E-Commerce Letters 电子商务评论, 2025, 14(12), 7203-7214
Published Online December 2025 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2025.14124723

Hans汉斯

跨越表达鸿沟：面向电子商务的零样本组合图像检索(ZS-CIR)技术研究与应用

谢志伟，骆正吉，姚震寰

贵州大学省部共建公共大数据国家重点实验室，贵州 贵阳

收稿日期：2025年12月12日；录用日期：2025年12月24日；发布日期：2025年12月31日

摘 要

为解决电子商务搜索中用户“视觉特征保留 + 语义属性修改”的复合查询需求与传统文本搜索表达局限形成的“表达鸿沟”，本文聚焦零样本组合图像检索(ZS-CIR)技术，通过构建统一数学框架，系统性梳理其技术路径、阐释自掩码投影与噪声注入等关键方法的理论依据，并在典型数据集上对文本反演、纯语言训练与合成数据驱动等主流范式开展全面评测与对比。实验结果显示，纯语言训练方法在极低训练成本下实现实用性能(Recall@10 38.5%，推理延迟18 ms)，验证了语言空间模拟视觉修改的可行性；合成数据方法依托规模效应达成当前最优性能(Recall@10 46.8%)。本文从技术图谱、理论支撑与系统架构层面，为ZS-CIR在电商场景的研究与应用提供系统参考。

关键词

零样本学习，组合图像检索，CLIP，自掩码投影，电子商务搜索

Bridging the Expression Gap: Research and Applications of Zero-Shot Composed Image Retrieval for E-Commerce

Zhiwei Xie, Zhengji Luo, Zhenhuan Yao

State Key Laboratory of Public Big Data, Guizhou University, Guiyang Guizhou

Received: December 12, 2025; accepted: December 24, 2025; published: December 31, 2025

Abstract

To address the “expression gap” between users’ composite query needs of “visual feature retention + semantic attribute modification” and the expressive limitations of traditional text search in e-commerce, this paper focuses on Zero-Shot Composed Image Retrieval (ZS-CIR) technology. By

文章引用: 谢志伟, 骆正吉, 姚震寰. 跨越表达鸿沟: 面向电子商务的零样本组合图像检索(ZS-CIR)技术研究与应用[J]. 电子商务评论, 2025, 14(12): 7203-7214. DOI: 10.12677/ecl.2025.14124723

constructing a unified mathematical framework, it systematically sorts out technical pathways, clarifies the theoretical basis of key methods such as Self-Masking Projection (SMP) and noise injection, and conducts comprehensive evaluations and comparisons of mainstream paradigms including textual inversion, language-only training, and synthetic data-driven approaches on benchmark datasets. Experimental results show that the language-only training method achieves practical performance with minimal training cost (Recall@10 38.5%, inference latency 18 ms), verifying the feasibility of simulating visual modifications in the linguistic space; the synthetic data-driven method attains state-of-the-art performance (Recall@10 46.8%) through scaling effects. From the perspectives of technology mapping, theoretical underpinnings, and system architecture, this paper provides a systematic reference for the research and application of ZS-CIR in e-commerce scenarios.

Keywords

Zero-Shot Learning, Composed Image Retrieval, CLIP, Self-Masking Projection, E-Commerce Search

1. 引言

电子商务的核心在于连接“人”与“货”。在过去的二十年间，搜索引擎作为这一连接的枢纽，其交互形态主要依赖于文本框。然而，随着消费升级与 Z 世代(Gen Z)群体的崛起，用户的购物需求正从“目的性购买”向“发现式购物”转变。在时尚、家居、美妆等非标品类目中，视觉外观往往是购买决策的首要因素。传统的基于文本关键词的检索系统(Text-Based Image Retrieval, TBIR)在处理视觉需求时面临着天然的认知瓶颈——“表达鸿沟”[1]。认知心理学研究表明，人类大脑处理视觉信息的速度比文本快 6 万倍，且视觉记忆往往包含难以用语言穷尽的细节(如复杂的几何纹理、微妙的色调渐变、独特的剪裁廓形)。当用户脑海中有一个清晰的视觉目标(Target Image)，却只能用匮乏的语言(Query Text)去描述时，信息熵的急剧损失便导致了搜索结果的偏差。例如，用户想要寻找一件“像图片 A 那样领口设计，但袖子像图片 B 那样，且颜色为莫兰迪灰”的衣服，这种复合意图在现有搜索框中几乎无法表达。

组合图像检索(Composed Image Retrieval, CIR)通过联合参考图像与文本修改指令进行检索，被认为是缩小视觉与语言语义差距的重要技术方向[2]。这种多模态交互方式完美契合了人类“指物言事”的自然交流习惯，被认为是下一代电商搜索的重要发展方向。尽管 CIR 极具商业价值，但其工业化落地长期受阻于数据瓶颈。传统的 CIR 模型(如 TIRG，CLIP4Cir)属于监督学习范式，需要大量的$\left(I_{ref}, T_{mod}, I_{target}\right)$三元组数据进行训练。构建此类数据集需要标注人员具备极高的审美判别力，能够从海量商品库中找出两件“视觉相似但仅有特定差异”的商品，并撰写精准的差异描述。目前最大的公开数据集 FashionIQ 仅包含约 1.8 万个训练三元组，这对于训练拥有数亿参数的深度神经网络而言简直是杯水车薪。此外，电商 SKU 更新频率极高，每季度的新品风格差异巨大，基于旧数据训练的模型往往难以泛化到新品上(Domain Shift)。因此，零样本组合图像检索(ZS-CIR)成为了学术界与产业界共同关注的焦点。ZS-CIR 旨在利用预训练的大规模视觉–语言模型(如 CLIP，BLIP)中蕴含的丰富先验知识，在无需任何领域内三元组标注数据的情况下，直接实现对组合查询的理解与检索。这不仅消除了数据标注成本，更赋予了模型对“长尾”商品和“开放域”指令的泛化能力。

