表明多模态大语言模型重排序器的提升主要集中在最前列，而长程排序则相对稳定。多模态大语言模型重排序器具有很强的一对一相关性判断能力，与第一阶段检索器较低的 Recall@1 形成对比，揭示出明显差距：当前检索模型难以进行细粒度、多条件语义推理，而生成式视觉语言模型在独立评估每一对候选项时能够弥合这一差距。这一对比也验证了将 MCMR 设计为基准测试的合理性：它能有效揭示跨模态融合和细粒度检索方面的弱点，为评估模型的多模态对齐与条件感知推理能力提供有意义的测试平台。

## 5. 分析

### 5.1. 候选侧模态分析

**模态贡献与不对称性**。 表3 中的模态消融结果显示出明显的模态依赖性不对称。在纯图像候选项下，GME 与 LamRA 在 R@10 上仍接近其融合性能，而 LLaVE 一旦移除文本便几乎崩溃，R@1 从 24.99 降至 0.90。切换至纯文本候选项同样会使所有模型相对于融合设置表现下降，其中 MM-EMBED 与 CORAL 在 R@10 上的降幅相对较小。

总体而言，纯文本候选项在 R@10 上的表现远差于融合候选项；在五个共有模型中的四个上，它们也落后于纯图像候选项，这表明在 MCMR 上，视觉线索比仅靠文本元数据更具判别力。然而，融合查询在 R@10 上仍比仅图像高出约 4–8 个点，说明文本元数据在视觉证据之外提供了互补约束。这与数据集设计相符：每个条目至少包含一个仅图像属性和一个仅文本属性，因此纯粹的单模态推理并不完整。

**模型级方差与设计启示**。移除模态会加剧跨模型方差。GME 保持稳定，而 LLaVE 和 CORAL 在早期排名中表现出明显退化，这表明它们严重依赖文本先验。相比之下，MM-EMBED 似乎更加稳健。这些观察结果凸显了当前检索模型的一项关键局限：在融合输入下实现较强的全局语义相似度，并不意味着当个别模态变弱时仍能保持稳健。未来的设计或许需要明确保留查询条件并逐一检查，同时减少模态之间的冗余重叠，以维持性能。

(a) 融合（图像+文本）

| 模型 | 规模 | R@1 | R@5 | R@10 | R@50 | R@100 | 平均倒数排名 | N@1 | N@5 | N@10 | N@50 | N@100 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LLaVE | 7B | 24.99 | 43.85 | 53.13 | **72.01** | **78.64** | 33.15 | 24.99 | 34.88 | 37.88 | 42.11 | 43.19 |
| GME-Qwen2VL | 7B | 21.23 | 38.20 | 45.74 | 64.71 | 73.52 | 28.35 | 21.23 | 30.06 | 32.48 | 36.66 | 38.08 |
| LamRA-Qwen2.5VL | 7B | 17.96 | 34.99 | 43.30 | 64.36 | 73.24 | 25.27 | 17.96 | 26.85 | 29.53 | 34.25 | 35.69 |
| MM-EMBED | 8B | 21.74 | 39.58 | 47.91 | 66.22 | 74.16 | 29.35 | 21.74 | 31.05 | 33.75 | 37.82 | 39.11 |
| CORAL | 3B | **26.57** | **46.69** | **53.34** | 70.90 | 77.73 | **34.94** | **26.57** | **37.20** | **39.35** | **43.27** | **44.37** |
| VLM2Vec | 4B | 1.83 | 4.88 | 7.03 | 14.38 | 18.96 | 3.11 | 1.83 | 3.33 | 4.02 | 5.63 | 6.38 |

(b) 仅图像

| 模型 | 规模 | R@1 | R@5 | R@10 | R@50 | R@100 | 平均倒数排名 | N@1 | N@5 | N@10 | N@50 | N@100 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LLaVE | 7B | 0.90 | 2.53 | 3.93 | 9.48 | 13.68 | 1.67 | 0.90 | 1.52 | 2.19 | 3.39 | 4.06 |
| GME-Qwen2VL | 7B | **21.79** | **41.30** | **51.10** | **71.36** | **78.86** | **30.13** | **21.79** | **31.91** | **35.08** | **39.63** | **40.86** |
| LamRA-Qwen2.5VL | 7B | 18.05 | 36.25 | 43.30 | 66.83 | 76.73 | 25.96 | 18.05 | 27.63 | 30.51 | 35.35 | 36.96 |
| MM-EMBED | 8B | 13.23 | 28.15 | 35.68 | 57.29 | 67.00 | 19.72 | 13.23 | 21.06 | 23.50 | 28.30 | 29.89 |
| CORAL | 3B | 11.51 | 25.99 | 33.53 | 54.72 | 64.15 | 17.83 | 11.51 | 19.11 | 21.54 | 26.19 | 27.72 |

(c) 仅文本

| 模型 | 规模 | R@1 | R@5 | R@10 | R@50 | R@100 | 平均倒数排名 | N@1 | N@5 | N@10 | N@50 | N@100 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LLaVE | 7B | 11.95 | 23.38 | 29.43 | 48.02 | 56.23 | 16.83 | 11.95 | 17.85 | 19.80 | 23.92 | 25.25 |
| GME-Qwen2VL | 7B | 10.62 | 22.65 | 29.60 | 48.55 | 57.50 | 16.00 | 10.62 | 16.95 | 19.21 | 23.42 | 24.87 |
| LamRA-Qwen2.5VL | 7B | 7.28 | 16.21 | 22.31 | 38.53 | 48.49 | 11.27 | 7.28 | 16.21 | 13.85 | 17.36 | 18.98 |
| MM-EMBED | 8B | **12.98** | **26.88** | **34.50** | **53.66** | **62.37** | **18.94** | **12.98** | **20.15** | **22.61** | **26.86** | **28.67** |
| CORAL | 3B | 8.58 | 17.10 | 22.88 | 39.30 | 47.73 | 12.37 | 8.58 | 12.98 | 14.83 | 18.43 | 19.80 |

表 3. 在三种候选可见性设定（融合/仅图像/仅文本）下，MCMR 上的检索结果。查询采用文本编码；我们报告 R@K（K=1, 5,10,50,100）、NDCG@K 和 MRR@10。数值为百分比。

