即将发表于《第19届 *IEEE Pacific Visualization Conference* 论文集》（*PacificVis 2026*），澳大利亚悉尼（会议专题）。

arXiv:2602.13402v1 [cs.HC] 13 Feb 2026

InfoCIR：面向组合图像检索的多媒体分析

Ioannis Dravilas*, †,‡ Ioannis Kapetangeorgis*,‡ Anastasios Latsoudis*,‡ Conor McCarthy§Gonc. alo Marcelino§ Marcel Worring§

阿姆斯特丹大学，荷兰

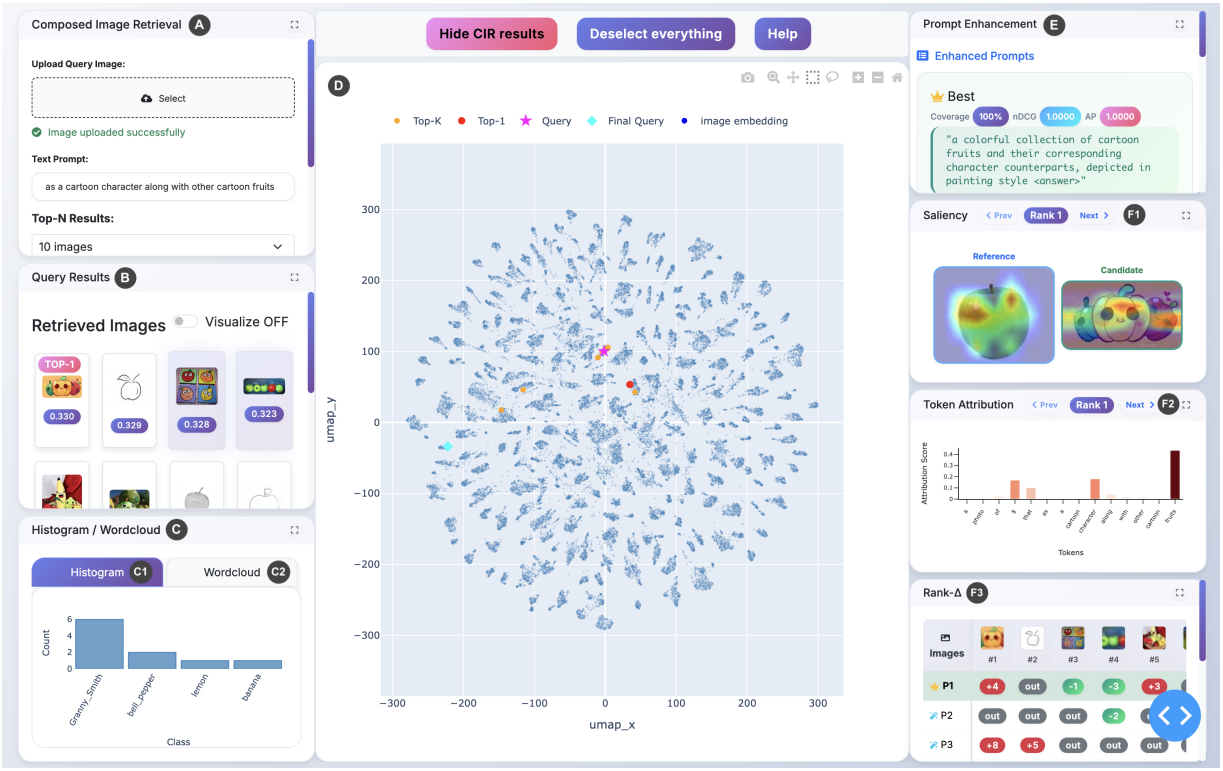

图1：我们的系统用户界面由六个主要面板组成。(A) 组合图像检索面板允许用户输入图像和文本提示，以启动检索过程并进行基线搜索。(B) 查询结果面板按相似度显示排名靠前的k 幅图像；用户可以点击任意图像，将其标记为提示词优化的理想目标。(C) 直方图/词云面板包含类别频率直方图（C_1）和词云（C_2），用于汇总当前排名靠前的k项中的标签。(D) 中央的嵌入视图展示数据集的二维UMAP投影，并突出显示参考图像、组合查询嵌入以及排名靠前的k 个结果。(E) 提示增强面板利用大语言模型，根据用户选定的理想目标图像提出备选提示，并展示检索指标。(F) 解释面板通过显著性图（F_1）、词元归因条形图（F_2）和 Rank-Δ 热力图（F_3）展示模型归因。

摘要

组合图像检索（CIR）允许用户通过将参考图像与描述所需修改的文本提示相结合来搜索图像。

*本文作者贡献相同。
†通讯作者：Ioannis Dravilas
‡电子邮件：{ioannis.dravilas、ioannis.kapetangeorgis、anasta-sios.latsoudis}@student.uva.nl §电子邮件： {c.t.mccarthy、g.barretoferreiramarcelino、m.worring}@uva.nl © 2026 IEEE。允许个人使用本材料。所有其他用途均须获得 IEEE 的许可，无论是在当前媒体还是未来媒体中，包括为广告或宣传目的重新印刷或再发布本材料、创作新的汇编作品、转售或重新分发至服务器或列表，或在其他作品中重复使用本作品的任何受版权保护的组成部分。

描述所需的修改。虽然 CLIP 等视觉语言模型通过将多种模态嵌入联合空间，让这项任务广受关注，但开发者仍缺少能够揭示多模态提示如何与嵌入空间交互，以及措辞上的细微变化为何会显著改变结果的工具。我们提出 InfoCIR，这是一种可视分析系统，通过在单一的交互式仪表板中结合检索、可解释性和提示工程来弥补这一不足。InfoCIR 集成了最先进的组合图像检索后端（SEARLE [1]），并配备六面板界面：(i) 允许用户组合图像 + 文本查询；(ii) 使用统一流形近似与投影（UMAP），将排名靠前的k 个结果投影到低维空间，以支持空间推理；(iii) 叠加基于相似度的显著性图和基于梯度的词元归因条，以提供局部解释；(iv) 利用大语言模型驱动的提示增强器生成反事实变体，并可视化这些变体如何影响用户选定目标图像的排名。基于 Plotly-Dash 构建的模块化架构，只需极少

的工作量即可接入新模型、数据集和归因方法。我们认为，InfoCIR 有助于诊断检索失败、指导提示增强，并加快模型开发过程中的洞察生成。所有源代码均可用于复现演示，详见 https://github.com/giannhskp/ InfoCIR。

索引词： 组合图像检索、提示增强、显著性图、词元级归因、神经排序。

1 引言

组合图像检索（CIR）旨在检索并返回一个排序后的图像列表，其中的图像与参考图像相似，但在自然语言提示所描述的特定方面有所不同 [43]。与通常仅依赖视觉相似性或文本查询的传统图像搜索不同，组合图像检索将这两种模态融合为一个查询。例如，用户可能会请求与森林照片相似的图像，但要求“呈现秋日色彩”或“处于夜晚”。完成这一任务需要系统理解提示中描述的变换，并将其应用于视觉概念。随着 CLIP 等视觉语言模型的问世，组合图像检索在多媒体检索和计算机视觉领域日益受到关注 [37, 28], ； 这类模型将图像和文本嵌入联合特征空间 [30], 。联合图像-文本嵌入构成了许多近期组合图像检索模型的基础，其中包括利用预训练组件、无需进一步任务专属训练的零样本系统 [12, 4, 50]。

在实践中，组合图像检索仍然难以使用。现有工作流严重依赖“盲目”试错：开发者在诊断出较低的 Recall@K 指标后，会在不了解底层失败机制的情况下随机调整提示或参考图像。尽管已有用于嵌入可视化 [36] 和单模态可解释性 [34], 的通用工具，但它们因两个不同原因而不足以应对组合图像检索。首先，通用嵌入投影将查询视为静态数据点，无法可视化由文本修饰引起的向量偏移，而这正是组合图像检索的核心机制。其次，标准显著性图往往孤立地可视化视觉特征，无法揭示检索模型何时以及为何忽略文本条件而偏向视觉主导（例如，尽管提示为“红色”，却检索出“青苹果”）。此外，正如我们在分析中所观察到的，UMAP 等标准降维技术在基于 CLIP 的空间中可能受到风格偏差的影响，即图像按艺术风格而非与用户相关的语义类别进行聚类。这些局限性既阻碍了系统调试，也阻碍了提示工程，尤其是在需要对图像语义进行精确控制的应用中 [38]。

