arXiv:2601.04571v1 [cs.AI] 8 Jan 2026

通过互补信息提取与对齐增强多模态检索

Delong Zeng[1], * **YuexiangXie**[2] **YaliangLi**[2] **Ying Shen**[1,3,†]
[1]中山大学智能系统工程学院
[2]阿里巴巴集团[3]广东省火灾科学与智能应急技术重点实验室
zengdlong@mail2.sysu.edu.cn, {yuexiang.xyx, yaliang.li}@alibaba-inc.com
sheny76@mail.sysu.edu.cn

摘要

近年来，多模态检索已成为一个前景广阔但极具挑战性的研究方向。现有的多模态检索研究大多侧重于捕捉多模态数据中与其配对文本相似的信息，却往往忽略了多模态数据所包含的互补信息。本研究提出了一种新颖的多模态检索方法 CIEA，该方法采用**互**补信息**提**取**与**对齐**方**法，将文档中的文本和图像映射到统一潜在空间，并采用互补信息提取器来识别和保留图像表征中的差异。我们通过两种互补的对比损失函数优化 CIEA，以确保语义完整性并有效捕捉图像中包含的互补信息。大量实验表明，CIEA 效果显著，性能优于分而治之模型和通用稠密检索模型。我们还提供了消融研究、进一步讨论和案例分析，以展示 CIEA 所取得的进展。为推动社区进一步研究，我们已在 https://github.com/zengdlong/CIEA发布源代码。



图 1：多模态检索示例。图像包含可辅助回答查询的关键信息，而这些信息可能不会出现在文本中。

1 引言

检索增强生成（RAG）（Lewis 等，2020b；Gao 等人，2023年；Yi 等人，2024年）近期因其能够通过提供最新信息并缓解幻觉问题，增强大语言模型（LLMs）（Tou-vron 等人，2023年；Bai 等人，2023年；Chung 等人，2024年；Brown 等人，2020年），而受到广泛关注。现有研究大多聚焦于检索文本信息（Cheng 等人，2023年；Shi 等人，2024年；Wang 等人，2024年；Zhu 等人，2025年），利用文本语料库为模型的回答提供事实依据。随着多模态数据的规模不断增长，如何有效支持多模态检索，从而提供文本之外的知识，已成为一个前景广阔但颇具挑战性的研究领域（Zhao 等人，2023年；Kuang 等人，2025）。

为推动多模态检索的发展，研究人员采用图像描述模型将多模态数据转换为文本（Wu 等人，2023年；Baldrati 等人，2023年；Mahmoud 等人，2024年），从而能够应用现有的、针对文本语料库开发的检索技术。然而，这些方法可能会受到图像描述模型效果的显著影响，导致转换过程中丢失关键信息（Che 等人，2023年；Li 等人，2024年）。近期研究提出了基于表示的方法：先对文本和多模态信息进行独立编码，再将其映射到统一嵌入空间中以进行知识检索（Radford 等人，2021年；Zhang 等人，2021年）；或者采用基于项目的联合编码过程（Zhou 等人，2024年；Li 等人，2024年）。

*在阿里巴巴集团实习期间完成的工作。
†通讯作者。

本研究聚焦于一种新颖的基于表示的方法，旨在将多模态数据迁移到语言模型的潜在向量空间中。我们指出现有基于表示的方法存在一个关键局限（Wang 等人，2024年；Zhou 等人，2024年）：它们往往着重捕捉多模态数据中与其文本对应内容（如图注或相关文本）相似的信息，却忽视了多模态数据中的互补信息。例如，如图1所示，与某张图像配对的文本描述可能提及彩虹大桥。虽然图像确实能够提供有关彩虹大桥的相似视觉细节，但夜景和桥梁颜色等补充信息也可能十分重要。对于基于表示的方法，提取并保留多模态数据中的这类互补信息，能够提升对仅凭文本信息无法充分解答的查询的响应质量。

受到上述洞见的启发与推动，我们提出了一种新颖的多模态检索方法，涉及**互**补**信**息提取**与**对齐，简称**CIEA**，即**CIEA**。具体而言，我们采用语言模型将文本转换到潜在空间，同时使用 CLIP 模型（Rad- ford 等人，2021年；Li 等人，2023）和投影器将图像信息映射到同一统一潜在空间。随后，我们设计了一个互补信息提取器，用于识别文档中图像与文本表示之间的差异。基于这些差异，我们更新图像的表示，以融入互补信息。此外，我们针对 CIEA 提出了一种新颖的优化方法，构建两个互补的对比损失函数：一个确保所学习表示的语义完整性，另一个则增强从图像中提取互补信息的能力。