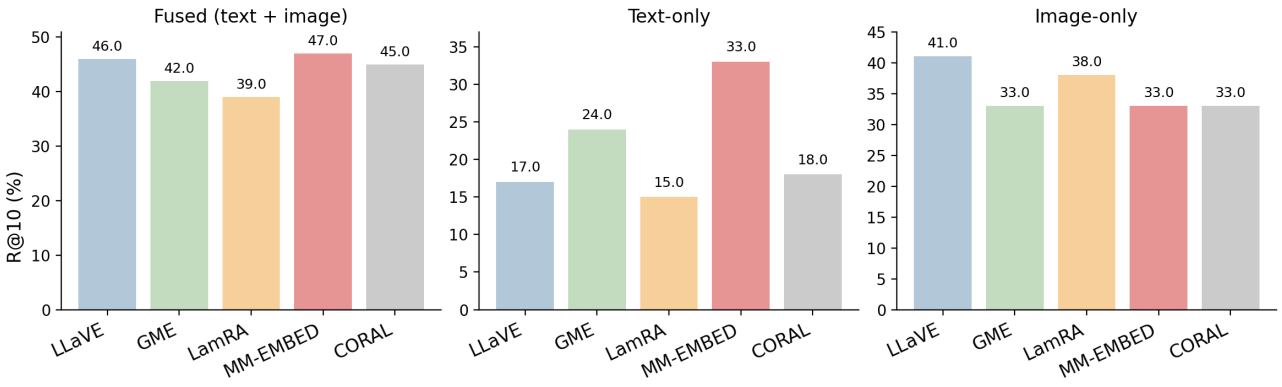



图 3. 不同查询模式下的检索性能（Recall@10）。完整结果见补充材料。

## 5.2. 查询侧组合效应

**查询侧模态消融。** 为分析查询侧的模态贡献，我们在一个包含 100 个查询的子集上构造了受控的查询变体。我们从融合的自然语言查询开始，这些查询由每个产品的图像特征列表和文本描述元数据共同生成——数据。对于仅文本条件，我们使用 ChatGPT-4o 将每个融合查询与图像特征列表进行比对，并删除所有对应于视觉属性的短语，从而得到仅保留由元数据派生的约束的版本。对称地，对于仅图像条件，我们将融合查询与文本描述进行比较，并移除 _

| 模型 | size | N@1 | N@5 | N@10 | N@50 |
|---|---|---|---|---|---|
| Qwen2.5-VL | 32B | 78.22 | 79.87 | 82.58 | 84.88 |
| Qwen2.5-VL | 7B | 74.16 | 77.26 | 80.26 | 82.84 |
| internVL | 8b | 80.28 | 81.95 | 84.66 | 86.61 |
| Qwen3-VL | 8B | 72.45 | 75.48 | 78.32 | 81.44 |
| Qwen3-VL | 4B | 69.92 | 73.39 | 76.81 | 79.95 |
| Qwen3-VL-Reranker-8B | 8B | 78.69 | 80.79 | 83.51 | 85.57 |
| lychee-reranker-mm | 8B | **92.35** | **93.41** | **94.42** | **94.86** |

表4. 在 LLaVE-7B 的前50个候选池上评估 MCMR 中逐点重排序器的 NDCG@K。

基于文本元数据的短语，仅保留有视觉依据的属性。所有得到的查询三元组（融合、仅文本和仅图像）均经人工检查其忠实性，同时候选池在所有镶嵌方式中保持固定，并带有融合的图像–文本元数据。

图 3总结了三种查询模式下 R@10 的检索性能：融合文本+图像查询、纯文本查询和纯图像查询。[3] 在纯文本查询中移除图像派生约束，会导致所有模型的早期排名性能严重下降：与融合设置相比，R@10 下降约 20–30 个点。相比之下，在纯图像查询中移除文本派生约束造成的损失要小得多，而且在若干情况下，纯图像得分仍接近融合性能，或仅低约 5–10 个点。不同模式下，系统排名也会发生变化：在融合查询下，MM-EMBED 和 LLaVE 表现最强；在纯文本设置下，CORAL 成为表现最佳的系统；而仅有视觉证据时，LLaVE 明显占据主导地位。总体而言，这些模式表明，所有系统在两种模态同时存在时获益最大，但在被迫采用单模态模式时，对图像线索的依赖远高于对文本元数据的依赖。

**查询约束数量的影响**。 我们进一步改变查询中组合约束的数量。令$k_T$ 和 $k_I$ 分别表示源自文本和源自图像的约束数量。我们考虑 $k_T = k_I \in \{1, 2, 3, 4, 5\}$的配置，将 1T+1I 作为低约束基线，将 5T+5I 作为约束最多的设置，同时固定使用融合的图像–文本元数据作为候选池，并保持评估协议不变。对于每个 $k_T = k_I$值，我们使用 ChatGPT-4o 从图像特征列表中选择$k$ 个有视觉依据的属性，并从文本描述中选择 $k$ 个元数据派生属性，再将它们组合成一条自然语言查询。所有生成的查询都经过人工检查，以确保其忠实度和流畅度；所有查询均以文本形式编码。

图 4 展示了每个查询中组合约束数量不同时的 Recall@10，其中文本约束数量相等，

[3]所有模型和查询模式的完整指标均见补充材料。

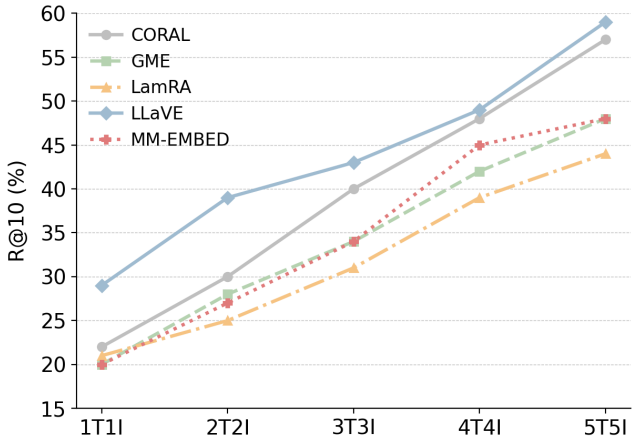



图 4. 不同组合约束数量下的 Recall@10（$k_T = k_I \in 1$，2，3，4，5）。候选项固定为融合的图像–文本元数据。

派生约束和图像派生约束的数量从 1T $k_T = k_I$ I 到 5T+1I 不等，候选项则使用融合的图像–文本元数据固定。+5对于所有模型，随着约束数量增加，性能始终有所提升；不过，从 4T[4] I 增至 5T+4I 的提升小于从 2T+5I 增至 3T+2I 的提升，表明约束数量较高时收益递减。

## 5.3. 逐点重排序性能

如表 4所示，所有逐点重排序器在靠前排名上的准确率都很高，归一化折损累计增益@1 达到 70–80；表现最佳的模型 InternVL-8B 达到 80.3。当多模态大语言模型逐一比较查询和候选项时，它能够通过明确的跨模态定位和条件验证，可靠地识别正样本。这与第一阶段检索器形成鲜明对比：其最佳归一化折损累计增益@1 仅为 26.57（CORAL），揭示了显著的性能差距。重排序与第一阶段检索之间的巨大差距表明，当前的多模态检索器在细粒度多条件匹配方面仍面临困难。全局嵌入模型侧重于整体语义相似性，但往往无法验证是否同时满足所有查询条件。相比之下，重排序器会独立评估每个查询-候选项对，从而实现细致的跨模态定位和明确的约束检查。其代价是计算效率：逐点重排序成本高昂，难以扩展到大型检索语料库。这些发现凸显了未来系统面临的一项核心挑战——将细粒度多条件的多模态融合整合进检索模型，同时保持真实应用所需的可扩展性。