为弥补这一空白，我们提出了 InfoCIR，这是一种旨在支持组合图像检索模型探索、诊断与改进的可视分析系统。InfoCIR 遵循设计研究方法开发，面向在诊断检索失败和迭代优化多模态查询方面反复面临挑战的开发者与研究人员。

如图 1所示，该界面集成了六个相互关联的组件（A-F），支持交互式探索。用户可以选择参考图像和文本提示，查看前 k 个检索结果，并通过统一流形近似与投影（UMAP）查看它们在嵌入空间中的分布 [24], , 识别影响模型行为的关键因素。系统叠加显示显著性图、词元归因和提示变体，帮助用户理解并改进搜索过程。各组件的详细说明请参见图 1 的图注。

我们将 InfoCIR 明确定位为一项系统贡献，强调其在弥合黑箱检索模型与人类意图之间的差距方面的完整性与实用性。我们的目标是将检索、可解释性和提示工程转化为连贯的工作流程，以支持人在回路推理。我们的具体贡献包括：

- A 统一的可视分析架构，集成了检索，归因和提示工程，从而形成“诊断—比较—增强”的反馈循环，以解决孤立且“盲目”的试错工作流程所存在的局限。
- 一种交互式相关性反馈工作流程，将大语言模型生成的提示变体与通过显著性图、词元级归因以及新颖的 Ideal-anchoredRank-Δ热力图进行的可视化验证相结合，以验证提示的影响。
- 一种采用去偏 UMAP 方法的监督式度量投影流程，可将结果、参考图像和查询投影到同一空间，专门用于克服风格偏差，从而实现更准确的邻域分析。
- 一项对照用户研究表明，与基线组合图像检索界面相比，InfoCIR 显著提高了检索效率和成功率。

这些功能共同使 InfoCIR 成为一款实用工具，能够以可解释且交互的方式诊断检索失败、理解模型行为并优化组合图像检索工作流。

2 相关工作

本节回顾了塑造本系统设计与功能的五个研究领域：（1）**面向检索系统的可视分析**提供用于检视排名动态、分析查询影响并优化结果的交互工具。（2）**降维与嵌入可视化**将高维嵌入投影到可导航空间，使用户能够检视结构模式与模型行为。（3）**用于检索解释的视觉显著性**突出对相似度分数影响最大的图像区域和文本线索，从而揭示模型的决策过程。（4）**提示增强与词元归因**将排序变化与单个词语关联，并为查询优化提供有针对性且可解释的建议。（5）**本研究的定位**展示了该系统如何统一并扩展这些方向，以提供可解释且交互式的组合图像检索工作流程。

2.1 检索系统的可视分析

近期的检索研究日益重视交互式可视化分析，以揭示排序行为、查询构建的影响以及反馈回路。早期系统（如 RankExplorer[35]）将排序随时间的演变可视化，而 Colorslope[44]则借助斜率图和热力图扩展了这一概念，以捕捉密集的排序轨迹。这些工具为后续系统奠定了基础，帮助用户理解排序器如何响应不断变化的查询而呈现动态行为。

一些系统探索了更高层次的基于提示的交互。PromptMagician [13] 支持关键词推荐和用于文生图的多级提示编辑，并提供了一个实用界面，便于用户试验组合查询。

另一个重要方向是反事实分析。CREDENCE [31] 引入了查询与文档编辑，以解释排序变化；近期的可编辑搜索解释框架 [49, 5]则在此基础上进一步拓展。在图像检索中，研究表明，即使只使用一个精心挑选的关键词，也能显著改善基于 CLIP 的结果，且无需重新训练 [25]。

这些进展表明，评估方式正从静态评估转向交互式系统，帮助用户分析排名动态、探索提示效应，并理解检索行为——这些原则正是我们系统所采纳并拓展的。

2.2 降维与 EmbeddingVis 可视化

要以有意义的方式可视化高维数据，需要采用能够保留

原始空间结构的降维技术。主成分分析（PCA）等经典方法 [18] 和等距映射（Isomap） [39] 分别通过保留方差或测地距离来降低维度。较新的方法，如 t 分布随机邻居嵌入（t-SNE）和 UMAP，侧重于保留局部邻域并提升可扩展性，因此适用于大规模视觉-语言嵌入 [41, 24]。其他技术包括三元组映射（TriMap） [3]，（采用三元组约束）以及成对受控流形近似与投影（PaCMAP） [46]，（通过分阶段优化平衡全局与局部结构），这些方法拓展了基于投影进行分析的设计空间。这些方法已应用于多媒体任务，范围涵盖交互式音乐探索 [40]和快速图像标注， [23]，直至密集文本检索中的聚类验证[22]。在我们的系统中，我们采用这一策略，通过调整后的 UMAP 投影，在单一且易于理解的视图中呈现嵌入空间，以支持检索驱动的诊断。

在这些降维方法的基础上，嵌入可视化工具使用户能够以交互方式探索和比较模型表示。TensorBoard 的 Embedding Projector 引入了支持刷选和标签着色的交互式 PCA、t-SNE 和 UMAP 图，为高维嵌入分析提供了首批实用界面之一 [36]。EmbeddingVis 将这一功能扩展为支持并排模型比较[21]，，而 WizMap 则借助四叉树摘要和地图式导航，将这些交互扩展至数百万个点 [47]。尽管这些工具很有用，但大多数工具仍与模型和任务无关，将每个嵌入视为孤立的点。它们缺少检索专用信号，例如排名位置、提示演变和理想图像锚点，而这些信号都已集成到我们的系统中 [17, 53]。

2.3 用于检索解释的视觉显著性

视觉显著性通过突出显示最具影响力的区域或特征，帮助用户理解检索模型为何将某些图像判定为相关。基于相似度的方法（如 [10] ）会扰动候选图像，以识别驱动相似度分数的区域；而像 [42] 这样的交互式框架，则会将这些映射整合到反馈回路中，从而支持实时提示或裁剪调整。基于梯度的技术将显著性扩展到多模态检索：Grad-ECLIP [54] 将 Grad-CAM [34]改编应用于 CLIP，通过对图像塔和文本塔进行反向传播，生成联合热力图和令牌级梯度；SANE [27] 则将图像块级显著性与语义属性关联起来，从而提升可解释性。在组合图像检索（CIR）场景中，查询结合了视觉和文本线索，这些技术有助于用户理解模型如何对齐不同模态。我们的系统引入并改编了 Grad-ECLIP，以支持检索后的解释：每次搜索后，热力图都会叠加在参考图像或任意检索图像上，将显著区域与起影响作用的提示词元关联起来，并引导有针对性的提示增强。

2.4 提示增强与词元归因

早期的反事实研究表明，微小的查询修改可以揭示某些项目为何出现在错误的排名位置 [5]。PromptBench[55] 和Counterfact [7]将这一思路拓展到语言模型，通过将排名变化归因于单个词元，而非整个提示词。

在多模态模型的背景下，Grad-ECLIP、SANE 和 PromptExp 等基于梯度的方法，分别针对 CLIP、图文任务和大语言模型任务，计算每个子词的带符号影响分数 [54, 27, 11]。这些研究表明，与提示级显著性相比，词元级信号能更有效地揭示幻觉属性等问题。

尽管 RankExplorer [35] 和 Colorslope [44] 等可视化摘要工具能够展示提示词编辑前后的排名变化，却无法将这些变化与词元级原因联系起来。我们采用其以颜色编码的 Rank-Δ 设计，同时为每一行配上一个条形图，展示各个词元的梯度，从而形成因果闭环，揭示排名变化及其对应的致因词语。

通过将词元归因与排名变化可视化相结合，我们的系统将最初为纯文本模型开发的解释技术拓展至组合图像检索的多模态场景。这种整合将反事实查询编辑与基于显著性的解释联系起来，使用户能够追溯排名变化所对应的具体词语及其相应的视觉证据。