我们进行了大量实验，以证明 CIEA 在多模态检索中的有效性。实验结果表明，与分而治之模型和通用稠密检索模型相比，所提出的方法取得了显著提升。我们还进行了消融研究，以突出 CIEA 各个组件的贡献。此外，我们进一步讨论了语言模型的影响，并提供了一些案例研究，以便更好地理解 CIEA。

2 相关工作

传统检索模型（Lewis 等人，2020年；Yu 等人，2023年）通过语言模型（Chen 等人，2024年）将查询和文档编码为向量，使用对比学习进行训练，并通过 K 近邻检索（Su 等人，2022）。与纯文本检索相比，多模态检索融入了多种信息模态，因此需要有效利用不同模态的信息（Yuan 等人，2021年）。一种方法是使用图像说明模型，将其他模态的信息转换为文本，从而将多模态检索任务有效地转化为纯文本检索任务（Baldrati 等人，2023年）。然而，这种方法可能会导致信息丢失（Che 等人，2023年；Li 等人，2024年）。

另一种方法是采用表示模型，在融合之前由视觉编码器和文本编码器分别对信息进行编码（Rad- ford 等人，2021年）。早期研究采用分而治之的策略，分别编码每种模态，再拼接向量以融合信息，这可能导致模态竞争。为解决这一问题，UniVL-DR（Liu 等人，2023b）提出了一种通用多模态检索框架，将查询和文档编码到统一嵌入空间中，以用于检索、路由和融合。一些现有研究（Wang 等人，2022年，b）通过训练大型模型来统一文本和视觉内容。然而，表示差异使得难以获取足够的数据来实现有效的语义理解（Lu 等人，2023）。

近期研究（Zhou 等人，2024年；Wang 等人，2024年；Li 等人，2024年，2023年）提出了一种基于投影的框架，该框架利用 CLIP（Radford 等人，2021年）等模型将视觉输入转换为特征序列，并引入投影层，使这些序列与语言模型嵌入对齐。这样的框架有助于在嵌入层面对视觉与文本内容进行比较，并将视觉信息转化为语言模型“词元”，从而利用语言模型训练过程中注入的知识。例如，MARVEL（Zhou 等人，2024年）利用基于投影的框架，在语言模型的输出空间中捕获多模态信息；MCL（Li 等人，2024年）则添加了一个检索词元，以提升模型在表示学习中的性能。