[4]补充材料中提供了涵盖所有模型和约束镶嵌方式的 Recall、平均倒数排名和归一化折损累计增益指标完整表格。

## 6. 结论

在本研究中，我们提出了 MCMR（多条件多模态检索），这是一个用于评估自然语言查询下细粒度多条件检索的大规模基准测试。通过明确区分视觉约束和文本约束，MCMR 能够系统地评估模型跨模态推理多个相互依存条件的能力。通过对多种多模态检索器和基于多模态大语言模型的重排序器进行全面实验，我们发现了持续存在的模态不对称性——视觉特征主导靠前排名的精确率，而文本元数据则有助于稳定长尾排序。此外，基于多模态大语言模型的逐点重排序器通过明确验证查询与候选项的一致性，显著提升了细粒度匹配能力，但计算成本也更高。这些发现揭示了当前多模态检索系统的一项根本局限：尽管它们具备较强的全局语义对齐能力，却难以大规模整合组合式、条件感知的推理。我们希望 MCMR 能成为一个具有诊断性和挑战性的基准测试，推动未来研究开发能够兼顾可扩展性、多模态落地和组合式理解能力的检索器。

## 参考文献


[1] Mohammad Mahdi Abootorabi、Amirhosein Zobeiri、Mahdi Dehghani、Mohammadali Mohammadkhani、Bardia Mo- hammadi、Omid Ghahroodi、Mahdieh Soleymani Baghshah 和 Ehsaneddin Asgari。《任意模态提问：多模态检索增强生成综合综述》。载于 计算语言学协会：*ACL 2025* 论文集，第 16776–16809 页，奥地利维也纳，2025年。计算语言学协会。 1

[2] Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang、Humen Zhong、Yuanzhi Zhu、Mingkun Yang、Zhao- hai Li、Jianqiang Wan、Pengfei Wang、Wei Ding、Zheren Fu、Yiheng Xu、Jiabo Ye、Xi Zhang、Tianbao Xie、Zesen Cheng、Hang Zhang、Zhibo Yang、Haiyang Xu 和 Jun- yang Lin。Qwen2.5-vl 技术报告。*arXiv* 预印本 *arXiv:2502.13923*，2025年。5[3] Yingshan Chang、Guihong Cao、Mridu Narang、Jianfeng Gao、Hisami Suzuki 和 Yonatan Bisk。Webqa：多跳多模态问答。载于 *IEEE/CVF*计算机视觉与模式识别会议，第16474–16483页，2022年。3[4] Yang Chen、Hexiang Hu、Yi Luan、Haitian Sun、Soravit Changpinyo、Alan Ritter 和 Ming-Wei Chang。预训练视觉与语言模型能否回答视觉信息检索问题？载于 *2023*年自然语言处理实证方法会议论文集，第14948–14968页，新加坡，2023年。计算语言学协会。3[5] Zhe Chen、Jiannan Wu、Wenhai Wang、Weijie Su、Guo Chen、Sen Xing、Muyan Zhong、Qinglong Zhang、Xizhou Zhu、Lewei Lu 等。Internvl：扩展视觉基础模型并进行对齐，以完成通用视觉语言任务。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第24185–24198页，2024年。5

[6] Wei Chow、Yuan Gao、Linfeng Li、Xian Wang、Qi Xu、Hang Song、Lingdong Kong、Ran Zhou、Yi Zeng、Yidong Cai、Botian Jiang、Shilin Xu、Jiajun Zhang、Minghui Qiu、Xiangtai Li、Tianshu Yang、Siliang Tang 和 Juncheng Li。Merit：通过交错多条件查询实现多语言语义检索。arXiv 预印本 arXiv:2506 03144，2025 年。提出了用于多模态检索的 CORAL微调框架。5[7] Wei Chow、Yuan Gao、Linfeng Li、Xian Wang、Qi Xu、Hang Song、Lingdong Kong、Ran Zhou、Yi Zeng、Yidong Cai、Botian Jiang、Shilin Xu、Jiajun Zhang、Minghui Qiu、Xiangtai Li、Tianshu Yang、Siliang Tang 和 Juncheng Li。Merit：通过交错多条件查询实现多语言语义检索，2025 年。2，3[8] Ziqi Dai、Xin Zhang、Mingxin Li、Yanzhao Zhang、Dingkun Long、Pengjun Xie、Meishan Zhang、Wenjie Li 和 Min Zhang。监督微调还是对比学习？迈向更好的多模态大语言模型重排序，2025 年。5[9] Manuel Faysse、Hugues Sibille、Tony Wu、Bilel Omrani、Gautier Viaud、C´eline Hudelot 和 Pierre Colombo。Colpali：利用视觉语言模型实现高效文档检索。载于《第十三届国际学习表征会议》，2025 年。3[10] Xintong Han、Zuxuan Wu、Phoenix X. Huang、Xiao Zhang、Menglong Zhu、Yuan Li、Yang Zhao 和 Larry S. Davis。自动化空间感知时尚概念发现。载于《IEEE国际计算机视觉会议》，ICCV 2017，第 1472–1480 页，意大利威尼斯，2017 年。IEEE计算机学会。2，3[11] Xintian Han、Honggang Chen、Quan Lin、Jingyue Gao、Xi- angyuan Ren、Lifei Zhu、Zhisheng Ye、Shikang Wu、Xiong-Hang Xie、Xiaochu Gan、Bingzheng Wei、Peng Xu、Zhe Wang、Yuchao Zheng、Jingjian Lin、Di Wu 和 Junfeng Ge。Lemur：大规模端到端多模态推荐，2025 年。1[12] Yupeng Hou、Jiacheng Li、Zhankui He、An Yan、Xiusi Chen 和 Julian McAuley。连接语言与物品，服务于检索与推荐。arXiv 预印本 arXiv:2403.03952，2024 年。3[13] Hexiang Hu、Yi Luan、Yang Chen、Urvashi Khandelwal、Mandar Joshi、Kenton Lee、Kristina Toutanova 和 Ming-Wei Chang。开放域视觉实体识别：迈向识别数百万维基百科实体。载于《IEEE/CVF国际计算机视觉会议》，ICCV 2023，第 12031–12041 页，法国巴黎，2023 年。IEEE。3[14] Chao Jia、Yinfei Yang、Ye Xia、Yi-Ting Chen、Zarana Parekh、Hieu Pham、Quoc Le、Yun-Hsuan Sung、Zhen Li 和 Tom Duerig。利用噪声文本监督扩展视觉及视觉语言表征学习。载于《国际机器学习会议》，第 4904–4916 页。PMLR，2021 年。1，3[15] Yiming Jia、Jiachen Li、Xiang Yue、Bo Li、Ping Nie、Kai Zou 和 Wenhu Chen。VisualWebInstruct：通过网络搜索扩展多模态指令数据。载于《