2.5 我们的工作定位

现有的组合图像检索模型 [1, 12] 往往侧重于孤立的组件，忽视了采用集成方法的必要性。视觉嵌入工具通常忽略检索上下文和提示语义，而排名界面强调准确性，却不提供可解释性。降维方法仍与模型无关，因此难以深入了解检索行为；提示增强工具通常也只支持单一模态，且缺乏充分说明。

我们的系统通过在单一分析平台中整合检索排名、提示演变和词元归因，弥补了这些不足，从而能够直接比较不同的组合图像检索模型。据我们所知，这是首个同时支持分析嵌入空间、查询动态和模型行为的系统。

3 设计依据与目标

InfoCIR 的设计源于这样一种需求：当查询在检索过程中不断变化时，我们需要理解复杂的模型行为。我们的系统并不将检索视为静态操作，而是将其看作一种迭代式对话，用户通过不断完善多模态输入，使模型的内部表征与其特定意图相契合。本节将形式化定义检索与增强任务，并概述指导系统交互功能设计的目标。

3.1 形式化任务

组合图像检索（CIR）旨在解决这样一个问题：检索与给定参考图像相似、但按照文本修改有所不同的图像。与传统图像检索不同，CIR 要求系统理解自然语言转换，并将其应用于视觉输入，将两种模态融合为一个语义明确的查询。此过程通常涉及将参考图像映射到自然语言嵌入空间，这一过程称为文本反演，以便与文本提示无缝融合。

形式上，给定：
- 一张参考图像 $I\mathrm{ref} \in \mathscr{D}$（查询图像），
- 一个自然语言修饰语T（相对描述），
- 一个包含 $\mathscr{D}$ 图像及其预计算嵌入的数据库，

目标是检索一个排名列表，其中包含与I_{ref}相似、但不包含T所描述修改的前k 张目标图像。这一任务可通过在联合多模态嵌入空间中进行最近邻搜索来实现：

$$\left(I_{(1)},\ldots,I_{(k)}\right) = \underset{I_j \in \mathscr{D}}{\arg\operatorname{sort}}\ \operatorname{sim}\Big(f_c(I_{\mathrm{ref}},T), f_i(I_j)\Big),$$

其中，f_c 是一个组合函数，用于将图像和文本融合为查询嵌入；f_i 是数据库图像的固定编码器。

提示增强是指改写原始文本修饰语T ，使组合查询能够更准确地体现用户意图。初次搜索后，用户会从排名前k 的结果中选出一张或多张理想图像I^* 。随后，基于参考图像I_{ref}的视觉上下文、原始文本修饰语T以及所选理想图像的语义内容，基于语言模型的编辑器 $\mathscr{R}$ 会生成一小组备选提示：

$$T' = \mathscr{R}\big(I_{\text{ref}}, T, I^{*}\big).$$

检索引擎会重新评估每个候选提示T'，以确定其相对于原始查询对I^{*} 排名的影响。这种单次改写旨在通过使文本修饰语更好地匹配目标图像的视觉特征来提升检索质量，无需模型重新训练。

3.2 交互分析

为了推动面向组合图像检索（CIR）的可视分析系统设计，我们分析用户在反复构建和优化多模态查询时如何进行推理。用户并非只提交一次查询并接受结果，而是持续进行意义建构循环，这一循环借鉴了 Pike 等人的交互科学 [26], ，要求用户诊断检索失败、比较可能的解决方案，并完善查询以改善结果。在这一循环中，每次交互都相当于一次微观假设检验，可更新用户对系统如何处理多模态输入的心智模型。这种结构化的“诊断—比较—完善”工作流程，既能带来更好的检索结果，也能加深用户对模型行为的理解。

在整个过程中，用户会追求一系列反复出现的高层次交互意图，这些意图构成了分析组合图像检索结果的框架。这些意图描述了用户的宏观目标，也反映了 Pike 等人所述意义建构的认知阶段。我们围绕所提出工作流程的三个主要要素，对这些目标进行归纳：

- 诊断：用户检查系统为何会表现出观察到的行为，以识别检索失败。这包括评估结果的语义连贯性、找出分布中的结构离群点，以及检测可能导致检索偏差的过度代表的概念。
- 比较：用户评估检索引擎的不同状态，以确定哪种表述最有效。这包括切换不同的提示变体，观察哪种能产生更连贯的语义邻域，并比较目标图像在嵌入空间中的位置变化。
- 增强：用户反复修改组合提示，以更好地表达目标概念。每次修改都代表一个新的假设，用户通过观察目标图像的排名变化，并借助自动化建议来调整搜索焦点，从而验证该假设。

这些要素紧密耦合，通常发生在同一个意义建构过程中。用户会在诊断、比较和增强之间灵活切换，以迭代改进组合图像检索的结果。

3.3 设计目标

基于上述“诊断—比较—增强”工作流程，我们制定了三个设计目标，将这些认知阶段转化为具体的系统能力。这些目标指导界面设计，确保每个视觉组件都直接支持用户推理过程中的一个特定阶段。

DG1 诊断 支持快速评估和诊断检索行为。系统必须提供可解释的视觉信号，呈现整体语义结构和局部模型决策，帮助用户迅速判断检索相关性，并找出推理障碍的原因。

DG2 比较 支持比较不同提示变体和检索结果。系统应明确呈现排名动态和语义邻近度的变化在不同查询之间进行比较，使用户能够对照结果，并确定哪种表述最能体现预期概念。

DG3 增强 支持迭代式提示增强。系统必须通过提供智能查询指导和对优化内容的即时反馈，促进快速实验，使用户能够高效测试新假设，而无需盲目反复试错。

4 方法

所提出的组合图像检索（CIR）交互式多媒体分析框架，将多模态深度学习与协同可视化表示相结合，以支持提示优化和模型内省。该系统以用户反馈、可解释可视化和提示增强策略为核心，并利用大语言模型（LLM）。系统遵循模块化流水线架构，如图 2所示，其设计受 Worring 等人启发。 [48], 该架构包含输入、模型选择、检索、投影、可视化和提示增强等不同阶段。以下将介绍系统的方法组成部分。

4.1 系统概述

该流程始于用户提供的输入查询，其中包含一张参考图像和一段自然语言提示。它集成了 SEARLE [1]组合图像检索模型，该模型使用 CLIP 计算联合图像-文本嵌入，以支持比较检索和可视化工作流程。

这些嵌入用于根据相似度对目标数据库中的图像进行排序。系统通过一组协同视图展示相似度最高的前k 个结果，其中包括二维嵌入视图、图像画廊、类别分布直方图和语义词云。集成的可解释提示增强模块使用大语言模型生成不同的文本变体，根据用户选择的理想图像对它们重新排序，并即时通过显著性图、紧凑的 Rank-Δ 热力图和词元影响摘要展示其影响，从而为用户提供快速、透明的迭代流程，以改进查询。

该系统使用 Plotly 和 Dash 框架实现，支持高度交互式的网页可视化，并可与基于 Python 的机器学习组件无缝集成，如系统概述图所示（图 1）。

4.2 输入与模型选择

系统接受两个主要输入：（1）上传的参考图像；（2）描述所需变换或特征的文本提示。第三个可选（但强烈建议提供）的输入是理想图像：用户从初始排名靠前的k 张图像中选择一张或多张符合其意图的图像。这些图像将成为固定锚点，用于评估所有提示变体，从而支持第 4.6 节所述的细粒度反事实诊断。

在检索后端，InfoCIR 实现了 SEARLE，这是一种零样本模型，通过将参考图像转换为 CLIP 文本嵌入空间中的伪词元来执行文本反演。这种架构能够将图像与提示进行组合融合，形成联合查询向量。随后，系统计算该向量与预先计算的数据库嵌入之间的余弦相似度，从而生成最终的图像排名。

4.3 检索与相似度计算

给定组合查询嵌入后，系统使用余弦相似度对所有数据库图像进行排序。选择排名前k 的图像进行进一步可视化，从而形成排序结果集。除图像外，系统还会返回相似度分数，以及可用时关联的类别标签或标签，以供进一步分析。