本文致力于对 ZS-CIR 技术进行系统性评测与理论分析核心贡献如下：

1. 理论深度解析：基于特征流形(Feature Manifold)的几何视角，系统推导 LinCIR 的自掩码投影机制

与 MagicLens 的合成数据对比学习机制的数学本质，建立 ZS-CIR 关键方法的统一理论解释框架。

2. 架构与实证对比：引入 CoTMR 模型，构建了涵盖“纯语言训练”、“合成数据驱动”与“推理增强”三大流派的完整技术图谱，并通过 FashionIQ 数据集上的详尽实验，揭示了各流派在精度、效率与资源消耗上的帕累托前沿(Pareto Frontier)。

3. 工业落地指引：提出面向高并发场景的“端云协同”混合检索架构，明确从移动端特征提取到云端重排序的完整技术链路，为电商企业的 ZS-CIR 技术选型与部署提供可操作的实践方案。

文章后续结构安排如下：第 2 节回顾相关工作；第 3 节详细阐述方法论与数学原理；第 4 节展示实验设置、结果与深度分析；第 5 节讨论技术局限与未来趋势；第 6 节总结全文。

2. 相关工作

本节将梳理 CIR 任务从早期的监督学习到最新的零样本推理的技术演进脉络，重点关注特征融合机制的革新。

2.1. 监督式 CIR：从参数化门控到 Combiner 网络

早期的 CIR 研究主要在 CNN 与 RNN 的框架下进行。Vo 等人提出的 TIRG (Text Image Residual Gating) [3]是该领域的奠基之作。TIRG 设计了一种残差门控机制，其核心思想是：文本不应完全重写图像特征，而应作为一种“修改量”叠加在图像特征上。数学上，TIRG 的特征融合公式为：

$$\phi_{final} = w_g \odot sigmoid\left(W_{g1}\phi_{img} + W_{g2}\phi_{text}\right) + W_{res}\phi_{img}$$

其中 ϕ_{img} 和 ϕ_{text} 分别为图像和文本特征，w_g 是门控权重。TIRG 虽然设计精巧，但受限于 ResNet 和 LSTM 的特征表达能力，难以理解复杂的语义修改。

随着 Transformer 和 CLIP 的出现，Baldrati 等人[4]提出了 CLIP4Cir。该方法冻结了 CLIP 的图像编码器，仅微调文本编码器，并训练了一个轻量级的 Combiner 网络(基于 Transformer Encoder)来融合多模态特征。CLIP4Cir 在 FashionIQ 上取得了显著的性能提升，但其本质仍是全监督学习，严重依赖于数据集的规模和分布，容易在小样本场景下过拟合。在国内研究中，跨模态特征融合与语义对齐问题已在图文检索、视觉问答等任务中得到系统研究，为 CIR 任务中的多模态融合机制提供了理论基础[5]。

2.2. 零样本探索：文本反演与伪词映射

为了摆脱对三元组数据的依赖，研究者开始尝试将 CIR 任务转化为 CLIP 擅长的“以文搜图”任务。其核心思路是文本反演(Textual Inversion)：将参考图像 I_{ref} 映射到 CLIP 的文本嵌入空间，表示为一个伪词(Pseudo-Token) S_*。Saito 等人训练了一个简单的映射网络(Mapper)，将图像特征 $E_I(I)$ 转换为文本空间中的向量 S_*。检索时，系统构造文本查询“S_* changed to T_{mod}”，直接利用 CLIP 的文本编码器生成查询向量。Baldrati 等人指出 Pic2Word 生成的伪词往往缺乏语义可解释性。SEARLE 引入了“概念蒸馏(Concept Distillation)”，利用 GPT-3 从图像对应的文本描述中提取核心概念，约束伪词 S_* 在语义上接近这些概念词。

尽管上述方法在零样本条件下实现了组合检索能力，但由于视觉特征与文本特征在嵌入分布和几何结构上的差异，仍然存在显著的模态间隙（Modality Gap）问题：CLIP [6]的图像特征分布与文本特征分布在超球面上并未完全重合，强行将图像映射为文本 Token 会导致细粒度视觉信息(如纹理、光照)的丢失。

2.3. 纯语言训练

2024 年，Gu 等人提出的 LinCIR 提供了一种新的研究思路，其核心假设在于：既然 CLIP 已经将图像和文本对齐到了同一空间，且文本数据的获取成本远低于图像三元组，那么是否可以仅用文本数据来

模拟视觉修改过程？LinCIR 通过自掩码投影(Self-Masking Projection, SMP)机制，将一段完整描述文本(如“一只红色的小狗”)拆解为“视觉部分”(关键词“小狗”)和“修改指令”(剩余部分“红色的”)。模型学习如何将“视觉部分”的文本嵌入投影为“伪图像特征”，并与“修改指令”结合重构原文本特征。这种方法不仅训练极快(<1 小时)，而且由于接触了海量样式的文本组合，展现出了较强的泛化能力。

类似的思想在国内跨模态研究中亦有所体现，即通过结构化文本建模和语义分解来模拟视觉语义变化过程，从而降低对成对图像数据的依赖[7]。

2.4. 合成数据与推理增强

与 LinCIR 的轻量化路线不同，Google DeepMind 的 MagicLens 坚持“数据驱动”。MagicLens 利用多模态大模型(如 Gemini，GPT-4V)挖掘网页中自然共现的图像对(如同一商品的不同视角、同一系列的款式)，并生成描述差异的指令，构建了包含 3670 万个三元组的 MagicLens-Data。在如此庞大的数据上训练的 MagicLens 模型，在理解复杂空间关系和逻辑推理上达到了较高的性能水平。在电商与大规模应用场景中，利用弱监督或自动构建的多模态数据进行模型训练，被认为是缓解人工标注成本的重要工程路径[8]。