2025年自然语言处理实证方法会议论文集》，第 1373–1393 页，中国苏州，2025 年。计算语言学协会。1[16] Ting Jiang、Minghui Song、Zihan Zhang、Haizhen Huang、Weiwei Deng、Feng Sun、Qi Zhang、Deqing Wang 和 Fuzhen Zhuang。E5-V：利用多模态大语言模型构建通用嵌入。CoRR，abs/2407.12580，2024 年。3[17] Ziyan Jiang、Rui Meng、Xinyi Yang、Semih Yavuz、Yingbo Zhou 和 Wenhu Chen。VLM2Vec：训练视觉语言模型以执行大规模多模态嵌入任务。载于《第十三届国际学习表征会议》，2024 年。1，2，3，5[18] Zhibin Lan、Liqiang Niu、Fandong Meng、Jie Zhou 和 Jinsong Su。LLaVE：采用难度加权对比学习的大语言与视觉嵌入模型。载于《计算语言学协会论文集：2025年自然语言处理实证方法会议》，第 13721–13735 页，中国苏州，2025 年。计算语言学协会。2，5[19] Junnan Li、Dongxu Li、Cai ming Xiong 和 Steven Hoi。BLIP：通过自举式语言图像预训练实现统一的视觉语言理解与生成。载于《国际机器学习会议》，第 12888–12900 页。PMLR，2022 年。1，3[20] Mingxin Li、Yanzhao Zhang、Dingkun Long、Keqin Chen、Sibo Song、Shuai Bai、Zhibo Yang、Pengjun Xie、An Yang、Dayiheng Liu、Jingren Zhou 和 Junyang Lin。Qwen3-VL-Embedding 与 Qwen3-VL-Reranker：用于先进多模态检索与排序的统一框架。arXiv，2026 年。5[21] Sheng-Chieh Lin、Chankyu Lee、Mohammad Shoeybi、Jimmy Lin、Bryan Catanzaro 和 Wei Ping。MM-EMBED：利用多模态大语言模型实现通用多模态检索。载于《第十三届国际学习表征会议》，2024 年。1，2[22] Tsung-Yi Lin、Michael Maire、Serge J. Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Doll´ar 和 C. Lawrence Zitnick。Microsoft COCO：上下文中的常见物体。载于《第13届欧洲计算机视觉会议》，ECCV 2014，第 740–755 页，瑞士苏黎世，2014 年。Springer。2，3[23] Weizhe Lin、Jingbiao Mei、Jinghong Chen 和 Bill Byrne。PreFLMR：扩展细粒度晚交互多模态检索器。载于《计算语言学协会第62届年会论文集》（第 1 卷：长篇论文），第 5294–5316 页，泰国曼谷，2024 年。计算语言学协会。3[24] Fuxiao Liu、Yinghan Wang、Tianlu Wang 和 Vicente Ordonez。Visual News：新闻图像说明的基准测试与挑战。载于《2021年自然语言处理实证方法会议论文集》，第 6761–6771 页，线上及多米尼加共和国蓬塔卡纳，2021 年。计算语言学协会。3[25] Qidong Liu、Jiaxi Hu、Yutian Xiao、Xiangyu Zhao、Jingtong Gao、Wanyu Wang、Qing Li 和 Jiliang Tang。多模态推荐系统：综述。《ACM计算调查》，57（2），2024 年。1

[26] Siqi Liu、Weixi Feng、Tsu-Jui Fu、Wenhu Chen 和 William Wang。EDIS：面向多模态网页内容的实体驱动图像搜索。载于《2023年自然语言处理实证方法会议论文集》，第4877–4894页，新加坡，2023年。计算语言学协会。3[27] Yikun Liu、Pingan Chen、Jiayin Cai、Xiaolong Jiang、Yao Hu、Jiangchao Yao、Yanfeng Wang 和 Weidi Xie。LamRA：将大型多模态模型用作高级检索助手。载于《IEEE/CVF计算机视觉与模式识别会议论文集》（CVPR），第4015–4025页。IEEE/CVF，2025年。2, 5[28] Zheyuan Liu、Cristian Rodriguez Opazo、Damien Teney 和 Stephen Gould。利用预训练视觉语言模型检索真实场景图像。载于《2021年IEEE/CVF国际计算机视觉会议》（ICCV 2021），第2105–2114页，加拿大蒙特利尔，2021年。IEEE。2, 3[29] Zheyuan Liu、Weixuan Sun、Yicong Hong、Damien Teney 和 Stephen Gould。通过文本提示学习对组合图像检索进行双向训练，2023年。1[30] Zhenghao Liu、Chenyan Xiong、Yuanhuiyi Lv、Zhiyuan Liu 和 Ge Yu。通用视觉语言稠密检索：为多模态检索学习统一表示空间。载于第十一届国际学习表征会议，2023年。3[31] Xuan Lu、Haohang Huang、Rui Meng、Yaohui Jin、Wenjun Zeng 和 Xiaoyu Shen。重新思考文档排序中的推理：为何思维链效果不佳。arXiv预印本 arXiv:2510.08985，2025年。1[32] Xuan Lu、Sifan Liu、Bochao Yin、Yongqi Li、Xinghao Chen、Hui Su、Yaohui Jin、Wenjun Zeng 和 Xiaoyu Shen。MultiConIR：迈向多条件信息检索。载于《计算语言学协会论文集：2025年自然语言处理实证方法会议》，第13471–13494页，中国苏州，2025年。计算语言学协会。2, 3[33] Xuan Lu、Haohang Huang、Rui Meng、Yaohui Jin、Wenjun Zeng 和 Xiaoyu Shen。工具文档记录不足：简单的文档扩展即可提升工具检索。载于第十四届国际学习表征会议，2026年。2[34] Man Luo、Zhiyuan Fang、Tejas Gokhale、Yezhou Yang 和 Chitta Baral。通过多模态查询进行端到端知识检索。载于《计算语言学协会第61届年会论文集》（第1卷：长篇论文），第8573–8589页，加拿大多伦多，2023年。计算语言学协会。3[35] Kenneth Marino、Mohammad Rastegari、Ali Farhadi 和 Roozbeh Mottaghi。OK-VQA：一项需要外部知识的视觉问答基准测试。载于《IEEE计算机视觉与模式识别会议》（CVPR 2019），第3195–3204页，美国加利福尼亚州长滩，2019年。计算机视觉基金会 / IEEE。3[36] Bryan A. Plummer、Liwei Wang、Chris M. Cervantes、Juan C. Caicedo、Julia Hockenmaier 和 Svetlana Lazebnik。Flickr30k实体：收集区域与短语的对应关系，以构建更丰富的图像到句子模型。载于《2015年IEEE国际计算机视觉会议》（ICCV