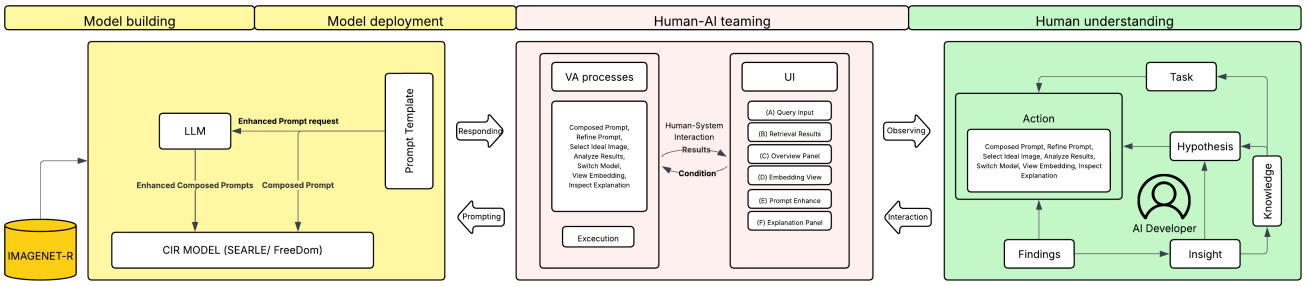

图2：该架构通过用户界面面板（A-F）将提示词组合、基于嵌入的检索和交互式可视化联系起来，使用户能够优化提示词、解读输出并获取见解。遵循 Wang 等人的做法， [45], 我们整合了语义融合（CLIP 空间）、视觉整合（显著性和词元叠加）以及跨视图链接（同步视图）。

4.4 降维

为了便于对嵌入结构进行定性拓扑检查，我们将排名前k 的结果对应的高维向量投影到二维空间。此投影旨在通过视觉邻域验证辅助诊断，而非基于嵌入距离进行精确的度量推理。

在这一领域，UMAP 和 t-SNE 是高维可视化的标准技术。基于工作流程的三项需求，我们采用 UMAP：

1. **参数稳定性（样本外扩展）：** 我们的系统会将新提示词实时嵌入固定的参考映射中。UMAP 的参数化形式支持对未见数据进行稳定的变换操作 [32], 而标准 t-SNE 是非参数方法，通常需要重新运行投影。

2. **全局结构保留：** 与 t-SNE 类似，UMAP 强调局部邻域保留，这对于识别语义相关的嵌入组至关重要 [24]。不过，UMAP 通常能生成更连贯、粒度较粗的簇组 [8]。

3. **用于快速语义验证的簇密度：** UMAP 往往能生成更密集、更紧凑的簇，且类间分离更加清晰 [19]。这直接支持我们的 DG1 目标，即让用户能够快速验证提示是否落入预期的语义类别。

所得的二维坐标用于在嵌入视图中定位图像。

为确保可视化支持语义调试，我们在最终投影前应用类别投影保持流程（图 3）。原始 CLIP 嵌入常因风格或背景特征而呈现显著的类别间重叠。我们明确指出，此流程并非模型训练或适配，而是一种在可视化前应用的启发式、任务驱动的语义预对齐策略。

该流程包含三个顺序过滤步骤：(A) 通过基于PCA的过滤进行**风格去偏，**保留语义可判别性较高的维度；(B)**对比去偏** [15], ，促使嵌入与其类别原型更紧密地对齐；以及 (C) 通过**独立成分分析 (ICA)** [29] ，将语义信号与风格噪声分离。

该启发式方法流程旨在提升类别级可分性，使所得可视化反映与用户相关的语义类别，而非无关的风格变化。最后，我们强调面板D用于定性的拓扑检查，而非精确的度量推理。二维投影中的距离和方向不应被解释为

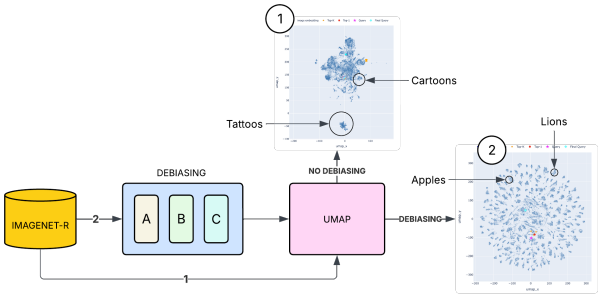

图3：类别投影保持流程的概念概览。ImageNet-R嵌入经过三阶段过滤过程：(A) 风格去偏、(B) 对比去偏，以及 (C) 基于ICA的替代投影，随后进行最终的UMAP可视化。与未处理的投影（1）相比，处理后的版本（2）展现出更高的类别可分性，并降低了风格伪影的影响，有助于对类别归属进行视觉验证。

语义幅度或线性嵌入轨迹的忠实表征。

4.5 交互式可视化

检索到的图像与投影后的图像呈现在六个主要面板（A–F）中，如图1所示，提供多个协同视图，以支持诊断—比较—增强工作流程并实现设计目标：

• **组合图像检索面板（A）：** 允许用户输入参考图像和文本提示，同时选择要检索的图像数量k 。

• **查询结果面板（B）：** 在图像画廊中按与查询的相似度排序显示检索图像。该视图支持对排序结果进行快速检查（DG1），并与嵌入视图保持同步。

•**直方图/词云面板（C）：** 包含一个类别频率直方图（$C1$），用于分析类别占主导的情况；以及一个词云（$C2$），用于概括当前 top-k中的语义趋势。这些小组件帮助用户诊断结果中过度呈现的概念或偏差，直接回应 DG1。

• **嵌入视图（D）：** 中央面板显示数据集的二维UMAP投影。空间上的邻近关系反映的是粗略的语义邻域关系，

而非精确的相似度大小，用户可据此观察类别层面的聚类、分离或离群点（DG1）。交互式选择可显示与相应投影相关的元数据。

•提示增强面板（E）： 根据用户选定的理想图像，提出由大语言模型生成的其他提示变体T'。用户可以交互式地比较这些变体，观察其对检索性能的影响，从而促进比较（DG2）和迭代优化（DG3）。

• 解释面板（F）： 提供多层次归因，以揭示模型行为的成因，支持诊断推理（DG1）。其中包括：

– 显著性图（F1）： 叠加基于梯度的视觉显著性，以突出显示影响较大的图像区域，将视觉证据与文本输入相关联，帮助用户理解模型关注的重点。

– 词元归因图（F2）： 量化提示词中的每个词元对检索理想图像的贡献程度。

– 排名Δ 热力图（F3）： 可视化每个变体如何影响最初检索到的排名靠前的k 张图像的位置。单元格颜色表示排名变化（绿色表示上升，红色表示下降），用户可以通过直接比较（DG2）找出最有效的提示词，从而改进搜索（DG3）。

这些面板相互关联，以支持多维度探索。在一个视图中的任何交互（例如点击查询结果面板中的图像）都会自动更新相关的分析视图和说明视图。

4.6 可解释的提示增强

InfoCIR 中的提示增强模块是由用户引导的细化函数实现 $\mathscr{R}$。这一工作流程中的一项具体挑战是“冷启动”问题：初始查询未能在前k 个结果中检索到任何相关图像，无法将其用作反馈锚点。InfoCIR 通过利用手动组合图像检索面板（面板 A）作为主要的切换工具来解决这一问题：在零命中场景中，用户首先调整初始修饰词 T，将搜索转向相关的语义邻域。一旦前k个结果中出现可行的目标，用户便选择一个或多个理想目标图像 I^*。这些选定的图像将作为增强锚点：系统利用开放权重的大语言模型（LLM）mistralai/Mistral-7B-Instruct-v0.2，生成一组备选提示词 T'，这些提示词以I_{ref}的内容、T以及I^*的视觉特征为依据。对于每个拟议变体 T'，系统都会重新排列结果，并通过理想目标锚定的排名-Δ 热力图将I^* 的排名变化可视化。该热力图能即时反馈每个新提示词如何影响目标的排名，帮助用户一目了然地比较提示词的效果。由于热力图将初始前k 个结果固定为稳定的参照集，因此能够展示局部排序变化，并在不同变体间保持视觉连续性。