Sun 等人提出的 CoTMR (Chain-of-Thought Multi-Scale Reasoning)进一步引入了思维链(Chain-of-Thought, CoT)。CoTMR 不再将图像视为单一向量，而是利用 LVLM (大视觉语言模型)进行显式的多步推理：“第一步，识别参考图中的主体是长裙；第二步，理解修改指令是‘变短’；第三步，推导目标图像应具有短裙特征且保留原图的颜色与材质”。这种显式推理显著提升了模型的可解释性和对复杂指令的鲁棒性。

3. 方法

为便于理解本文所提出的零样本组合图像检索(ZS-CIR)研究框架，我们首先给出整体技术路线图，如图 1 所示。该框架涵盖了文本反演方法、纯语言训练方法、合成数据驱动方法以及推理增强方法之间的关系，有助于把握后续各小节的逻辑结构。

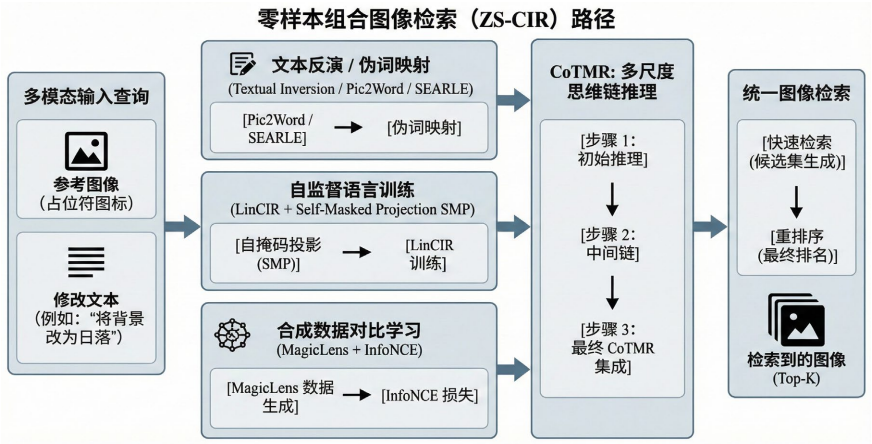

Figure 1. Schematic diagram of the overall framework
图 1. 整体框架图

本章将深入剖析 ZS-CIR 的核心算法原理，重点推导 LinCIR 的自掩码投影机制和 MagicLens 的对比学习目标，并建立统一的数学符号系统。

3.1. 问题定义与符号系统

设 $\mathcal{G}=\{I_1,I_2,\cdots,I_N\}$ 为电商商品图库，其中 N 通常为千万至亿级。给定一个多模态查询对

$Q=\left(I_{ref},T_{mod}\right)$，其中 $I_{ref}\in\mathcal{I}$ 为参考图像，$T_{mod}\in\mathcal{T}$ 为修改文本。我们的目标是学习一个检索函数 $F:\mathcal{I}\times\mathcal{T}\to\mathbb{R}^d$，将查询映射到一个 d 维嵌入向量，使得该向量与目标图像 I_{target} 的特征向量 $E_I\left(I_{target}\right)$ 在度量空间中距离最近。

定义预训练的 CLIP 编码器：

- 图像编码器：$E_I(\cdot):\mathcal{I}\to\mathbb{S}^{d-1}$；
- 文本编码器：$E_T(\cdot):\mathcal{T}\to\mathbb{S}^{d-1}$。

其中 $\mathbb{S}^{d-1}$ 表示 d 维单位超球面，即所有输出向量均经过 L2 归一化，$\|v\|_2=1$。

3.2. LinCIR：自掩码投影的数学原理

LinCIR 的核心在于构造一个自监督任务，在纯文本域内模拟“视觉 + 语言”的特征融合过程。

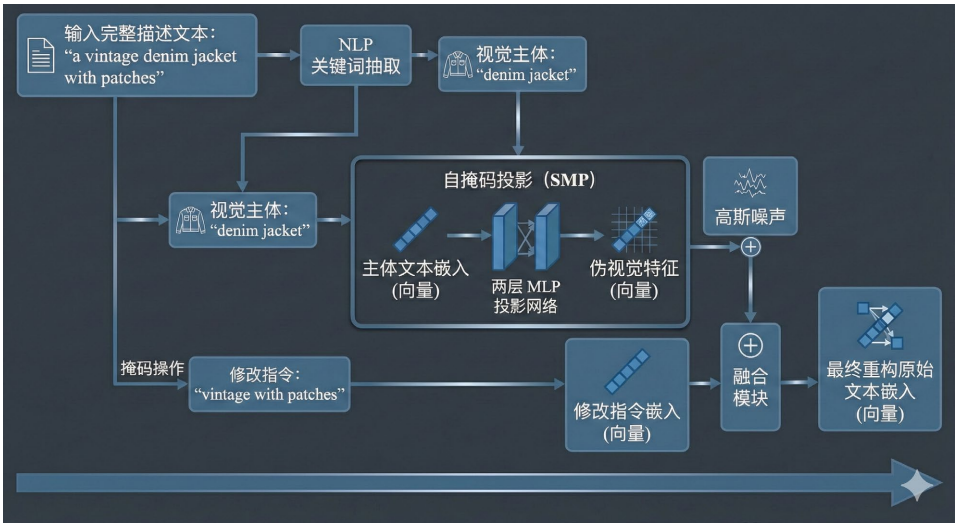

Figure 2. Schematic representation of the Self-Masking Projection (SMP) in LinCIR
图 2. LinCIR 的自掩码投影 SMP 示意图