2015年），第2641–2649页，智利圣地亚哥，2015年。IEEE计算机学会。2, 3[37]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack 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。PMLR，2021年。1, 3[38] Xiaoyu Shen、Svitlana Vakulenko、Marco del Tredici、Gianni Barlacchi、Bill Byrne 和 Adri`a de Gispert。开放域问答的低资源稠密检索：全面综述，2022年。1[39] Xiaoyu Shen、Akari Asai、Bill Byrne 和 Adria De Gispert。xPQA：支持12种语言的跨语言商品问答。载于《计算语言学协会第61届年会论文集》（第5卷：工业赛道），第103–115页，加拿大多伦多，2023年。计算语言学协会。1[40] Xiaoyu Shen、Svitlana Vakulenko、Marco del Tredici、Gianni Barlacchi、Bill Byrne 和 Adria de Gispert。用于开放域问答的弱监督神经排序：综述。载于《计算语言学协会论文集：EACL 2023》，第1736–1750页，克罗地亚杜布罗夫尼克，2023年。计算语言学协会。1[41] Tianshi Wang、Fengling Li、Lei Zhu、Jingjing Li、Zheng Zhang 和 Heng Tao Shen。跨模态检索：方法与未来方向的系统综述，2024年。1[42] Cong Wei、Yang Chen、Haonan Chen、Hexiang Hu、Ge Zhang、Jie Fu、Alan Ritter 和 Wenhu Chen。UniIR：通用多模态信息检索器的训练与基准测试。载于第18届欧洲计算机视觉会议，第387–404页，意大利米兰，2024年。施普林格出版社。3[43] Hui Wu、Yupeng Gao、Xiaoxiao Guo、Ziad Al-Halah、Steven Rennie、Kristen Grauman 和 Rog´erio Feris。Fashion IQ：一个旨在通过自然语言反馈检索图像的新数据集。载于《IEEE计算机视觉与模式识别会议》（CVPR 2021），第11307–11317页。计算机视觉基金会 / IEEE，2021年。3[44] Bingbing Zhang、Yi Han 和 Xiaofei Han。基于预训练模型的电子商务平台多模态检索系统研究。《人工智能技术研究》，2(9)，2025年。1[45] Xin Zhang、Mingxin Li、Yanzhao Zhang、Dingkun Long、Yongqi Li、Yinghui Li、Pengjun Xie、Meishan Zhang、Wenjie Li、Min Zhang 等。SSRB：以自然语言直接查询海量异构半结构化数据。载于第三十九届神经信息处理系统年会数据集与基准赛道，2025年。2[46] Xin Zhang、Yanzhao Zhang、Wen Xie、Mingxin Li、Ziqi Dai、Dingkun Long、Pengjun Xie、Meishan Zhang、Wenjie Li 和 Min Zhang。跨越模态：借助多模态大语言模型改进通用多模态检索。载于《计算机视觉与模式识别会议论文集》，第9274–9285页，2025年。2, 5[47] Junjie Zhou、Zheng Liu、Shitao Xiao、Bo Zhao 和 Yongping Xiong。VISTA：用于通用多模态检索。在计算语言学协会第*62*届年会论文集（第*1*卷：长篇论文），第3185–3200页，泰国曼谷，2024年。计算语言学协会。3[48] Jianqun Zhou、Yuanlei Zheng、Wei Chen、Qianqian Zheng、Hui Su、Wei Zhang、Rui Meng 和 Xiaoyu Shen。《超越内容相关性：评估检索模型中的指令遵循》。 *arXiv* 预印本 *arXiv:2410.23841*，2024年。 2，3[49] Tianshuo Zhou、Sen Mei、Xinze Li、Zhenghao Liu、Chenyan Xiong、Zhiyuan Liu、Yu Gu 和 Ge Yu。《MARVEL：通过视觉模块插件释放稠密检索的多模态能力》。载于计算语言学协会第*62*届年会论文集（第*1*卷：长篇论文），第14608–14624页，泰国曼谷，2024年。计算语言学协会。 3

# 超越全局相似性：迈向细粒度多条件多模态检索

## 补充材料

## 7. 基线方法详情

对于本文使用的每个模型，表5 总结了参数规模、架构、是否使用显式检索指令、最大输入上下文长度以及底层骨干模型检查点。我们根据系统在两阶段流水线中的作用，将其分为检索器和重排序器。

表 5：实验中使用的检索器和重排序模型详情。规模表示每个模型的参数数量。架构表示我们设置中使用的模型类型（所有模型均以仅解码器模式运行）。指令说明在编码查询时是否添加任务专用自然语言前缀（例如，检索提示或生成 <emb>标记的模板），而非直接输入原始查询。最大长度表示实验中每个模型使用的最大输入长度（以词元计），通常与骨干网络的上下文窗口相符。骨干网络指明了构建各检索器或重排序器所基于的预训练检查点。

## 8. MCMR示例

表 6展示了五个MCMR子数据集中的示例，涵盖上装、下装、鞋类、珠宝和家具。

## 9. 完整结果

表7 报告了查询端模态消融研究的完整数值结果。对于主实验中评估的五种检索器，我们提供了三种查询模式下的详细分数：(i) 同时包含图像派生约束和文本派生约束的融合查询；(ii) 移除所有基于图像的约束后得到的纯文本查询；以及 (iii) 移除所有基于文本的约束后得到的纯图像查询。每个子表均报告了不同截断位置下的 Recall@K、MRR 和 NDCG@K，从而能够细致考察早期排名的精确率以及长程排序行为。这些完整结果清楚地表明，不同模型对跨模态证据移除的响应各不相同。

表 8 报告了 §4.3 中所述查询端组合约束研究的完整数值结果。对于每种检索器（CORAL、GME、LamRA、LLaVE 和 MM-Embed）以及文本和图像派生约束数量的每种配置，我们以百分比形式列出 Recall@K、平均倒数排名和 NDCG@K。我们考察了匹配的约束镶嵌方式，其中 $k_T{=}k_I \in \{1,2,3,4,5\}$ 同时保持融合候选池和评估协议与主要实验保持一致。正文重点讨论 $k_T{=}k_I \in \{2,3,4,5\}$ 的趋势；为求完整，此处还纳入了 1T+1I 配置。

## 10. 提示

### 10.1. 用于构建 MMR 的提示

**步骤 1：图像侧属性提取** 图 5展示了用于从每张产品图像生成细粒度视觉属性描述的提示模板。指示模型仅输出有图像依据的结构化特征，不推断隐藏或功能属性。

**步骤 2：文本侧描述生成** 图 6 展示了根据元数据生成产品文本描述的提示设计。该设计严格要求内容与视觉证据分离，并对价格和发布日期等关键属性的措辞进行标准化。

**步骤 3：通用泄漏检查器** 图7 展示了描述生成后用于检测视觉到文本泄漏的验证提示。它会标记 图像特征与 文本描述之间的任何精确重叠或改写重叠，以及类别不匹配。此步骤可确保在查询合成之前，文本侧描述严格以文本为依据。

**步骤 4：查询生成** 图8 展示了用于模拟真实客户查询的提示。模型按照严格的组合规则融合视觉和文本属性，确保每条查询都能自然表达跨模态信息，同时保持语言多样性和真实性。

**步骤 5：查询检查**。图9 展示了用于筛除低质量或不忠实查询的验证提示。模型会检查模态覆盖是否均衡（至少包含两个图像衍生匹配项和一个文本衍生匹配项）、价格和日期表述是否有效，以及规范化问题（例如单位不一致或品牌泄漏）。只有通过所有逻辑门槛的查询才会保留用于数据集构建。