除了提供提示词建议外，该模块还提供了用于诊断某些提示词为何表现更佳的工具。词元归因图（使用基于梯度的分析 [11]）突出显示了T'中对相似度分数影响最大的词语$s = \mathrm{sim}(f_c(I_{\text{ref}}, T'), f_i(I^*))$。具体而言，我们计算相对于每个输入词元的$s$梯度。这些梯度的幅度表示每个词语对匹配结果的贡献，并通过条形图进行可视化，以引导用户关注高影响力词项。这有助于了解模型对文本查询的推理过程，并提示用户在进一步优化时应强调或调整哪些关键词（见图 4）。

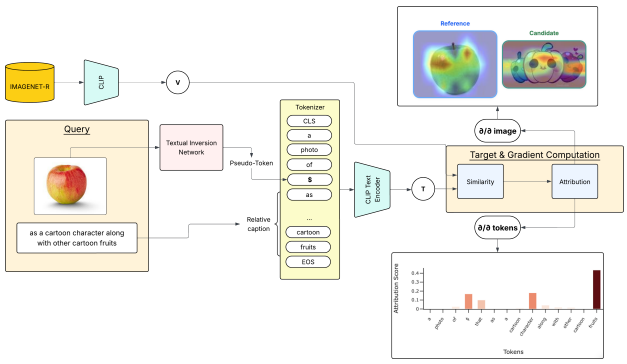

图 4：组合图像检索中视觉归因与文本归因的解释流程。系统接收一张参考图像I_r 和一段相对描述C，并将二者组合成提示词“a photo of $\varphi(I_r)$ that$C$”，其中 $\varphi(I_r)$ 是通过文本反演网络生成的伪词元。该提示词经分词后，由 CLIP 的文本编码器编码，生成文本嵌入T。数据集（例如 ImageNet-R）中的候选图像被编码为图像嵌入V，检索则基于余弦相似度$s = \cos(V,T)$进行。为生成解释，系统计算相似度分数s 相对于图像或文本词元的梯度。由此得到的梯度生成视觉显著图（右上）和词元归因分数（右下），帮助用户了解图像或提示词中的哪些部分对检索结果影响最大。

作为词元级解释的补充，我们生成显著性图，以展示I_{ref} 或I^*中最具影响力的区域。我们采用基于梯度的视觉解释技术，更具体地说，是从头实现了 Grad- ECLIP 的一种适配 [54]，用于我们的基于 CLIP 的多模态编码器（SEARLE），从而将f_c 的输出归因回图像像素。通过沿模型的图像塔反向传播s 的梯度，我们生成热力图叠加层，标示出对查询结果相似度至关重要的区域。显著性高亮使用户能够看到模型在匹配查询时“关注”了图像的哪些部分，将具体的视觉证据与语义查询反馈联系起来（见图 4）。

这些可解释的提示增强功能共同构成一个由用户掌控的交互式反馈循环。大语言模型生成的变体T'（及其排名变化指标）与归因可视化（词元条形图和显著性图）相结合，让用户能够以更高的透明度和信心引导检索。这种方法与新兴的交互式解释范式相契合，强调人在回路中的优化和更深入的模型透明度 [31]，，确保提示词调整既基于模型内部运作机制，又以用户意图为依据，最终通过降低有效实验的门槛来实现 DG3。

4.7 系统模块化与可扩展性

系统在设计时充分考虑了模块化。只要检索模型能够在 CLIP 嵌入空间中提供图像和文本嵌入，就可以轻松替换或添加。可视化模块彼此解耦且可配置，支持添加新的视图或编码策略。这种灵活性使用户能够尝试不同的检索算法、数据集和用户交互范式。

4.8 使用场景

为了展示 InfoCIR 如何支持组合图像检索中的交互式精炼，可以设想用户上传一张青苹果的图像，并输入提示词“红苹果”。目标是检索语义上

相关的图像，在保留物体身份（苹果）的同时修改其属性（颜色）。

执行查询后，SEARLE 检索器返回的结果集中既有青苹果，也有红苹果，还包括一些无关的水果。在 UMAP 投影中，结果分布散乱，没有形成明显的红苹果簇。用户打开归因面板，发现“红色”一词对排名的贡献较弱，这表明 CLIP 表征存在风格偏差。

为纠正这一问题，用户选择一张或多张符合其意图的检索图像（例如，任何明显呈深红色的苹果），并打开提示增强模块。该组件会提供语义引导的提示变体，例如“熟透的苹果”“深红色的苹果”和“深红色水果”，这些变体均以语料嵌入和视觉相似度为依据。

选择“深红色的苹果”后，更新后的检索结果在嵌入空间中聚成一个紧凑的红苹果簇。如 Rank-Δ 热力图所示，所选图像的排名从第 12 位跃升至第 2 位，而归因权重也显示“红色”及相关图像区域的激活更强。

该场景展示了 InfoCIR 如何通过提示适配和嵌入检查，帮助用户调试并引导检索过程，从而获得忠实于语义、符合组合意图的结果。

5 评估

我们的评估策略主要以提示增强工作流作为分析单元。InfoCIR 包含六个协同视图（A-F），而提示增强是“关键”任务，需要综合整个仪表板中的信息。要成功优化查询，用户必须先诊断初始失败（使用嵌入视图和显著性图），再提出假设（使用大语言模型建议），最后验证结果（使用 Rank-Δ 热力图）。因此，我们认为，以提示增强上的表现作为衡量指标，可以全面反映各诊断支持组件的效用。

我们开展了两部分评估，以考察 InfoCIR 如何通过交互式分析组件，帮助用户高效检索目标图像并理解模型行为。这项研究结合了定量性能任务与定性反馈分析，以考察客观效率和主观用户体验。我们的主要假设是，与缺少此类支持的基线系统相比，InfoCIR 的交互式解释功能，尤其是提示增强面板，能够提升检索性能、用户信心以及对图文关系的理解。

本研究遵循既定的研究伦理标准。所有参与者均获知研究相关信息，并提供了书面知情同意。参与完全自愿，未收集任何可识别个人身份的数据，所有回答均经过匿名化处理并安全存储。根据《通用数据保护条例》的规定，项目完成后将永久删除这些数据。

5.1 定量研究：性能与检索效率

5.1.1 流程

我们招募了阿姆斯特丹大学的八名人工智能硕士生，他们均具有计算机科学背景。参与者此前曾使用网页开发、设计或信息检索工具，并在过去的专业任务中有搜索和检索图像的经验。下文将参与者称为 P1 至 P8。

对于该数据集，我们使用 ImageNet-R 模拟难度和视觉抽象程度相近的真实检索场景。

任务 A 和任务 B 的参考图像取自公开可用的在线资源[1] [2]。我们设计了两个图像检索任务（任务 A 和任务 B），使其在语义和感知难度上具有可比性。在这两个任务中，参与者需要查看一张真实的参考图像，并尝试检索同一物体类别的风格化（卡通风格）版本。任务 A 涉及一只 **豪猪**，任务 B 则涉及一只**波士顿梗**（见图 5）。选择这两个类别，是因为它们的视觉复杂度相近，且都具有鲜明、易于识别的特征。



(a) 参考任务 A **(b) 目标任务 A**

(c) 参考图像任务 B **(d) 目标图像任务 B**

图 5：定量评估中使用的参考图像与目标图像对。任务 A（豪猪）和任务 B（波士顿梗）均要求检索现实参考图像的风格化版本。参与者反复优化文本提示，直到目标图像出现在检索结果的前 3 位中。

设计与流程 我们比较了参与者使用工具两个不同版本时的表现：一个基础版本和一个增强版本，以了解小组件如何影响图像检索任务。

1. **场景 1 - 基线：** 参与者仅使用组合图像检索面板和查询结果面板，隐藏 InfoCIR 的所有小组件。2. **场景 2 - InfoCIR：** 参与者使用完整的 InfoCIR 界面，包括提示增强面板、显著性图、直方图/词云、嵌入视图、词元归因和排序-Δ 热力图。