尽管已取得显著进展，视觉信息仍是语言模型中一个新颖且尚未得到充分

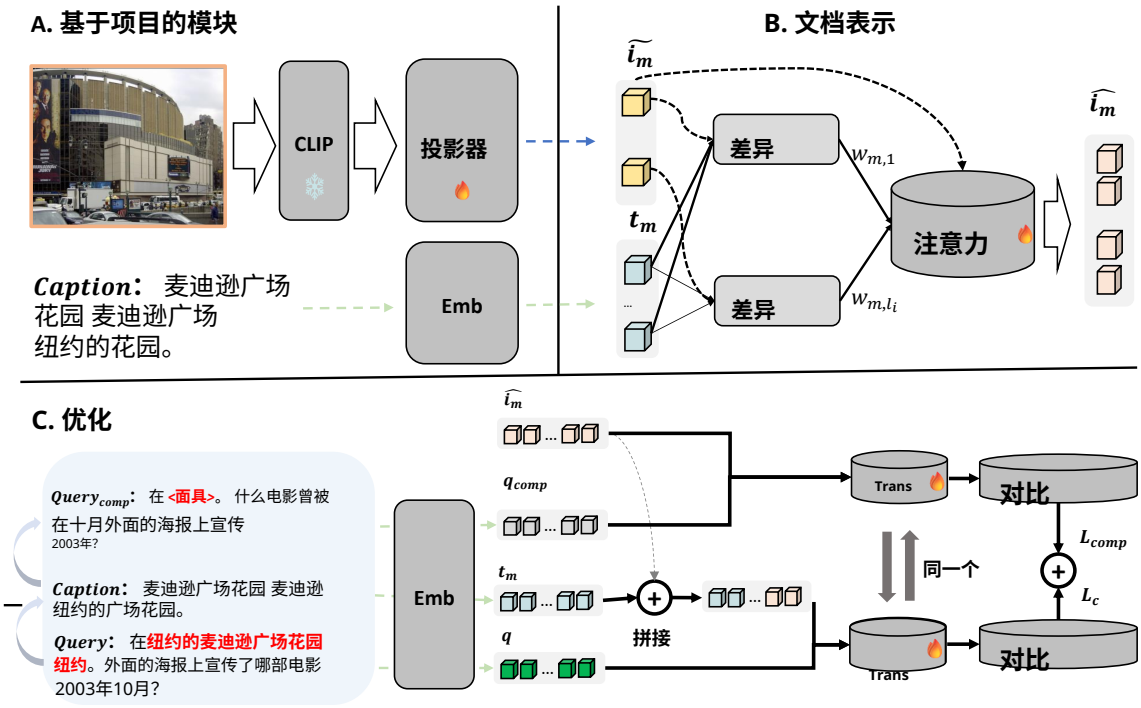


图2：CIEA的整体架构。上半部分展示了查询和多模态文档如何转换到统一潜在空间，下半部分则详细介绍了CIEA的优化过程。

探索的组成部分，这促使我们为处理视觉信息提供有效方案，从而获得全面的表示，以提升多模态检索效果。

3 方法

3.1 预备知识

本研究聚焦于多模态检索，其目标是根据给定的文本查询 q，从知识库中查找一个或多个相关文档。知识库包含总计 N 个文档，可表示为 $D = \{d_1, d_2, \ldots, d_N\}$。每个文档可能包含文本和图像，即 $d_m = \{t_m, i_m\}$，其中 t_m 和 i_m 分别表示文档中的文本和图像。

多模态检索的主要目标是将查询的表示与相应的多模态文档对齐，确保相似项目在统一潜在空间中彼此紧密匹配。为此，采用多模态编码器对查询和文档进行编码，将它们转换为稠密表示，表示形式如下：

$$\begin{aligned} \mathbf{q} &= \text{Encoder}(q), \\ \mathbf{d}_m &= \text{Encoder}(d_m), \forall d_m \in D. \end{aligned} \tag{1}$$

随后，通过余弦相似度测量查询与文档之间的相似度：

$$\cos(\mathbf{q}, \mathbf{d}_m) = \frac{\mathbf{q} \cdot \mathbf{d}_m}{\|\mathbf{q}\| \|\mathbf{d}_m\|}. \tag{2}$$

在以下各小节中，我们将详细介绍所提出的**互补信息提取与对齐方法**，记为**CIEA**。CIEA的整体架构如图2所示。具体而言，我们首先通过语言模型和CLIP模型对查询和多模态文档进行编码，如第3.2节所述。然后，在第3.3节中，我们设计了一个互补信息提取器，用于捕获图像中包含的互补信息。第3.4节介绍了针对所提出的CIEA设计的优化方法。

3.2 查询和多模态文档的表示

首先，如图2A 部分所示，我们将查询和多模态文档都转换为稠密表示。受以往研究的启发（Zhou 等人，2024年；Li 等人，2024年，2023年），我们采用基于 Transformer 的语言模型作为编码器的骨干网络，对文本和图像进行编码。基于 Transformer 的

语言模型通常由一个嵌入层和若干 Transformer 块组成。我们将嵌入层记为 Emb(·)，将 Transformer 块记为 Trans(·)。

文本查询 q 可以通过语言模型转换为稠密表示，表示如下：

$$\mathbf{q} = \text{Trans}(\text{Emb}(q)), \tag{3}$$

其中$\mathbf{q} \in \mathbb{R}^{l_q \times d}$、$l_q$ 表示查询中的词元数量，而 d 表示稠密表示的维度。

对于同时包含文本和图像的文档，即 $d_m = \{t_m, i_m\}$，文本 t_m 和图像 i_k 会在融合前分别进行处理。具体而言，我们首先将文档中的文本输入嵌入层：

$$\mathbf{t}_m = \text{Emb}(t_m), \tag{4}$$

其中 $\mathbf{t}_m = [\mathbf{t}_m^{(1)}、\mathbf{t}_m^{(2)}, \ldots\ldots, \mathbf{t}^{(l_t)}{}_m] \in \mathbb{R}^{l_t \times d}$，$l_t$ 表示文本中的词元数量。

文档中的图像通常可以表示为 RGB 多通道矩阵。我们采用冻结的 CLIP（Radford 等人，2021年）视觉编码器，将这些矩阵转换为一组语义向量。此过程涉及将一幅图像划分为多个图像块，每个图像块表示图像中的不同区域，形式化表示如下：

$$\mathbf{i}_m = \text{CLIP}_{\text{Visual}}(i_m), \tag{5}$$

其中 $\mathbf{i}_m \in \mathbb{R}^{l_i \times d_{\text{clip}}}$、$l_i$ 表示图像块的数量，d_{clip} 表示 CLIP 视觉编码器的隐藏维度。

请注意，这些图像表征的隐藏维度可能与文本表征的隐藏维度不同，即 $d_{\text{clip}} \neq d$。因此，我们采用线性层作为投影器，以便进一步对齐：

$$\widetilde{\mathbf{i}_m} = \text{Proj}(\mathbf{i}_m), \tag{6}$$

where $\widetilde{\mathbf{i}_m} = [\widetilde{\mathbf{i}_m}^{(1)}, \widetilde{\mathbf{i}_m}^{(2)}, \ldots, \widetilde{\mathbf{i}_m}^{(l_i)}] \in \mathbb{R}^{l_i \times d}$.