**逐点**。图。10 展示了重排序过程中用于为查询—候选项对打分的逐点提示。

| 模型 | Size | 架构 | 指令 | 最大长度 | 骨干网络 |
|---|---|---|---|---|---|
| 检索器 | | | | | |
| GME-Qwen2-VL-7B-Instruct | 7B | 解码器 | Yes | 32K | Qwen2-VL-7B-Instruct |
| LLaVE-7B | 7B | 解码器 | Yes | 32K | LLaVA-OneVision-7B |
| LamRA-Ret-Qwen2.5VL-70亿参数 | 7B | 解码器 | Yes | 128K | Qwen2.5-VL-7B |
| MM-Embed | 8B | 解码器 | Yes | 32K | NV-Embed |
| CORAL | 3B | 解码器 | No | 128K | Qwen2.5-3B-Instruct |
| VLM2VEC | 4B | 解码器 | No | 131K | Phi3.5V |
| 多模态大语言模型重排序器 | | | | | |
| Qwen2.5-VL-7B-Instruct | 7B | 解码器 | Yes | 128K | Qwen2.5-VL-7B-Instruct |
| Qwen2.5-VL-32B-Instruct | 32B | 解码器 | Yes | 128K | Qwen2.5-VL-32B-Instruct |
| InternVL3-8B-Instruct | 32B | 解码器 | Yes | 32K | InternVL3-8B-Instruct |
| Qwen3-VL-4B-Instruct | 4B | 解码器 | Yes | 262K | Qwen3-VL-4B-Instruct |
| Qwen3-VL-8B-Instruct | 8B | 解码器 | Yes | 262K | Qwen3-VL-8B-Instruct |

表5. 实验中使用的检索器和重排序模型的详细信息。

**图像属性提取提示词**

你是一位一丝不苟的视觉标注员。
Task
1. 用一两个小写单词指出你看到的产品类别。2. 仅列出简短的外观描述词，准确描述产品的细微、客观视觉细节。规则

- 仅输出一个 JSON 数组：先写类别，再写描述词。
- 尽量提供 4–10 个描述词；如果能确定的更少，就只输出那些描述词（不要猜测）。
- 每个描述词都必须是 $\leq 8$ 个英文单词，全部使用小写和美式拼写。
- 只描述清晰可见的内容；不要提及尺寸、价格、发布日期、性能或舒适度相关的说法。
- 如果某个词暗示功能或技术（例如防水、透气、保暖、带缓冲的），除非该词的原文清晰可见地印在产品上，否则不要使用。

•重点关注：颜色、材料/纹理、局部图案/图形、制作工艺/部件/边缘、接缝/叠层/缝线、闭合方式/五金件（例如拉链、搭扣、鞋带、扣环）、形状/轮廓，以及可辨认的标志/文本。

- 如果出现可辨认的标志或文本，请准确写出其小写文本（不加引号）。不要编造品牌或印记。
- 排除背景或不属于产品的其他物品（例如其他服装、道具、身体部位）。
- 跳过任何无法百分之百确定可见的特征；不要推断被遮挡的细节。输出前，请确保在可见的情况下，至少涵盖以下 4 项： {金属颜色/表面处理}、 {宝石颜色及形状/切工}、 {镶嵌方式/排列方式}、 {带/链条/手链细节}、 {闭合方式/扣环}、 {雕刻/图案/叠层}、{标志/文本/印记}。

仅返回 JSON 数组（不含说明文字）。

图 5. 图像侧标注中使用的图像属性提取提示词。

## 11. 人工评估方案

我们开展了一项小规模人工研究，以验证生成的查询的自然度和属性保真度。我们从所有领域中随机抽取了100个目标产品。对于每个产品，我们向标注员A提供产品图像，其结构化元数据以及内部提取的细粒度属性，并要求标注员撰写一条用户在实际情境中可能会输入的自然语言搜索查询——且不向其提供模型生成的查询。

随后，标注者 B 在双盲条件下评估人工撰写的查询和模型生成的查询。

| 类别 | 查询文本 | 目标文本 | 目标图像 |
| --- | --- | --- | --- |
| 上装 | 我在寻找一件灰色的男士夹克，带有格纹和绿色点缀，L码。它应具备防水功能，带有连帽和前口袋，采用耐用的尼龙斜纹布制成，需要手洗。最好是于2013年左右推出，价格约为200美元。 | ”title”: ”哥伦比亚 男士 Whirlibird III 可拆卸夹克”, ”description”: ”三件夹克合而为一——一件温暖、防泼水的Omni-Heat 内衬；一件防水透气的且关键部位接缝压胶的外壳；以及两者的组合——为你带来极致灵活应对，保持温暖、干燥、得到保护，并在多变的冬季环境中感到舒适。”, “价格”: 200.0, ... |  |
| 下装 | 我想找一条男士棕色棉质牛仔裤，版型修身直筒。我如平前设计和拉链门襟，由柔软的100%棉质牛仔布，耐用，and可机洗。非常适合休闲或工作服，最好低于30美元，并发布于2021年前后。 | ”title”: ”World of Leggings 大码氨纶及膝男孩短裤 - 选购16种颜色”, ”description”:我们的及膝无缝大码男孩短裤为均码无缝尼龙氨纶加大码裤腿是每位女性打底裤衣橱中的必备单品。它们采用优质缝线，弹性极佳..... |  |
| 鞋类 | 我在找男士棕色皮革系带闭合的高帮工作靴并配有缓震 OrthoLite 鞋垫。它们应配备防滑鞋底，并符合ASTM EH 安全标准，以提供电气危险防护。我想买一双价格低于260 美元。 | ”标题”: ”Danner 男士 Bull Run 8 Ẅork 工装靴”, ”描述”: ”美国制造秉承可追溯至1932年的品质承诺，Bull Run 是一款实用型工装靴，设计经典永不过时，让你下班后依然保持时尚风格。全粒面皮革上装兼具坚韧与柔软。搭配坚固耐用的Danner 楔形外底，带来全天候持久舒适。”, “价格”: 259.95, ”features”: ”100% 皮革”, ... |  |
| 珠宝 | 我想找一条银色手链，配有古巴链和滑扣。请展示14K金镀925纯银的款式，价格约为450美元，约于2022年推出。 | ”title”: ”SAVEARTH DIAMONDS 5.80 Ct至10.90 Ct实验室培育莫桑石钻石 6 毫米宽古巴链14K金包925纯银男士项链纯银，16至30长度，颜色：G-H色，清晰度：VVS1净度”, ”描述”: ”珠宝有能力成为这样一件小小的事物，它能让你感到独一无二，.....”, ..... |  |
| 家具 | 我正在寻找一些可爱的乡村风万圣节家居装饰，我想要一个配有小型木制标牌的多层托盘布置以及价格低于 20 美元的秋季主题装饰。 | ”title”: ”CYNOSA 万圣节装饰万圣节多层托盘装饰 秋季装饰《女巫也疯狂》 我闻到了孩子的气息 呜！ 木制标牌和橙色格纹地精毛绒玩偶农舍乡村风多层托盘装饰，适用于家用桌子”, ”描述”: ”包装包括：1 个黑橙格纹格纹地精；1 个圆形标牌Ï闻闻孩子¨；1 个万圣节主题方形标牌十月 31¨；1 个黑橙格纹长方形标牌（呜！）；..... |  |