为避免顺序偏差，参与者被分为两组，分别以相反的顺序完成相同的任务，从而确保任务顺序不会影响表现。

- **A组：** 任务 A（基线） → 任务 B（InfoCIR）
- **B组：** 任务 B（基线） → 任务 A（InfoCIR）

每项任务要求参与者不断完善文本提示，直到目标出现在检索结果前三图像中，限时 5 分钟。每次查询会返回十张图像。实验过程计时，并由两名实验人员记录结果。评估指标包括：

1. 成功率：将目标置于检索结果前三的参与者比例。2. 任务平均用时：成功检索所需的平均时间。3. 唯一查询数：提示词提交次数的平均值。

5.1.2 定量结果

如表 1所示，使用完整 InfoCIR 界面的参与者，其检索表现显著优于基线。成功率从 37.50% 提高到 87.50%，

[1]任务 A 的参考图像，检索自：https://brandywinezoo.org/wp-content/uploads/2024/10/north-american-豪猪-brandywine-zoo-2023.jpg

[2]任务 B 的参考图像，来源：https://dogtime.com/wp-content/uploads/sites/12/2023/09/GettyImages-1500664833.jpg

平均完成时间从 277 秒降至 133 秒，唯一查询数的平均值从 7.4 降至 3.3。一位参与者（P6）在两种条件下均未能成功，因为最初提供的提示词过于宽泛，目标图像未出现在排名前十的检索图像中，因而没有有意义的反馈可用于指导优化。这也表明，有效使用 InfoCIR 仍需要有目的地调整提示词。

这些结果表明，InfoCIR 的小组件通过引导用户使用更有效的类别化语言，加快了收敛。各项指标证实了设计目标的实现：耗时和查询次数减少，表明系统支持快速试验（DG3）；而较高的成功率则表明用户成功诊断并纠正了检索失败（DG1）。

5.2 定性研究：用户体验与可解释性

5.2.1 流程

完成两项检索任务后，参与者填写了一份问卷，对 InfoCIR 的各个组件进行五点李克特量表评价。随后进行了一场简短讨论，参与者在其中描述了自己的工作流程、认为哪些部分易于理解，以及遇到了哪些困难。此次定性研究旨在了解用户在执行检索时如何体验该界面、最依赖哪些组件，以及系统如何支持他们进行推理。

5.2.2 定性研究结果

参与者普遍认为，增强后的界面既提升了检索效率，也增进了他们对系统如何处理提示词的理解。不过，几位参与者（P1、P3、P5、P8）指出，初期需要一段时间熟悉界面；在此期间，他们会探索各个组件，之后逐渐确定一组较少但始终依赖的功能。这表明存在学习曲线，尤其是对于旨在提升可解释性、而非直接帮助完成任务的组件。

提示增强面板始终被认为是提升检索效果最有用的功能。大多数参与者认为其建议清晰且相关，62.50%的参与者将其选为总体上最有帮助的小组件。参与者表示：“这让任务变得更容易了”（P4），以及“我不需要手动精心设计一个强有力的初始提示词，就能在前 3 个结果中检索到目标图像”（P7）。此外，参与者指出，使用该面板让他们学会了如何更有效地组织基于类别的提示词，而不是依赖目标图像的描述性属性（图见补充材料），这验证了该面板在支持迭代式提示增强方面的作用（DG3）。

直方图和词云通过展示模型将哪些概念与当前提示词相关联，帮助参与者识别更有效的词汇。参与者表示，这有助于他们从描述性措辞转向基于类别的术语。例如，P2 解释说，它帮助其将“小型黑白狗”改为“波士顿梗”，从而直接改善了检索效果；P5 则将其描述为发现模型偏好术语的有用指南（图示见补充材料）。通过揭示这些语义不匹配，这些小组件成功实现了支持快速诊断的目标（DG1）。然而，参与者也指出，尽管该小组件有助于理解，但与任务期间的快速提示词优化相比，它对整体方向把握的帮助更大。

嵌入视图和显著性图都对模型可解释性和事后推理发挥了实质作用，而非直接影响即时检索决策。参与者通常能理解嵌入视图，并主要用它观察系统如何对图像进行分组，以及如何组织相似度关系。一些参与者表示，它帮助他们理解了模型的内部结构，但并不能直接帮助他们快速优化提示词。P5指出，它“解释了结构，但我无法快速利用它”。与此同时，显著性图帮助用户确认哪些图像区域影响了检索结果，常被用作正确性检查（图见补充材料）。P5认为它有助于确认系统是否聚焦于主要对象，这与 DG1 的诊断需求相符。不过，参与者普遍提到它会带来时间成本，认为使用它所需的时间和注意力超出了任务允许的范围。P3表示，它“处理起来太费时间”，P4则提到，尽管它阐明了模型注意力，却没有直接帮助提示词调整。

词元归因和**排名Δ 热力图**在时间受限的检索任务中使用频率较低，因为许多参与者认为它们解读起来更慢，因此难以应用于快速的提示词调整。不过，对于采取更具探索性方法的参与者而言，这些小组件支持了更深入的分析。尤其是 P3 和 P8，他们在场景 B 中额外花时间检查这些可视化图，并表示这些图帮助他们理解细微的措辞变化如何影响排名结果（图示见补充材料）。这表明，当用户有足够时间进行探索时，这些小组件对于详细的可解释性分析和比较（DG2）具有重要价值。

6 讨论与未来工作

InfoCIR 的开发凸显了组合图像检索（CIR）领域对可解释、交互式系统日益增长的需求。尽管近期的组合图像检索模型展现出颇具前景的零样本和免训练检索能力 [1, 12], ，但其行为往往仍不透明，尤其是在多模态查询（图像 + 文本）产生意外结果时。本系统通过提供紧密集成的工作流程来应对这一挑战，让用户能够以可解释且直观的方式追踪检索动态。

一项关键的设计洞见是，将检索结果与嵌入空间的整体结构以及针对各个提示令牌的详细反馈结合起来，能够带来益处。通过 UMAP 投影嵌入可帮助用户探索邻域连贯性和聚类结构，而词元归因和显著性图则揭示了模型决策背后的内部推理。相互关联且可交互的可视化视图使用户能够发现并处理错位情况，例如某个提示词尽管在语义上并不相关，却显得占主导地位，从而更便于用户快速形成并检验新想法。

另一项重要贡献是提出了理想引导的提示词增强理念。用户可以将一个或多个目标图像设为基准，InfoCIR 由此将提示增强转化为可通过 Rank-Δ 热力图观察的约束优化任务。这种方法能够即时提供针对性的反馈，帮助评估提示词的有效性，弥补了模型调试工作流程中长期存在的不足：以往通常只能根据最终检索排名判断提示词质量，却无法了解其背后的因果动态 [6, 20]。

不过，仍存在一些局限：

- **嵌入可视化的局限性**将高维 CLIP 嵌入投影到二维空间不可避免地会造成度量失真，这意味着在原始空间中有效的线性向量运算，在投影空间中并不能严格成立。因此，嵌入视图旨在用于可视化簇验证，确认语义簇内的拓扑包含关系，而非作为用于轨迹分析的精确度量图。尽管我们的流程通过生成清晰的簇，便于快速目视检查，从而支持这项任务，但我们也承认流形近似固有的权衡。未来的工作将探索 PaCMAP 和 TriMap 等替代技术 [46, 3, 14]，以期在局部保真度与全局结构一致性之间取得更好的平衡。

表 1：参与者在两种检索场景中的表现。场景 A 为基线（手动提示词优化）。场景 B 为 InfoCIR 辅助条件（启用所有分析组件和提示增强面板）。成功表示目标图像是否出现在前 3 条检索结果中。在场景 B 中，每次提示词优化操作也计为一次查询。平均值基于全部 8 位参与者。

		场景 A（基线）			场景 B（InfoCIR 辅助）		
ID	**组**	**时间（秒）**	**查询次数**	**成功率（%）**	**时间（秒）**	**查询次数**	**成功率（%）**
P1	A	243	6	100.00	103	3	100.00
P2	A	300	9	0.00	87	2	100.00
P3	A	300	7	0.00	134	3	100.00
P4	A	281	5	100.00	91	3	100.00
P5	B	300	11	0.00	149	4	100.00
P6	B	300	8	0.00	300	6	0.00
P7	B	194	4	100.00	55	2	100.00
P8	B	300	9	0.00	142	3	100.00
平均值	–	**277**	**7.4**	**37.50**	**133**	**3.3**	**87.50**