3.3 互补信息提取器

所提出的 CIEA 旨在捕捉图像中包含的互补信息，即文档中文本未涵盖的信息。为此，我们设计了一种互补信息提取器，用于捕捉图像与文本内容之间的差异。具体而言，我们计算文本与图像表示之间的图像块级距离，并取最大值作为差异度量，这与 BLIP-2 中采用的方法相似（Li 等人，2023年）。形式化地，图像的第j个图像块与文档m中的文本之间的差异，记为$r_{m,\ j}$，可计算如下：

$$r_{m,j} = -\max_{c \in [l_t]} \cos(\widetilde{\mathbf{i}_m}^{(j)}, \mathbf{t}_m^{(c)}). \tag{7}$$

所得差异被归一化到 $[0, 1]$ 范围内，以便有效地对图像块进行加权，如下所示：

$$w_{m,j} = \frac{1 + r_{m,j}}{2}. \tag{8}$$

我们根据计算得到的 $\mathbf{w}_m = w_{m、\ j}\ \forall j \in [l_i]$，采用注意力层对图像的图像块进行重新加权。嵌入的注意力分数与相应的权重相乘以进行归一化，具体如下：

$$\mathbf{Q}_m = \widetilde{\mathbf{i}_m}\mathbf{W}_Q,\ \mathbf{K}_m = \widetilde{\mathbf{i}_m}\mathbf{W}_K,\ \mathbf{V}_m = \widetilde{\mathbf{i}_m}\mathbf{W}_V,$$

$$\widehat{\mathbf{i}_m} = \text{softmax}(\frac{\mathbf{Q}_m\mathbf{K}_m \cdot \mathbf{w}_m}{\sqrt{d}})\mathbf{V}_m, \tag{9}$$

其中，$\mathbf{W}_Q$、$\mathbf{W}_K$、$\mathbf{W}_V$为可学习矩阵，此处“·”表示广播方法。

随后，如图2B 部分所示，图像和文本的嵌入被拼接在一起，并输入 Transformer 块以进行多模态知识融合。所获得的文档稠密表示 d_m 可表示为：

$$\mathbf{d}_m = \text{Trans}([\mathbf{e}_{\text{start}} \oplus \widehat{\mathbf{i}_m} \oplus \mathbf{e}_{\text{end}} \oplus \mathbf{t}_m]), \tag{10}$$

其中，$\oplus$ 表示拼接操作，$\mathbf{e}_{\text{start}}$ 和$\mathbf{e}_{\text{end}}$ 分别表示特殊标记 *<start>* 和 *<end>*的嵌入，用于明确区分文本嵌入和图像嵌入。

最后，对于给定的查询，我们根据文档表示（参见公式（10））与查询表示（参见公式（3））之间的相似度对文档进行排序，以获得相关文档。

3.4 优化

CIEA框架中的可训练参数包括投影器（定义见公式（6））、新增的注意力权重 $\mathbf{W}_Q$、$\mathbf{W}_K$、$\mathbf{W}_V$（用于重新加权，定义见公式（9）），以及语言模型的 Transformer 层（定义见公式（10））。本节将介绍如何优化 CIEA 中这些可训练参数。

表 1：WebQA-Multi 数据集上的比较结果。最佳结果以粗体标出。

	MRR@10 归一化折损累积增益@10		MRR@20 归一化折损累积增益@20		REC@20	REC@100
BM25	22.11	22.92	22.80	25.41	46.27	62.82
CLIP-DPR	37.35	37.56	37.93	40.77	69.38	85.53
BM25 与 CLIP-DPR	42.27	41.58	42.79	44.69	73.34	87.50
BM25 与 CLIP（oracle）	61.05	58.18	61.37	60.45	80.82	90.83
VinVL-DPR	38.14	35.43	38.74	37.79	53.89	69.42
CLIP-DPR	48.83	46.32	49.34	49.11	69.84	86.43
UniVL-DR	62.40	59.32	62.96	61.22	80.37	89.42
T5-ANCE	64.13	62.03	64.40	64.02	83.81	92.07
MARVEL	65.43	63.07	65.67	65.00	84.32	92.27
CIEA（我们的方法）	**66.16**	**63.89**	**66.41**	**65.85**	**85.43**	**92.75**