表6. 我们的MCMR基准测试中的检索样例。

(a) 融合(文本+图像)

| 模型 | R@1 | R@5 | R@10 | R@50 | MRR | N@5 | N@10 | N@50 |
|---|---|---|---|---|---|---|---|---|
| LLaVE | 22.00 | 38.00 | 46.00 | 73.00 | 29.09 | 30.53 | 33.10 | 39.39 |
| GME-Qwen2VL | 9.00 | 34.00 | 42.00 | 64.00 | 19.50 | 22.32 | 24.91 | 29.81 |
| LamRA-Qwen2.5VL | 13.00 | 31.00 | 39.00 | 63.00 | 21.14 | 22.73 | 25.43 | 30.67 |
| MM-EMBED | 20.00 | 40.00 | 47.00 | 64.00 | 27.58 | 29.94 | 32.19 | 35.88 |
| CORAL | 24.00 | 39.00 | 45.00 | 71.00 | 30.67 | 32.22 | 34.11 | 39.86 |

(b) 仅文本

| 模型 | R@1 | R@5 | R@10 | R@50 | MRR | N@5 | N@10 | N@50 |
|---|---|---|---|---|---|---|---|---|
| LLaVE | 2.00 | 12.00 | 17.00 | 30.00 | 6.54 | 7.40 | 9.02 | 11.98 |
| GME-Qwen2VL | 6.00 | 16.00 | 24.00 | 45.00 | 11.09 | 11.65 | 14.08 | 18.43 |
| LamRA-Qwen2.5VL | 2.00 | 8.00 | 15.00 | 27.00 | 4.77 | 4.92 | 7.10 | 9.68 |
| MM-EMBED | 12.00 | 29.00 | 33.00 | 47.00 | 18.72 | 20.90 | 22.18 | 25.47 |
| CORAL | 9.00 | 16.00 | 18.00 | 30.00 | 12.00 | 12.82 | 13.45 | 16.05 |

(c) 仅图像

| 模型 | R@1 | R@5 | R@10 | R@50 | MRR | N@5 | N@10 | N@50 |
|---|---|---|---|---|---|---|---|---|
| LLaVE | 20.00 | 33.00 | 41.00 | 65.00 | 26.04 | 27.06 | 29.55 | 35.07 |
| GME-Qwen2VL | 13.00 | 25.00 | 33.00 | 56.00 | 18.07 | 18.91 | 21.57 | 26.41 |
| LamRA-Qwen2.5VL | 12.00 | 30.00 | 38.00 | 60.00 | 19.11 | 20.94 | 23.57 | 28.42 |
| MM-EMBED | 14.00 | 28.00 | 33.00 | 50.00 | 19.42 | 21.09 | 22.64 | 26.59 |
| CORAL | 10.00 | 31.00 | 37.00 | 54.00 | 18.82 | 21.27 | 23.21 | 26.81 |

表 7. 检索 在 MCMR 上采用三种查询可见性机制（融合 / 仅文本 / 仅图像）的结果。数值为百分比 .

每个查询均按三个维度使用 1–5 分的李克特量表进行评估：(i) 属性正确性、(ii) 跨模态覆盖，以及 (iii)自然度和清晰度。标注者 B 还会选择哪个查询与目标产品更匹配。

总体而言，模型生成的查询在所有标准上的得分均与人工撰写的查询相当，仅在自然度方面略有差距。偏好分布几乎相同，表明标注者并未明显偏向任何一方。这些结果证实，所提出的生成流程能够生成高质量的多条件查询，适用于跨模态检索研究。

| 模型 | $k_T=k_I$ | Recall@K（%） | | | | 平均倒数排名（%） | NDCG@K（%） | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 5 | 10 | 50 | | 5 | 10 | 50 |
| CORAL | 1t1i | 9.00 | 15.00 | 22.00 | 41.00 | 12.38 | 12.30 | 14.61 | 18.84 |
| | 2t2i | 15.00 | 31.00 | 38.00 | 57.00 | 22.09 | 23.64 | 25.89 | 30.11 |
| | 3t3i | 21.00 | 36.00 | 40.00 | 62.00 | 27.25 | 29.06 | 30.34 | 35.10 |
| | 4t4i | 26.00 | 42.00 | 48.00 | 68.00 | 33.21 | 34.86 | 36.77 | 41.13 |
| | 5t5i | 33.00 | 50.00 | 57.00 | 79.00 | 40.19 | 41.94 | 44.19 | 48.93 |
| GME | 1t1i | 2.00 | 18.00 | 20.00 | 36.00 | 7.92 | 10.22 | 10.87 | 14.32 |
| | 2t2i | 10.00 | 23.00 | 28.00 | 46.00 | 15.75 | 17.03 | 18.70 | 22.67 |
| | 3t3i | 12.00 | 32.00 | 34.00 | 52.00 | 19.43 | 22.32 | 23.01 | 27.13 |
| | 4t4i | 18.00 | 39.00 | 42.00 | 59.00 | 25.78 | 28.73 | 29.73 | 33.55 |
| | 5t5i | 21.00 | 42.00 | 48.00 | 68.00 | 29.40 | 31.93 | 33.87 | 38.36 |
| LamRA | 1t1i | 5.00 | 12.00 | 21.00 | 42.00 | 8.51 | 8.49 | 11.38 | 15.94 |
| | 2t2i | 8.00 | 19.00 | 25.00 | 50.00 | 13.06 | 13.90 | 15.90 | 21.38 |
| | 3t3i | 8.00 | 24.00 | 31.00 | 53.00 | 14.78 | 16.31 | 18.67 | 23.80 |
| | 4t4i | 12.00 | 29.00 | 39.00 | 63.00 | 19.48 | 20.81 | 24.08 | 29.41 |
| | 5t5i | 19.00 | 35.00 | 44.00 | 70.00 | 26.66 | 27.84 | 30.79 | 36.52 |
| LLaVE | 1t1i | 5.00 | 23.00 | 29.00 | 50.00 | 12.55 | 14.57 | 16.49 | 21.14 |
| | 2t2i | 16.00 | 32.00 | 39.00 | 57.00 | 21.83 | 23.49 | 24.54 | 29.67 |
| | 3t3i | 19.00 | 37.00 | 43.00 | 68.00 | 25.98 | 28.06 | 30.04 | 35.45 |
| | 4t4i | 21.00 | 42.00 | 49.00 | 69.00 | 28.93 | 31.40 | 33.72 | 38.10 |
| | 5t5i | 27.00 | 50.00 | 59.00 | 72.00 | 35.89 | 38.50 | 41.35 | 44.28 |
| MM-EMBED | 1t1i | 8.00 | 18.00 | 20.00 | 40.00 | 12.06 | 13.36 | 13.98 | 18.57 |
| | 2t2i | 14.00 | 22.00 | 27.00 | 44.00 | 17.31 | 18.03 | 19.55 | 22.09 |
| | 3t3i | 19.00 | 29.00 | 34.00 | 49.00 | 22.96 | 23.96 | 25.55 | 28.80 |
| | 4t4i | 23.00 | 40.00 | 45.00 | 64.00 | 30.04 | 31.96 | 33.65 | 37.83 |
| | 5t5i | 30.00 | 43.00 | 48.00 | 68.00 | 36.30 | 37.43 | 39.15 | 43.67 |