制约因素。

• **冷启动脆弱性：**自动提示增强依赖用户从初始结果中至少选出一张相关的“理想”图像。这使其效果取决于初始查询的表现。如果零样本检索未能返回任何相关目标（即“零命中”情形），用户就必须通过手动迭代，将搜索转向相关的语义邻域，然后才能使用自动增强工具。

• **泛化与过拟合：** 将提示词优化锚定于用户选定的特定图像，可能会导致生成的提示词过拟合于所选锚点的视觉特征，而无法泛化到更广泛的语义类别。能够成功检索到特定锚点的提示词，可能会无意中排除目标类别的其他相关变体。

• **模型可操作性：** 尽管 InfoCIR 能有效诊断检索失败，但这些洞见主要有助于提示工程和查询加权，而非模型修正。在大型基础模型时代，开发者通常无力根据该工具揭示的细粒度实例级观察结果，重新训练庞大的骨干网络（例如 SEARLE 或 CLIP）。

• **大语言模型启发式方法与验证：** 提示词建议由开放权重大语言模型生成，仍属于启发式方法；它们无法保证改善检索效果，偶尔还可能产生属性幻觉。我们通过 Rank-Δ热力图缓解这一问题，它充当验证闭环，使用户能够通过实测检验并弃用无效变体。未来工作可以纳入结构化提示（例如场景图 [52]）或多轮优化 [51]，以提高生成稳定性 [9]。

• **归因保真度：**词元归因图依赖梯度近似，可能无法充分捕捉基于 Transformer 的架构中复杂的注意力机制 [16]。

• **局部与全局可解释性：**基于归因的解释只能提供关于特定查询图像对的局部洞见。扩展系统以支持反事实视觉生成（例如，对查询图像进行假设性编辑）或概念级抽象，或可实现全局可解释性，并在更高层面支持模型调试 [2]。

• **反馈闭环集成：** 当前的优化仅包含一次基于大语言模型的改写，且只由用户选定的理想图像提供引导。显著性图和词元归因条形图仅用作解释性叠加层，不会反馈至改写函数。未来的迭代可以让明确的用户标注（例如区域草图、自定义注意力掩码，或强化学习引导的评分）直接影响提示生成器，从而提供更精细的控制和更丰富的交互体验。

尽管存在这些限制，InfoCIR 仍展现出，使检索结果与以人为中心的推理相一致既切实可行，又具有重要影响。该系统推动组合图像检索工具超越黑盒指标，迈向交互式、富含解释的工作流程，从而更好地支持模型开发、评估与优化。

7 结论

我们介绍了 InfoCIR，这是一个旨在提升组合图像检索（CIR）的透明度、可解释性与优化能力的可视分析系统。通过在协调统一的交互式界面中整合检索排名、嵌入投影与可视化图、提示增强及多层级解释，我们的系统弥补了 CIR 可解释性方面的关键不足，并帮助开发者更有信心地探索、理解、改进和验证检索模型行为。

我们的贡献包括一个可解释的提示增强流程，它将大语言模型生成的提示变体与基于梯度的归因可视化相结合；一个用于检查嵌入动态的协同环境；以及一种模块化架构，只需少量工作即可支持新的组合图像检索模型。通过引导式界面交互，用户可以快速评估提示的影响、探索嵌入空间，并找出失败案例或性能瓶颈。

与将嵌入检查或提示工程彼此割裂的现有工具不同，InfoCIR 将两者融为一体，支持一种以检索为中心的多模态分析新范式。

未来的工作可以扩展模型支持、可扩展性和个性化，进一步弥合不透明的模型内部机制与以用户为中心的检索工作流程之间的差距。随着多模态人工智能领域不断发展，我们认为 InfoCIR 既是一个诊断工具包，也是未来系统的框架，旨在优先实现由人类引导的改进、透明度以及对检索流程的信任。

补充材料

已提交补充材料，其中包括：(1) 经 UMAP 调整的结果以及评估过程中的小组件附录；(2) InfoCIR 系统的演示；(3) 指向 GitHub 仓库的超链接，该仓库包含 InfoCIR 系统的源代码。