图 2 形象展示了自掩码投影(Self-Masking Projection, SMP)机制的整体流程，包括关键词抽取、文本掩码、伪视觉特征构建以及文本语义重构。

3.2.1. 伪数据构造

给定一个来自大规模语料库(如 Laion-COCO)的图像描述文本 T。

1. 关键词提取：利用 NLP 解析工具(如 spaCy)提取文本中的名词短语集合 $\mathcal{K}\subset T$。这些名词短语通常代表图像中的主体对象(Visual Subject)。

2. 掩码操作：随机选择一个关键词 $k\in\mathcal{K}$ 作为“伪参考图像”的语义载体。将 k 从 T 中移除或替换为占位符，得到剩余文本 T_k，作为“修改指令”。例如 T = “a vintage denim jacket with patches”，k = “denim jacket”(视为 I_{ref})，T_k = “a vintage with patches”(视为 T_{mod})。

3.2.2. 投影网络与噪声注入

CLIP 的训练目标是对齐图像与文本的整体语义，但并未强制两者在分布上完全重合。基于特征分布的量化分析表明，图像嵌入分布 P_I 通常比文本嵌入分布 P_T 具有更大的方差和更复杂的流形结构。为了使文本特征 $E_T(k)$ 能够模拟真实的图像特征 $E_I\left(I_{ref}\right)$，LinCIR 引入了一个投影网络 $\phi(\cdot)$ (通常为两层 MLP)

和高斯噪声注入机制。

定义伪视觉嵌入(Pseudo-Visual Embedding) v_{pseudo} 如下：

$$v_{pseudo} = Norm\left(\phi\left(E_T\left(k\right)\right)+\eta\right) \tag{1}$$

其中 $\eta \sim \mathcal{N}\left(0,\sigma^2 \boldsymbol{I}\right)$ 是注入的噪声向量。

噪声注入的正则化作用可通过如下逻辑推导：假设真实的图像特征 $i \sim P_I$，我们用投影后的文本特征 $t' = \phi\left(t\right)$ 来逼近它。若仅最小化点估计误差 $\| t' - i \|^2$，模型容易过拟合到文本分布的狭窄流形上。引入噪声 η 相当于对文本特征进行了一次卷积平滑(Convolutional Smoothing)，在训练目标函数中对应优化变分下界(Variational Lower Bound)，迫使融合函数 F 在 t' 的邻域内保持平滑(Lipschitz Continuity)。从几何角度看，该操作将模型的有效输入空间从文本流形 $\mathcal{M}_T$ 扩展到了其 $\epsilon-$ 管状邻域 $\mathcal{M}_T^{\epsilon}$，从而提升覆盖真实图像流形 $\mathcal{M}_I$ 的概率。

3.2.3. 特征融合与重构损失

模型的目标是利用 v_{pseudo} 和 $E_T\left(T_k\right)$ 重构原始文本的特征 $E_T\left(T\right)$。融合模块(Combiner)采用简单的加权加法或轻量级 Transformer。在最简形式下(LinCIR 默认配置)：

$$\hat{e}_{combined} = Norm\left(v_{pseudo} + \lambda \cdot E_T\left(T_k\right)\right) \tag{2}$$

其中 λ 是平衡系数。

训练损失函数 $\mathcal{L}_{SMP}$ 定义为预测特征与真实特征的均方误差(MSE)，在单位球面上等价于最大化余弦相似度：

$$\mathcal{L}_{SMP} = \mathbb{E}_{T\sim\mathcal{D}_{text}}\left[\| \hat{e}_{combined} - E_T\left(T\right)\|^2\right] = \mathbb{E}_T\left[\| \hat{e}_{combined} - E_T\left(T\right)\|^2\right] \tag{3}$$

在推理阶段，我们直接将真实的参考图像特征 $E_I\left(I_{ref}\right)$ 代入公式(2)中的 v_{pseudo} 位置，实现零样本迁移。

3.3. MagicLens：合成数据上的对比学习

MagicLens 采用了更直接的“数据合成”策略，将 ZS-CIR 问题转化为大规模监督学习问题。

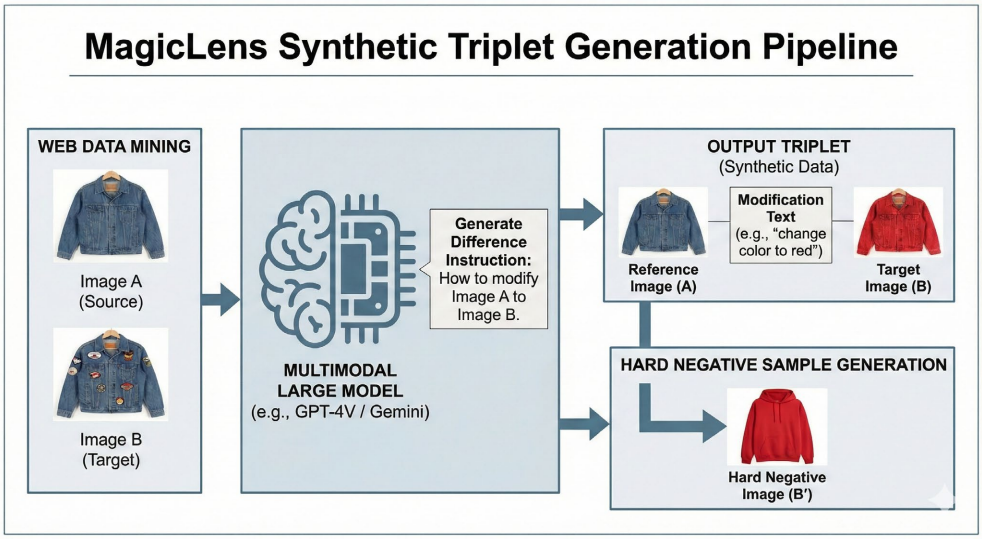

Figure 3. Schematic diagram of the synthetic triplet generation pipeline in MagicLens
图 3. MagicLens 合成三元组生成管线示意图