首先，我们从训练集的查询中找出与真实文档中文本相对应的文本片段。例如，如图2所示，我们将图注中出现的“麦迪逊广场花园 *in* 纽约”屏蔽。为进行识别，我们匹配语言模型分词器为查询和文档中文本生成的词元 ID。

之后，我们将这些识别出的文本片段替换为特殊标记 *<mask>*，同时保留查询的其余部分。我们将处理后的查询记为q_{comp}，其表示可以通过与式（3）中定义的计算类似的方式获得：

$$\mathbf{q}_{\text{comp}} = \text{Trans}(\text{Emb}(q_{\text{comp}})). \tag{11}$$

我们还为查询构造负样本，以便在优化时应用对比损失。标注为与查询相关的文档记为 D^+，而从知识库中采样的 N 负样本文档记为 $D^- = \{d_1^-, d_2^-, \ldots, d_N^-\}$。对比损失可定义为：

$$\mathcal{L}_c = -\log \frac{e^{\cos(\mathbf{q},\mathbf{d}^+)/\tau}}{e^{\cos(\mathbf{q},\mathbf{d}^+)/\tau} + \sum_{D^-} e^{\cos(\mathbf{q},\mathbf{d}^-)/\tau}}, \tag{12}$$

其中，τ 表示温度，$\cos(\cdot)$表示余弦相似度函数，$\mathbf{q}$ 表示查询的表示，$\mathbf{d}^+$ 和$\mathbf{d}^-$ 分别表示正样本文档和负样本文档的表示。

同时，为了增强从文档中的图像提取互补信息的能力，我们构建了另一种对比损失函数，该函数基于$\mathbf{q}^{\text{comp}}$ 和图像的表示，可表示为：

$$\mathcal{L}_{\text{comp}} = -\log \frac{e^{\cos(\mathbf{q}_{\text{comp}},\mathbf{d}^+_{\text{img}})/\tau}}{e^{\cos(\mathbf{q}_{\text{comp}},\mathbf{d}^+_{\text{img}})/\tau} + \sum_{D^-} e^{\cos(\mathbf{q}_{\text{comp}},\mathbf{d}^-_{\text{img}})/\tau}}, \tag{13}$$

其中，$\mathbf{d}^{\text{img}} = \text{Trans}(\mathbf{i}_k)$表示文档中图像的表示。上述损失函数背后的直觉是，我们促使图像的表示与查询中包含、但文档文本中未包含的信息对齐。

最后，所提出的 CIEA 的训练目标可表示为：

$$\mathcal{L} = \mathcal{L}_c + \lambda \cdot \mathcal{L}_{\text{comp}}, \tag{14}$$

其中，λ 是一个超参数，用于平衡两种对比损失的贡献。

4 实验

4.1 设置

数据集 我们在 WebQA（Chang 等人，2022）和 EDIS（Liu 等人，2023a）数据集上进行了大量实验，以将所提出的方法与基线方法进行比较。WebQA 是一个开放领域的多模态问答数据集，其中每个查询都关联一个或多个相关文档。该数据集包含 787,697 个仅含文本信息的文档，以及 389,750 个同时含有文本和图像的文档。为清晰起见，在本研究中，我们将该数据集称为 **WebQA-Multi**。此外，为了更好地展示多模态检索的有效性，我们提取所有包含图像的文档，并构建了**WebQA-Image**

表 2：WebQA-Image 和 EDIS 上的比较。最佳结果以粗体标出。

	WebQA-Image				EDIS			
	MRR@10	NDCG@10	MRR@20	NDCG@20	MRR@10	NDCG@10	MRR@20	NDCG@20
CLIP-DPR	59.78	61.05	60.2	63.28	62.52	39.19	62.90	44.24
UniVL-DR	65.95	67.33	66.24	69.01	62.89	39.50	63.29	44.54
T5-ANCE	64.38	65.65	64.73	67.52	63.95	39.74	64.30	43.76
MARVEL-DPR	62.51	63.48	62.90	65.54	63.60	37.76	63.98	41.89
MARVEL	66.43	67.60	66.76	69.40	67.00	42.19	67.28	46.63
CIEA（我们的方法）	**67.40**	**68.77**	**67.74**	**70.51**	**68.11**	**42.57**	**68.42**	**47.19**

数据集。**EDIS** 通过文本查询检索图像，其数据来自包含 100 万张图像的图像语料库（每张图像都配有文本说明），包括 26,000 个训练样本、3,200 个开发样本和 3,200 个测试样本。更多数据集统计信息请见附录A。

指标 为了评估所提出方法和基线方法在多模态检索中的有效性，我们采用了六种广泛使用的指标，包括 MRR@10、MRR@20、NDCG@10、NDCG@20、REC@20 和 REC@100。MRR（Bajaj 等人，2016年）和 NDCG（Taylor 等人，2008）指标使用官方代码计算[1]。