参考文献

[1] L. Agnolucci、A. Baldrati、M. Bertini 和 A. D. Bimbo。isearle：改进用于零样本组合图像检索的文本反演，2024 1, 3, 4, 8 。A. L. Alfeo、A. G. Zippo、V. Catrambone、M. G. Cimino[2]、N. Toschi 和 G. Valenza。从局部反事实到全局特征重要性：针对脑连接网络的高效、稳健且与模型无关的解释。《生物医学计算机方法与程序》，236:107550，2023。doi：10.1016/j.cmpb.2023.107550 9[3] E. Amid 和 M. K. Warmuth。TriMap：使用三元组进行大规模降维，2022 3, 8 。A. Baldrati、L. Agnolucci、M. Bertini[4] 和 A. D. Bimbo。通过文本反演实现零样本组合图像检索，2023 2 。B. Chandna 和 P. Sen[5]。面向检索模型的反事实解释框架。CoRR，abs/2409 00860, 2024 2, 3 。. B. Chandna 和 P. Sen[6]。面向检索模型的反事实解释框架，2025 8 。Z. Chen、F. Silvestri、J. Wang、Y. Zhang[7]、Z. Huang、H. Ahn 和 G. Tolomei。Grease：为基于图神经网络的推荐生成事实解释和反事实解释，2022 3 。Y. chin Ivan Chang。综述：潜在的降维方法，2025 5 。J. Choi[8]。通过基于大语言模型的混淆矩阵反馈，高效优化相关性评估提示。《应用科学》，15(9)，2025。doi：10 3390 15095198 9 。/app B. Dong、R. Collins[9] 和 A. Hoogs。基于内容的图像检索的可解释性。发表于 IEEE/CVF 计算机视觉与模式识别会议论文集，研讨会，页码95 98, 2019 3 – 。X. Dong、S. Wang、D. Lin、G. K. Rajbahadur、B. Zhou、S. Liu[10] 和 A. E. Hassan。PromptExp：大语言模型的多粒度提示解释。CoRR，abs/2410 13073, 2024 3, 6 。. N. Efthymiadis、B. Psomas、Z. Laskar、K. Karantzalos、Y. Avrithis、O. Chum[11] 和 G. Tolias。用于免训练域转换的组合图像检索，2024。2，3，8[12] Y. Feng、X. Wang、K. K. Wong、S. Wang、Y. Lu、M. Zhu、B. Wang[13] 和 W. Chen。PromptMagician：用于文本生成图像的交互式提示工程。《IEEE可视化与计算机图形学汇刊》，30(1):295–305，2024。doi：10.1109/TVCG.2023.3327168 2[14] J. Fu 和 D. Zhao。MPAD：一种在高维向量搜索中保留最近邻的新型降维方法，2025。8[15] T. Ghosh 和 M. Kirby。使用质心编码器进行有监督降维与可视化，2020 5 。S. M. V. Hatefi、M. Dreyer、R. Achtibat、T. Wiegand、W. Samek[16] 和 S. Lapuschkin。重新审视通过解释进行剪枝：优化归因方法以剪枝卷积神经网络和 Transformer，2024 9 。Z. Huang、D. Witschard[17]、K. Kucher 和 A. Kerren。Va + embeddings star：嵌入在可视分析中应用的最新进展报告。《计算机图形学论坛》，42:539–571，06 2023。doi：10. 1111 14859 3 /cgf。I. T. Jolliffe。《主成分分析》。施普林格统计丛书[18]。施普林格出版社，纽约，2002 10 1007 98835 3 。doi：。/b D. Kobak[19] 和 G. Linderman。初始化对于在 t-SNE 和 UMAP 中保留全局数据结构至关重要。《自然·生物技术》，39:1–2，02 2021。doi：10.1038/s41587-020-00809-z 5[20] Q. R. Lauro、S. Shankar、S. Zeighami 和 A. Parameswaran。无延迟的检索增强生成：检索增强生成流程的交互式调试，2025 8 。Q. Li、K. S. Njotopra wiro、H. Haleem[21]、Q. Chen、C. Yi 和 X. Ma。EmbeddingVis：用于比较网络嵌入检查的可视分析方法，2018 3 。Z. Liu、H. Zhang、C. Xiong、Z. Liu、Y. Gu[22] 和 X. Li。通过条件自编码器实现高效密集检索的降维，2022。3[23] F. Luus、N. Khan 和 I. Akhalwaya。基于 t-SNE 的交互式监督。发表于第十届知识捕获国际会议论文集，K-CAP '19 ，第 85–92 页。计算机协会机器，纽约州纽约，美国，2019 年。doi: 10.1145/3360901.3364414 3[24] L. McInnes、J. Healy 和 J. Melville。Umap：用于降维的统一流形近似与投影，2020 2, 3, 5 。R. Nara[25]、Y. C. Lin、Y. Nozawa、Y. Ng、G. Itoh、O. Torii 和 Y. Matsui。重新审视基于 CLIP 的交互式图像检索中的相关性反馈。CoRR，abs/2404 16398, 2024 2 。. W. A. Pike、J. Stasko、R. Chang[26] 和 T. A. O'Connell。交互科学。《信息可视化》，8(4):263–274，2009年12月。doi：10.1057/ivs.2009.22 4[27] B. A. Plummer、M. I. Vasileva、V. Petsiuk、K. Saenko 和 D. Forsyth。这些为何匹配？解释图像相似度模型的行为。arXiv，2019 3 。B. Psomas、I. Kakogeorgiou[28]、N. Efthymiadis、G. Tolias、O. Chum、Y. Avrithis 和 K. Karantzalos。遥感领域的组合图像检索，2024 2 。B. Poczos[29]。独立成分分析。https: //www.cs.cmu.edu/~bapoczos/Classes/ML10715_2015Fall/ slides/ICA.pdf，2015。讲义，CMU 10-715：机器学习高级导论。5[30] A. Radford、J. W. Kim、C. Hallacy、A. Ramesh、G. Goh、S. Agarwal、G. Sastry、A. Askell、P. Mishkin、J. Clark、G. Krueger 和 I. Sutskever。通过自然语言监督学习可迁移视觉模型，2021。2[31] J. Rorseth、P. Godfrey、L. Golab、M. Kargar、D. Srivastava 和 J. Szlichta。CREDENCE：文档排名的反事实解释。载于第39届国际数据工程会议论文集，2023 2, 6 。T. Sainburg、L. McInnes 和 T. Gentner。用于表征和半监督学习的参数化 UMAP 嵌入。《神经计算》，33:2881–2907，2021年10月。doi：10.1162/neco a 01434 5[32][33] M. Sedlmair、M. Meyer 和 T. Munzner。设计研究方法论：来自实践与文献的反思。《IEEE可视化与计算机图形学汇刊》，18(12):2431–2440，2012。doi：10.1109/TVCG.2012.213 2[34] R. R. Selvaraju、M. Cogswell、A. Das、R. Vedantam、D. Parikh 和 D. Batra。Grad-CAM：通过基于梯度的定位实现深度网络的可视解释。《国际计算机视觉杂志》，128(2):336–359，2019年10月。doi：10.1007/s11263-019-01228-7 2, 3[35] C. Shi、W. Cui、S. Liu、P. Xu、W. Chen 和 H. Qu。RankExplorer：大型时间序列数据中排序变化的可视化。《IEEE可视化与计算机图形学汇刊》，18 2678，2012 12 2669 10 1109 2012 ( ): – 。doi：. /TVCG 253 2, 3[36] D. Smilkov、N. Thorat、C. Nicholson、E. Reif、F. B. Vi´egas 和 M. Wattenberg。Embedding Projector：嵌入的交互式可视化与解读，2016 2, 3 。X. Song、H. Lin、H. Wen、B. Hou[37]、M. Xu 和 L. Nie。组合图像检索综合综述，2025 2 。Y. Tang、J. Yu、K. Gai[38]、J. Zhuang、G. Xiong、G. Gou 和 Q. Wu。利用世界模型预测缺失的目标相关信息，以实现准确的零样本组合图像检索，2025 2 。J. B. Tenenbaum、V. De Silva[39] 和 J. C. Langford。非线性降维的全局几何框架。《科学》，290 5500 2319 2323, 2000 10 1126 290 5500 2319 3 ( ): – 。doi：. /science. . . P. Tovstogan、X. Serra[40] 和 D. Bogdanov。用于音乐探索与再发现的深度音频嵌入可视化。载于第19届声音与音乐计算会议（SMC）论文集，第493–500页。法国圣艾蒂安，2022年6月。可从 UPF Digital Repository 获取。3[41] L. van der Maaten 和 G. Hinton。使用 t-SNE 可视化数据。《机器学习研究杂志》，9 86 2579 2605, 2008 3 ( ): – 。B. Vasu[42]、B. Hu、B. Dong、R. Collins 和 A. Hoogs。可解释的交互式基于内容的图像检索。《应用人工智能快报》，2 41，2021 4 10 1002 2 41 3 ( ):e 。doi：. /ail .[43] N. Vo、L. Jiang、C. Sun、K. Murphy、L.-J. Li、L. Fei-Fei 和 J. Hays。组合文本与图像用于图像检索——一次实证探索之旅，2018。2[44] H. Wang、X. Jiang、A. Nagarajan、X. Guo、L. Ding、D. Wan、J. Zhao 和 Y. Chen。Colorslope：一种平衡概览与

排名随时间变化细节的平衡可视化。《视觉智能》，1(7):1–18，2023。doi：10 1007 44267 023 00008 9 2, 3 。/s - - - J. Wang、Z. Deng、D. [45] Deng、X. Wang、R. Sheng、Y. Cai 和 H. Qu。借助可视化赋能多模态分析：综述。《计算机科学评论》，57:100748，2025。doi：10.1016/j.cosrev.2025.100748 5[46] Y. Wang、H. Huang、C. Rudin 和 Y. Shaposhnik。理解降维工具的工作原理：一种解读 t-SNE、UMAP、TriMap 和 PaCMAP 数据可视化方法的实证方法，2021。3, 8[47] Z. J. Wang、F. Hohman 和 D. H. Chau。WizMap：用于探索大型机器学习嵌入的可扩展交互式可视化，2023。3[48] M. Worring、J. Zah´alka、S. van den Elzen、M. T. Fischer 和 D. A. Keim。基础模型时代的多媒体分析模型，2025。4[49] Z. Xu、H. Lamba、Q. Ai、J. Tetreault 和 A. Jaimes。用于搜索结果解释的反事实编辑。CoRR，abs/2301 10389, 2023 2 。. Z. [50] Yang、D. Xue、S. Qian、W. Dong 和 C. Xu。LDRE：用于零样本组合图像检索的基于大语言模型的发散推理与集成。载于第47届国际计算机协会信息检索研究与开发会议论文集，SIGIR '24，第80–90页。计算机协会，美国纽约州纽约，2024 10 1145 3626772 3657740 2 。doi：. / . S. [51] Ye、Z. Sun、G. Wang、L. Guo、Q. Liang、Z. Li 和 Y. Liu。提示词炼金术：自动优化提示词以增强代码生成，2025。9[52] H. Yin、X. Xu、Z. Wu、J. Zhou 和 J. Lu。SG-Nav：面向基于大语言模型的零样本物体导航的在线三维场景图提示，2024 9 。J. [53] Zahalka´ 和 M. Worring。迈向交互式、智能化与集成化的多媒体分析。载于2014年IEEE可视分析科学与技术会议（VAST）论文集，第3–12页，2014。doi：10. 1109/VAST.2014.7042476 3[54] C. Zhao、K. Wang、J. H. Hsiao 和 A. B. Chan。Grad ECLIP：基于梯度的 CLIP 视觉与文本解释。CoRR，abs/2502 18816，2025。. [55] 3, 6 K. Zhu、Q. Zhao、H. Chen、J. Wang 和 X. Xie。PromptBench：用于评估大语言模型的统一库，2024。3