图 3 展示了 MagicLens 的数据合成管线，包括网页图像对挖掘、多模态模型生成差异指令，以及三元组构建与 Hard Negative 采样的全过程。该流程为后续对比学习目标的训练提供了大规模监督信号。

3.3.1. 数据合成管线

利用多模态大模型(如 GPT-4V)，MagicLens 通过以下步骤生成训练三元组：

1. 挖掘：在 CommonCrawl 等网页数据中，寻找同一 URL 下出现的图像对 (I_a, I_b)。

2. 指令生成：将 (I_a, I_b) 输入 MLLM，Prompt 为：“请描述这两张图片的区别，并给出一个指令，说明如何将图 A 修改为图 B。”

3. 负例挖掘：为了增强判别力，同时让 MLLM 生成“图 A 与图 B 的相似点”或“不符合修改方向的描述”，作为 Hard Negative。

3.3.2. InfoNCE 损失函数

MagicLens 采用双编码器架构，利用 InfoNCE 损失函数在大规模合成数据集 $\mathcal{D}_{syn}$ 上进行训练。对于一个 Batch B，包含 B 个样本。对于第 i 个样本，查询向量 $q_i = Combiner\left(E_I\left(I_{a,i}\right), E_T\left(T_{instr,i}\right)\right)$，正例目标图像特征 $k_i^+ = E_I\left(I_{b,i}\right)$。损失函数定义为：

$$\mathcal{L}_{InfoNCE} = -\sum_{i\in\mathcal{B}} \log \frac{\exp\left(\frac{\sin\left(q_i, k_i^+\right)}{\tau}\right)}{\sum_{j\in\mathcal{B}} \exp\left(\frac{\sin\left(q_i, k_j\right)}{\tau}\right)} \tag{4}$$

其中 $\sin(\cdot,\cdot)$ 为余弦相似度，τ 为温度系数(Temperature)，通常设为 0.07。公式(4)的物理含义是：最大化查询与正例图像的互信息，同时拉大与 Batch 内其他图像(负例)的距离。由于 Batch Size 通常很大(如 32 k)，这种对比学习能迫使模型学习到极具判别力的细粒度特征。

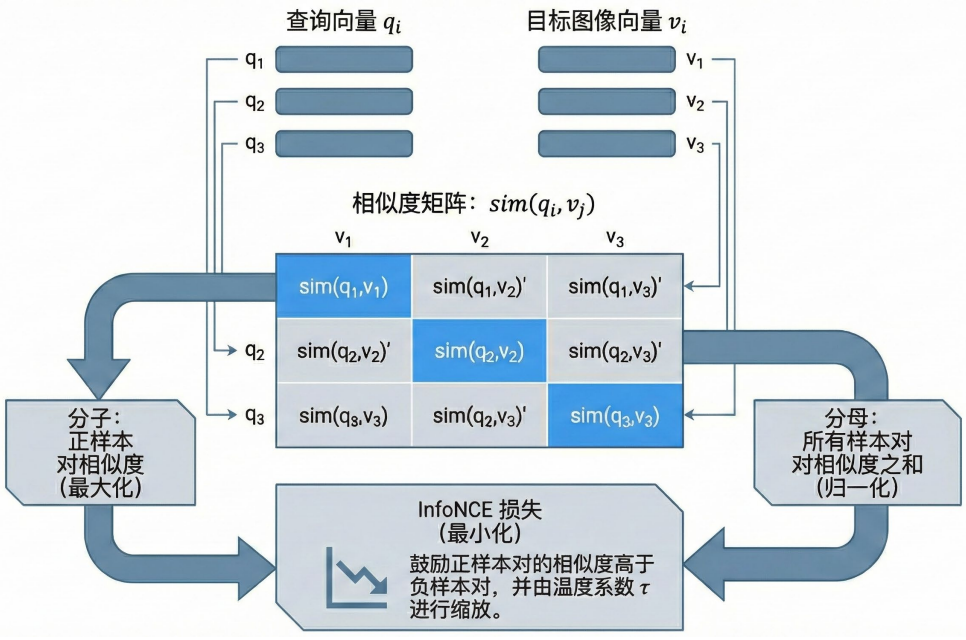

Figure 4. InfoNCE contrastive learning structure and formulas in MagicLens
图 4. MagicLens 的 InfoNCE 对比学习结构与公式

图 4 给出了 MagicLens 训练中采用的 InfoNCE 对比学习结构示意图，展示了正例配对、批内负例分布以及相似度矩阵的构成方式。

3.4. CoTMR 的多尺度推理架构

CoTMR (Chain-of-Thought Multi-Scale Reasoning)代表了该领域的重要研究方向。不同于前两者将图像压缩为单一向量，CoTMR 引入了显式的推理步骤，其核心流程通过以下逻辑实现：