实现细节 在实验中，我们采用 T5-ANCE（Yu 等人，2023年）作为基础语言模型，并使用 CLIP 作为视觉理解模块，以实现所提出的 CIEA 模型，从而增强图像理解能力。通过使用 ANCE（Xiong 等人，2021年）基于 T5（Raffel 等人，2020年）采样负样本文档，增强了 T5-ANCE 的文档表示能力。我们将文本输入长度截断为 128 个词元。在训练过程中，我们采用 AdamW（Loshch ilov 和 Hutter，2019年）优化器，最多训练 40 个 epoch，批次大小为 64，学习率为 5e-6，并将温度超参数 τ 设置为 0.01。超参数 λ 分别为 WebQA-Multi 设置为 0.0011、WebQA-Image 设置为 0.019，以及 EDIS 设置为 0.001。更多实现细节请参见附录B。

4.2 基线方法

我们将 CIEA 与两类不同的基线方法进行比较：分而治之模型和通用稠密检索模型。

分而治之模型建议分别检索图像和文本文档，然后合并检索结果。我们采用各种单模态检索器来实例化分而治之模型，包括 VinVL-DPR（Zhang 等人，2021年）、CLIP-DPR（Radford 等人，2021年）和 BM25（Robertson 等人，2009年）。多模态检索的结果基于单模态倒数排名或预言机模态路由进行合并。后者展现了我们的分而治之模型在检索任务中的最大潜在性能。

关于**通用稠密检索模型**，我们采用以下近期研究作为基线方法：(i) 预训练多模态对齐模型 VinVL-DPR（Zhang 等人，2021年）和 CLIP-DPR（Radford 等人，2021年），它们使用 DPR 框架进行训练；(ii) UniVL-DR（Liu 等人，2023b），该模型采用模态均衡难负样本训练文本和图像编码器，并使用图像语言适配技术弥合图像与文本之间的模态差距；(iii) MARVEL（Zhou 等人，2024年），这是一种基于项目的方法，使用 CLIP 对图像文档进行编码，并将其投影到语言模型的嵌入层中以进行融合。

比较 We-bQA 和 EDIS 上的实验结果分别见表 1和 2。从这些结果可以看出，与分而治之模型和通用稠密检索模型相比，所提出的 CIEA 在多项指标上均取得了显著提升。CIEA 在 WebQA 和 EDIS 上均表现出色，证明了其能够从文本和图像中捕捉关键信息，从而增强多模态检索能力。如表1所示，与采用投影器进行语言模型 grounding 的通用稠密检索模型相比（例如，MARVEL 相较于纯文本方法有所提升，其 MRR@10 比 T5-ANCE 高出 1.3），CIEA

[1]https://github.com/microsoft/MSMARCO-Passage-Ranking/blob/master/ms_marco_eval.py

表 3：使用不同语言模型时的实验结果。仅文本 设置表示训练时仅使用文档标题，不加入图像信息；而基于项目 设置是最常见的投影方法，不进行视觉对齐。

语言模型	设置	mrr@10	ndcg@10	mrr@20	NDCG@20	REC@20	REC@100
T5-ANCE	仅文本	64.13	62.03	64.40	64.02	83.81	92.07
	基于项目	65.43	63.07	65.67	65.00	84.32	92.27
	CIEA	**66.16**	**63.89**	**66.41**	**65.85**	**85.43**	**92.75**
BERT	仅文本	61.39	58.20	61.69	60.32	79.59	89.10
	基于项目	63.03	60.20	63.31	62.37	**81.80**	**90.78**
	CIEA	**63.63**	**60.46**	**63.90**	**62.49**	81.51	90.57
BART	仅文本	60.14	56.82	60.49	59.14	78.41	89.25
	基于项目	62.67	59.70	62.98	62.04	81.77	90.77
	CIEA	**63.73**	**60.43**	**64.04**	**62.72**	**81.96**	**90.98**
GPT2	仅文本	54.68	51.49	55.07	53.77	72.91	84.86
	基于项目	58.59	55.00	58.92	57.15	75.84	86.79
	CIEA	**59.25**	**55.35**	**59.59**	**57.62**	**76.47**	**86.92**
GPT-2-LARGE	仅文本	63.58	60.44	63.86	62.45	81.24	89.98
	基于项目	64.33	61.98	64.60	62.04	83.97	92.11
	CIEA	**65.38**	**63.00**	**65.62**	**64.96**	**84.96**	**92.68**