表8. 查询端　　在组合约束（$k_T=k_I \in \{1,2,3,4,5\}$）下的模态消融。数值为百分比。

| 标注者 | 背景 | 英语熟练度 |
|---|---|---|
| 标注员A | 语言学学士 | 雅思 7.0 |
| 标注者 B | 计算机科学硕士 | 托福 99 |

表 9. 参与人工评估研究的标注者背景和资历。

| 指标 | 人工 | 生成 |
|---|---|---|
| 属性正确性 | 4.45 | 4.37 |
| 跨模态覆盖 | 4.29 | 4.28 |
| 自然度与清晰度 | 4.48 | 4.33 |
| 平均分 | 4.41 | 4.33 |
| 偏好率 | 49% | 47% |

表10. 人工撰写查询与生成查询的人工评估总结。

**文本侧描述生成**

任务：根据产品的 JSON 元数据生成简洁的英文描述。输入（仅使用文本字段；最后一个数组是禁止提及的列表，而非需要描述的内容）：- title：标题 - description：描述 - features：特征 - price：价格 - Date First Available：日期 - 禁止提及的视觉元素（来自图像特征；不要提及或改述）：图像 硬性规则

• 撰写 2–4 个清晰的句子，总字数为 80–120 个英文单词，且仅使用一个段落。
• 仅使用以下字段：title、description、features、price、Date First Available。忽略图像及任何视觉特征。
• 始终提及价格和首次上市时间，并严格使用以下措辞：– 价格：严格使用“priced at $X.XX”。如果缺失：“price information not provided”。– 日期：严格使用“released ¡Month¿ ¡Day¿, ¡Year¿”。如果缺失：“release date not provided”。
• 不得包含任何视觉属性（颜色、图案、图形、标志、口号、角色、形状、轮廓、表面处理、光泽、哑光、风格线索）。

• 将禁止提及的视觉元素中的每一项都视为禁用内容；不得包含这些词语或其改述。
• 仅关注文本来源的事实：产品类型、用途、材料、组件、制作工艺、闭合方式、护理、耐用性、明确标注的尺寸、测量数据、认证、标准、包装、保修、产地、制造、价格和发布日期。
• 排除运输、服务和营销口号；必要时将冗长的尺寸表压缩为简洁的范围。
• 仅输出最终段落，不要添加备注或推理。质量检查（内部）

• 删除任何与禁止提及的视觉元素重叠的内容，以及残留的视觉描述用语。
• 确保长度为 80–120 个词；如果内容过短且有相关事实，则补充文本中的非视觉事实（例如护理、尺寸、认证、产地）。

• 如果字段之间存在冲突，优先采用“features”而非“描述”。切勿编造事实。输入 JSON 模板 { "title": {title}, "description": {description}, "features": {features}, "price": {price}, "首次上市日期": {date}, "image_feature": {image} }

图6. 用于文本侧描述生成的提示模板。

**用于跨模态泄漏检测的提示**

仅返回一个有效的 JSON 对象，不要添加任何说明文字。

任务：检查图像特征中的任何内容是否以完全相同的短语、明确的释义或同义词出现在文本描述中。使用不区分大小写的匹配。将连字符和空格视为等价。处理简单的单数或复数形式。

类别：如果第一个图像特征项目是通用类别，则在统计泄漏时忽略它，但仅当文本描述明确指出了不同类别时，才将类别冲突=设为 true。 -

消歧：不要将“925纯银”等材料规格视为与视觉色调“银色金属”的匹配。应采取保守标准：仅当含义相同时，才标记为释义。

输出 JSON 架构：“leak”：true、false，“matches”：[“feature”：str，“type”：“exact、paraphrase”，“evidence”：str], “category conflict”：true、false

输入：图像特征：[...] 文本描述：“...”

现在返回 JSON。

图7. 跨模态泄漏检测提示模板。

**查询生成提示**

你是一位真实的购物者，正在大型电子商务网站的搜索框中输入内容。
Task
用自然的第一人称语气写出一个流畅的搜索查询。使用 2-3 个句子，全文长度控制在 35-60 个词。
来源
• 描述：desc
• 图像标签：图像
硬性规则——必须全部满足
• 混合两个来源 -¿ 恰好使用 2-3 个视觉标签和恰好 2-3 个描述事实，且不得有重叠。
  – 如果移除重叠内容后剩余的描述事实少于 3 个，则全部使用剩余内容，不得自行编造。
  – 如果图像标签中出现清晰的图案，请指定底色和图案颜色，并且只能使用一个通用的图案术语，例如“印花图案”“图形图案”“条纹图案”。否则仅提及底色。
  – 仅当存在目标尺码时，准确提及一个目标尺码，切勿列出多个尺码或尺码范围。
• 价格格式 仅当来源中出现价格时，使用“大约”或“低于”添加一个近似价格，切勿使用确切价格。
• 发布时间格式 仅当来源中出现日期时，使用相对或近似表达添加一个大致发布时间，来源中 , 切勿使用确切日期。
• 全程使用第一人称语气。
• 不要提供这两个来源以外的任何细节。
• 只输出查询文本，不要使用项目符号、引号或反斜杠。
材质描述
• 当描述列出具体的纤维含量百分比时，请将其改写为定性短语，不含数字，也不含百分号，例如“富含棉”“含有一些聚酯纤维”“略含氨纶”“全棉”“聚酯纤维混纺”、“高尼龙含量”、“羊毛混纺”。每个此类材质短语都算作一个描述事实。
风格提示
• 优先突出面料、版型、闭合方式和护理信息，少写营销套话。
• 变换措辞，让每条查询都显得独特。
• 形容词应保持客观，例如“轻薄的棉质针织布”，而不是“超级舒适”。
输入 JSON
“描述”：描述 ,
“图像特征”：图像
输出
<仅查询文本>

图 8. 查询生成的提示模板。

**Prompt for Query Quality Judge**

You are a strict e-commerce query judge. Output ONE valid JSON object only, no extra text.
Inputs: query, image_feature (list of visual phrases), text_desc (metadata), candidate_price, candidate_year.
1. Coverage Check whether the query is well supported by image feature and text desc. Copy phrases from these sources; do not invent new facts. Mark coverage as OK only if you can find at least two image-based cues and one text-based cue that clearly match the query.
2. Price If the query mentions a price or budget, decide whether candidate_price is roughly consistent with that intent (not obviously too cheap or too expensive). If no clear price phrase or candidate_price is missing, treat price as in_range.
3. Date If the query mentions a year or "new this year", decide whether candidate_year and YEAR_NOW are broadly consistent. Otherwise treat date as in_range.
4. Norm flags and rewrite Set simple flags for unit mixing, brand leakage, overly generic or overly narrow queries. rewrite_minimal may contain a lightly edited version of the query that only fixes clear issues without adding new information.
Return the final JSON object now.

图 9. 查询质量评判器的提示词。

**Pointwise Prompt**

System.
You are a strict relevance judge. Given a user query and a candidate (image + textual attributes), answer with a single token: True or False. True means the candidate matches the query faithfully. False otherwise. No other words, no punctuation.
User. <image>
Query: {query_text} Candidate: {candidate_text} Does the candidate match the query? True or False

图 10. 用于逐点评分的提示词。