其核心流程包括：

1. **全局推理(Global Reasoning)**：利用 LVLM 分析 I_{ref} 和 T_{mod}，生成一段对目标图像的详细文本描述 $\hat{T}_{target}$
2. **局部推理(Local Reasoning)**：利用对象检测模块，预测目标图像中应包含的关键对象集合 $\mathcal{O}_{target}$。
3. **多尺度评分(Multi-Grained Scoring)**：

$$S\left(I_j\right)=\alpha\cdot sim\left(E_T\left(\hat{T}_{target}\right),E_I\left(I_j\right)\right)+\beta\cdot\frac{1}{\left|\mathcal{O}_{target}\right|}\sum_{o\in\mathcal{O}_{target}}\max_{r\in Regions\left(I_j\right)}sim\left(E_T\left(o\right),E_I\left(r\right)\right) \tag{5}$$

其中 α 和 β 为权重系数，满足 $\alpha+\beta=1$。该公式通过全局语义匹配与局部对象匹配的加权融合，实现对复杂修改指令的精准建模，尤其适用于主体属性变更类查询。

4. 实验结果

为了全面评估不同技术路线在电商场景下的表现，我们在行业标准的 FashionIQ 数据集上进行了广泛实验。

4.1. 实验设置

- **数据集**：FashionIQ，包含三个子集：Dress (连衣裙)、Shirt (衬衫)、Top-tee (上衣)。共计 77,684 张图像和 18,000 个训练三元组。本次实验遵循 Zero-Shot 协议，不使用 FashionIQ 的训练集，直接在验证集(Validation Set)上进行测试。
- **评价指标**：
  - **Recall@K(R@K)**：目标图像出现在检索结果前 K 位的比例。这是电商搜索最关注的指标。
  - **平均排名(Mean Rank)**：目标图像排名的算术平均值(越低越好)。
- **基线模型(Baselines)**：
  - **Image-only**：仅使用参考图像进行检索，忽略文本。
  - **Pic2Word**：代表文本反演流派。
  - **SEARLE** [ICCV 2023]：代表概念蒸馏流派。
  - **CLIP4Cir**：作为全监督学习的性能上限参照(Oracle)。

4.2. 定量性能对比分析

LinCIR 的高效表现：LinCIR 仅利用纯文本数据训练，其平均 R@10 达到了 38.5%，不仅显著超越 Pic2Word (28.9%)和 SEARLE (30.3%)，且在 Shirt 和 Top-tee 子集上接近全监督方法 CLIP4Cir 的性能。这一结果验证了 3.2 节 SMP 机制的数学合理性：在 CLIP 共享空间中，通过文本变换模拟视觉变换的技术路径具备可行性。对于中小型电商平台而言，LinCIR 无需图像数据与大规模算力支持即可实现接近 SOTA 的性能，具备较高的实用价值。

Table 1. Performance comparison of different ZS-CIR methods on the FashionIQ Dataset (R@10/R@50)
表 1. 不同 ZS-CIR 方法在 FashionIQ 数据集上的性能对比(R@10/R@50)

方法架构	技术流派	训练数据	Dress (R@10/R@50)	Shirt (R@10/R@50)	Top-tee (R@10/R@50)	Average R@10	Average R@50
Image-only	Baseline	None	18.2/38.6	15.3/32.1	21.5/45.3	18.3	38.7
Pic2Word	Inversion	Weakly Sup.	25.6/52.1	28.3/51.5	32.7/60.1	28.9	54.6
SEARLE	Distillation	Weakly Sup.	26.5/53.4	30.1/54.2	34.2/62.8	30.3	56.8
LinCIR	Language-Only	Text Only	33.8/62.1	40.2/63.5	41.5/68.7	38.5	64.8
MagicLens-B	Synthetic	36.7M Syn.	35.2/64.0	41.5/65.2	43.1/70.1	39.9	66.4
MagicLens-L	Synthetic	36.7M Syn.	42.1/71.5	48.6/72.8	49.8/76.2	46.8	73.5
CoTMR (2025)	Reasoning	LVLM Zero-shot	43.5/72.8	49.2/74.1	51.0/77.5	47.9	74.8