表4：WebQA-Multi 上的消融研究。这里，“w/o”表示“不含”，而“基线”指移除两个模块的设置。

	MRR@10	NDCG@10	REC@20	REC@100
CIEA	**66.03**	**63.70**	**85.14**	**92.63**
不含图像查询	65.66	63.44	84.74	92.54
无注意力机制	65.90	63.39	84.95	92.60
Base	65.30	63.19	84.49	92.40

的 MRR@10 进一步提高了 0.74，表2中也呈现了类似的提升。这些由 CIEA 带来的进步凸显了互补信息对齐的关键优势——这一框架能够更好地建模视觉信息，并增强基于项目的模型的表征能力。

消融研究我们进行了消融研究，以展示图像查询和注意力层的作用。具体而言，我们分别移除了用于重新加权的注意力层（参见公式（9））以及图像查询的使用（参见公式（13））。此外，我们还通过移除这两个组件进行基线设置。

实验结果如表 4 所示。尽管移除图像查询并未导致 REC@100 下降，但其他指标均出现不同程度的下降，尤其是 REC@20，从 85.14 降至 84.74。另一方面，移除注意力机制则会导致 NDCG@10 指标显著下降，凸显了互补信息提取的重要性。此外，我们还将基线配置与同时移除两个模块的情况进行了比较。结果表明，与基线相比，移除单个模块会使多个指标有所提升，这说明图像查询和注意力层均有助于提高多模态检索的准确率。更多结果见附录 C。

4.3 进一步讨论

语言模型在 CIEA 中的作用为了进一步讨论所提出的 CIEA 中所使用的语言模型的作用，我们使用不同的语言模型作为骨干网络进行实验，包括 T5-ANCE （Yu 等人，2023 年）、 BART （Lewis 等人，2020a）、BERT（Devlin 等人，2019年）、GPT-2 和 GPT-2-LARGE（Radford 等人，2019年）。这些语言模型涵盖了三种主流架构，即编码器-解码器、仅编码器和仅解码器。更多实现细节请参见附录 D。

表3 中的实验结果表明，骨干语言模型对多模态检索效果有显著影响。对于纯文本检索，使用

表5：WebQA-Multi 上不同权重的实验结果。

权重	MRR@10	NDCG@10	REC@20	REC@100
不相似	**66.03±0.012**	**63.70 ± 0.030**	**85.14±0.052**	**92.63±0.008**
相似	65.60±0.038	63.35±0.005	85.00±0.012	92.61±0.003

更大参数规模的语言模型（如 GPT-2 LARGE 和 T5-ANCE）能够取得更好的整体性能。基于项目的模型和 CIEA 均受益于强大的骨干网络，并通过图像衍生信息提升性能。值得注意的是，与基于投影的方法相比，CIEA 在不同骨干网络下都能保持有竞争力的结果，证实了其稳健性。

注意力权重选择 采用最大余弦相似度计算图像块级相似度，旨在提取与文本描述对应程度较低的图像区域。为确定最佳提取方法，我们进行了实证实验，保留原始余弦值而非其补值。如表 5所示，显式计算不相似度已被证明更有利于捕获细粒度的图像块级图文映射。

计算效率 尽管所提出的双重损失可能会增加训练复杂度，但损失的计算相对独立，不会导致整体复杂度呈倍数增长。与同样基于项目的方法 MARVEL 相比，所提出的方法实现了相近的计算效率。例如，在示例设备上，使用 MARVEL 在包含 4,966 个样本的 WebQA-Multid数据集上进行训练约需 6.1 分钟，而 CIEA 约需 6.5 分钟。

案例研究 为了更好地理解，我们在WebQA数据集上进行了案例研究。我们使用余弦相似度，从语言模型的词汇表中找出与图像投影得到的嵌入最接近的词，以展示从图像中提取了哪些信息。我们将提出的CIEA与最强基线 MARVEL进行比较。结果如图 3所示，这表明 MARVEL关注的是与文本相近的图像信息，而CIEA则能从图像中捕获更多补充信息。例如，CIEA捕获了一些术语，如 绿色、叶子和时钟，它们在语义上与图像相关，但未在图注中提及。第二个案例中的术语时钟 和第三个案例中的花 与查询相关，有助于模型高效地定位和检索相关内容。这些结果进一步证实了CIEA在多模态检索方面的优势。关于失败案例的讨论见附录E，这进一步展示了模型在具有挑战性的场景中的行为，并为未来的改进提供启发。

5 结论

本文提出了 CIEA，以提升多模态检索的有效性。CIEA 的核心思想是增强对多模态数据中互补信息的捕捉能力。具体而言，CIEA 利用语言模型和 CLIP 将多模态文档转换到统一潜在空间。为提取互补信息，我们计算文本与图像之间的图像块级距离，并据此对图像表征进行重新加权。在 CIEA 的优化方面，除了使用对比损失来学习文本的语义，我们还促使图像表征与查询中的互补信息对齐。我们进行了一系列实验，展示了与两种不同类型的多模态检索方法相比，CIEA 的优势。进一步探讨语言模型的影响，结果表明 CIEA 具有鲁棒性；此外，我们还纳入了若干案例研究，以便更好地理解该方法。