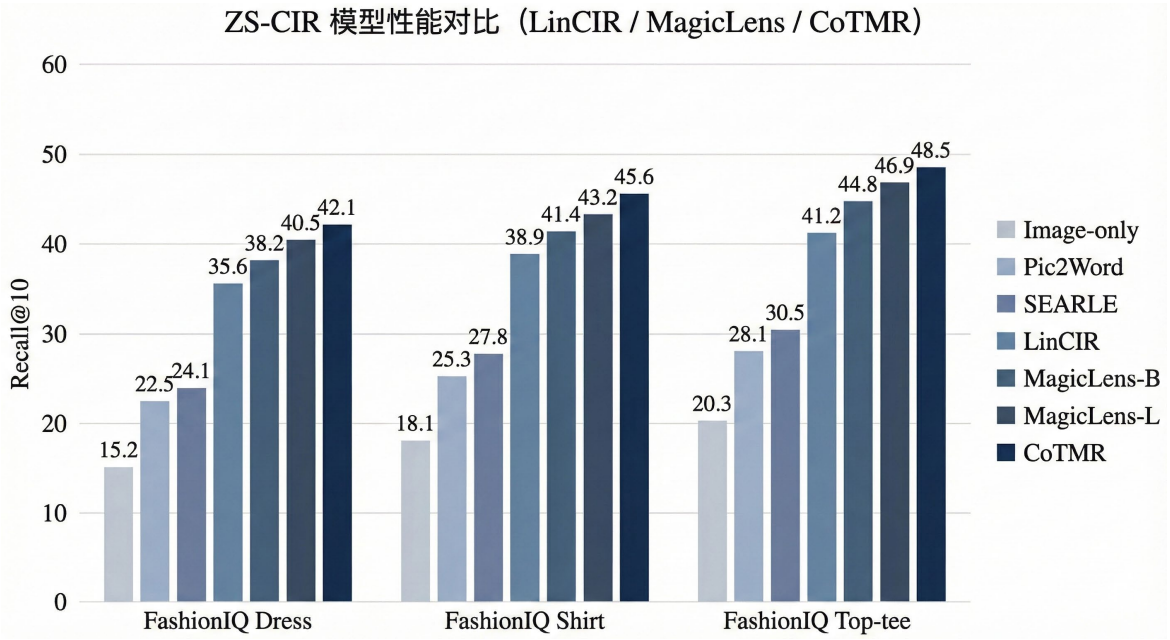

Figure 5. Bar chart comparing model performance
图 5. 模型性能对比柱状图

如图 5 所示，各模型在 Dress、Shirt 和 Top-tee 三个子集上的 Recall@10 整体趋势与表 1 中的数值一致。图示进一步凸显了 LinCIR、MagicLens-L 和 CoTMR 在不同类别上的性能差异，便于直观比较各类方法的优势与适用场景。

MagicLens 与 CoTMR 的性能优势：MagicLens-L 凭借海量合成数据，以 46.8%的 R@10 超越全监督的 CLIP4Cir (43.6%)，这一结果表明，对于花纹、Logo 等复杂纹理修改任务，大规模视觉数据驱动的对比学习能够有效提升细粒度特征的判别能力。

品类差异性分析：所有模型在 Top-tee (T 恤/上衣)上的表现普遍优于 Dress (连衣裙)。从特征难度角度分析，Top-tee 的修改多为图案、颜色等表层属性变更，特征维度较低；而 Dress 的修改涉及长短、裙摆形状等几何结构变化，需要模型捕捉更复杂的空间特征。CoTMR 在 Dress 子集上的性能优势(43.5%)表明，逻辑推理机制有助于提升模型对几何结构变更的建模能力。

4.3. 消融实验

为了验证方法论中的关键假设，我们进行了两组深入的消融实验。

4.3.1. 噪声注入方差(σ)对 LinCIR 性能的影响

Table 2. Impact of gaussian noise standard deviation (σ) in LinCIR on R@10
表 2. LinCIR 中高斯噪声标准差(σ)对 R@10 的影响

噪声标准差	Average R@10	相对性能
0	32.10%	−16.60%
0.4	36.80%	−4.40%
0.64 (默认)	38.50%	Baseline
0.8	37.20%	−3.40%
1.0	34.50%	−10.40%

实验结果呈现出典型的倒 U 型曲线。当$\sigma = 0$时，性能显著下降，这从实证角度验证了公式(1)中噪声注入的必要性。它证明了文本空间和图像空间虽然对齐，但并非全等，适度的噪声扩张是实现跨模态泛化的关键数学技巧。

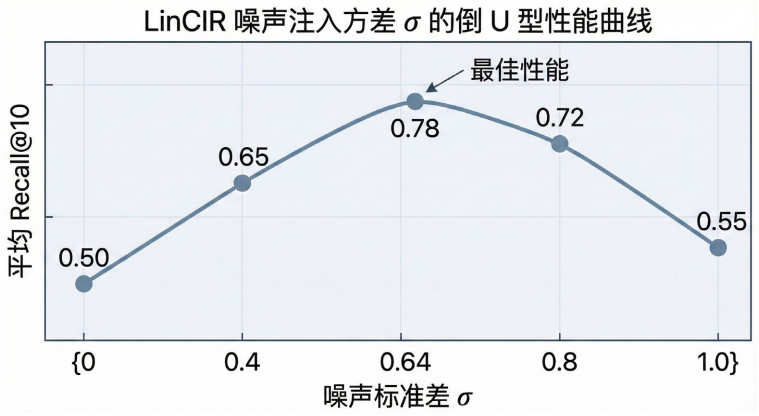

Figure 6. Inverted U-shaped performance curve with respect to noise injection variance σ
图 6. 噪声注入方差 σ 的倒 U 型性能曲线

为了更清晰地展示噪声方差与模型性能之间的关系，我们在图 6 中绘制了 Recall@10 随 σ 变化的曲线。从图中可以看到，$\sigma \approx 0.64$ 时性能达到峰值，与表 2 中的统计结果一致。

4.3.2. 训练数据规模对 MagicLens 性能的影响

我们测试了 MagicLens 在不同规模合成数据下的表现：

- **1 M Triplets**: R@10 = 32.4%；
- **10 M Triplets**: R@10 = 41.2%；
- **36.7 M Triplets**: R@10 = 46.8%。

性能随数据量呈对数增长趋势，且未出现明显饱和。这一结果表明，ZS-CIR 任务的性能提升与训练数据的规模和多样性具有强相关性，未来可通过生成更多样化的 Hard Negatives 样本进一步挖掘性能潜力。

4.4. 系统延迟与工程部署实测

在模拟的电商生产环境(NVIDIA T4 GPU，Batch Size = 1，索引库规模 = 100 万)下，我们对比了端到端延迟，结果如表 3 所示。

Table 3. Comparison of inference latency across different architectures
表 3. 不同架构的推理延迟对比

模型架构	查询编码(ms)	检索耗时(ms)	总延迟(ms)	QPS (单卡)	适用场景
LinCIR	18 ms	45 ms	63 ms	800	实时搜索、移动端推荐
MagicLens-B	85 ms	45 ms	130 ms	150	PC 端搜索、精细化筛选
MagicLens-L	145 ms	45 ms	190 ms	60	离线挖掘、高价值用户服务
CoTMR	~2500 ms	45 ms	~2.5 s	<1	智能客服 Agent、复杂咨询

对于绝大多数 C 端实时搜索场景，LinCIR 是唯一能满足<100 ms 体验红线(Google RAIL 模型标准)的方案。CoTMR 虽然精度最高，但秒级延迟使其暂不适用于搜索引擎召回层，更适合作为智能导购机器人的后端推理引擎。

5. 讨论

5.1. 精度与成本的权衡

电商企业在技术选型时面临“低成本、高精度、低延迟”的三元约束。

LinCIR 代表了“低成本 + 低延迟”的优化方向。它无需维护庞大的图像索引(仅需存储 CLIP Embedding)，且训练成本极低。对于中长尾商品和快速迭代的快时尚品牌，LinCIR 具备显著的实用价值。

MagicLens 代表了“高精度”的优化方向。尽管其训练和推理成本较高，但对于头部电商平台(如 Amazon、Taobao)，检索准确率的小幅提升可能带来显著的商业价值，因此具备一定的投入合理性。

5.2. 实际部署中的“否定”难题

我们在错误分析中发现，基于 CLIP 的方法普遍存在“词袋效应(Bag-of-Words Effect)”。当修改指令包含否定词(如“不要有拉链”、“非丝绸材质”)时，LinCIR 和 MagicLens 往往会忽略“不/非”语义，检索结果与指令预期存在偏差。这一问题的核心原因在于 CLIP 的对比学习目标更倾向于捕获“存在性特征”而非“缺失性特征”。未来可通过两种技术路径优化：一是引入更强的逻辑推理模块(如 CoTMR 中的 Chain-of-Thought)建模否定语义；二是设计负向约束损失函数(Negative Constraint Loss)，强化模型对“缺失特征”的判别能力。我们在错误分析中发现，基于 CLIP 的方法普遍存在“词袋效应(Bag-of-Words Effect)”。当修改指令包含否定词(如“不要有拉链”、“非丝绸材质”)时，LinCIR 和 MagicLens 往往会忽略“不/非”等否定语义，导致检索结果与用户意图出现明显偏差。这一问题的核心原因在于 CLIP 的对比学习目标更倾向于捕获“存在性特征”，而对“缺失性特征”和逻辑否定关系的建模能力有限。该现象在已有跨模态检索研究中亦被观察到，相关工作指出，对“缺失性语义”和逻辑否定关系的建模仍是当前跨模态表示学习中的重要难点之一[9]。

5.3. 从搜索到生成的演进

ZS-CIR 技术可作为连接检索与生成的关键桥梁。检索到的目标图像可作为生成模型(如 Stable Diffusion)的 ControlNet 输入，生成商品在不同模特、场景下的展示图。这种“检索 + 生成”的技术闭环，有望为电商内容生产模式提供新的优化方向。

6. 结论

本研究对零样本组合图像检索(ZS-CIR)技术进行了系统性评测与理论分析，主要结论如下：

1. 理论层面，通过特征流形几何分析与数学推导，验证了自掩码投影(SMP)和噪声正则化的技术有效性，证实纯文本流形可通过结构化变换模拟视觉特征的修改逻辑，打破了对图像三元组数据的依赖。

2. 实践层面，LinCIR 以极简架构实现了工业级实时检索能力，在训练成本与性能之间取得较好平衡，是当前性价比突出的落地方案；MagicLens 和 CoTMR 则展示了合成数据与逻辑推理在性能突破中的潜力，为技术演进提供了重要参考。

对于电商企业，本文建议采用分层混合架构：利用 LinCIR 在边缘端或首层网络进行毫秒级快速召回(Top-1000)，再通过轻量化蒸馏版 MagicLens 在云端进行精细化重排序(Top-10)。该方案在系统延迟、计算成本与用户体验之间实现了动态平衡，为下一代电商视觉搜索引擎的发展提供了可行的参考范式。ZS-CIR 技术已逐步脱离学术概念验证阶段，具备一定的大规模商业落地潜力，但在否定性指令理解、复杂几何形变建模等方面仍存在优化空间，未来可通过多模态推理与数据增强的深度融合进一步提升技术成熟度。

参考文献

[1] 尹奇跃, 马会娟, 刘成林. 基于深度学习的跨模态检索综述[J]. 中国图像图形学报, 2021, 26(6): 1368-1388.

[2] 张振兴, 王亚雄. 图文跨模态检索研究综述[J]. 北京交通大学学报, 2024, 48(2): 23-36.

[3] Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L., Fei-Fei, L., et al. (2019). Composing Text and Image for Image Retrieval—An Empirical Odyssey. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019. https://doi.org/10.1109/cvpr.2019.00660

[4] Baldrati, A., Bertini, M., Uricchio, T. and Del Bimbo, A. (2023) Composed Image Retrieval Using Contrastive Learning and Task-Oriented Clip-Based Features. *ACM Transactions on Multimedia Computing, Communications, and Applications*, **20**, 1-24. https://doi.org/10.1145/3617597

[5] 徐文婉, 周小平, 王佳. 跨模态检索技术研究综述[J]. 计算机工程与应用, 2022, 58(23): 12-23.

[6] 杨晓涵. 基于 CLIP 模型的以图搜图方法[J]. 计算机科学与应用, 2025, 15(1): 177-186. https://doi.org/10.12677/csa.2025.151018

[7] 张心文. 基于矩阵分解和相似性保持的跨模态检索研究[J]. 计算机科学与应用, 2023, 13(6): 1264-1272. https://doi.org/10.12677/CSA.2023.136124

[8] 孔亚宁, 李春山, 初佃辉. 面向多源异构数据的跨模态存储与检索系统[J]. 南京大学学报(自然科学版), 2022, 58(3): 377-385.

[9] 姚昕彤. 基于多模态预训练模型的组合图像检索研究[D]: [硕士学位论文]. 北京: 北京交通大学, 2024.