局限性

多模态检索正处于快速发展阶段。本文的探索仅限于只包含文本的查询，并重点关注从图像中提取信息。将所提出的方法扩展到音频和视频等更多类型的多模态数据，是未来研究的一个有前景的方向。此外，尽管我们使用多种语言模型进行了实验，但受计算资源限制，尚未探索更大的模型。



图注： 东京台场彩虹大桥 - 索尼 A7R (11812224916)

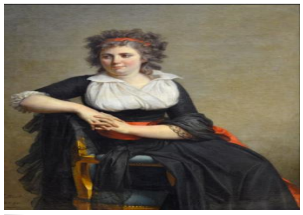
查询： 美洲冬青或臭千里光的花，哪种更大?
MARVEL 最邻近词元： ['麦芽汁'、'种子'、'假'、'细菌'、'代表'、'浆果'、'细菌'、'细菌的'，'藻类'、'本土的'、'朗姆酒']
CIEA 最近邻词元： ['麦芽汁', '绿色', '种子', 'it', '种子', '假', '代表', '叶子', '芽', '花', 'cum', '浆果']



图注： Durgin Park 2009年 Durgin Park 的入口照片，它是最古老的餐馆之一位于波士顿，靠近法尼尔厅。
查询： 波士顿 Durgin Park 正门外的时钟是什么颜色?
MARVEL 最近邻词元： ['啤酒厂', 'the', 'Lake', 'bou', 'stau', 'restaurant', 'restaur', '面包店', '以下', '居民', '商店', '乳制品', '开幕', '历史的', '坐落于']
CIEA 最近邻词元： ['啤酒厂', 'bou', 'stau', 'restaurant', '时钟', '面包店', '时钟', 'locaux', 'operates', '地点', '蛋黄', '商店', '出口', '查看', '博物馆']



图注： 苔生虎耳草（习性）阿莱奇冰川附近的苔生虎耳草，瑞士
查询： 苔生虎耳草的每根花茎上生长多少朵花?
MARVEL 最近邻词元： ['细菌'、'the'、'代表'、'浆果'、'细菌'、'omato'、'cum'、'本土的'、'rum']
CIEA最近邻词元： ['cum', '花', '花', 'alb', '代表', '浆果', 'timp', '个人', '年轻', '蓝色', '担忧', 'omato', '芽', 'um', 'acum', 'rum']



图注： 奥维利耶侯爵夫人肖像，雅克-路易·大卫，1790年 肖像日内瓦银行家雅克·里耶的两姐妹之一
查询： 《伊丽莎白·维杰·勒布伦自画像》中的女性和《奥维利耶侯爵夫人肖像》中的女性是否都穿着带红丝带的黑色连衣裙?穿着带红丝带的黑色连衣裙?
MARVEL 最近邻词元： ['占主导地位'、'该'、'代表'、'经常'、'玛丽'、'acum'、'黛博拉'、'ographer'，'祭坛'、'母亲'、'弗朗西斯'、'女演员'、'Sfant']
CIEA最近邻词元： ['connaissance'，'连衣裙'，'代表'，'富有表现力'，'$75'，'多萝西'，'黛博拉'，'ographer'，'母亲'，'祭坛'，'弗朗西斯'，'acum'，'女演员'，'Sfant']

图3：案例研究。最近邻词元表示词汇表中与视觉嵌入最接近的嵌入，重复项已移除。红色文字表示与图像相关、但在 CIEA 中 MARVEL 未找到的术语。

致谢

本研究得到广东省重点领域研发计划（编号：2024B1111100001）资助。

参考文献

白金泽、白帅、储云飞、崔泽宇、党凯、邓晓东、范阳、葛文斌、Yu Han、黄飞等。2023年。通义千问技术报告。*arXiv*预印本 *arXiv:2309.16609*。

Payal Bajaj、Daniel Campos、Nick Craswell、邓力、Jianfeng Gao、刘晓东、Rangan Majumder、Andrew McNamara、Bhaskar Mitra、Tri Nguyen 等。2016年。MS MARCO：一个人工生成的机器阅读理解数据集。*arXiv*预印本 *arXiv:1611.09268*。

Alberto Baldrati、Lorenzo Agnolucci、Marco Bertini和Alberto Del Bimbo。2023年。通过文本反演实现零样本组合图像检索。载于*IEEE/CVF*国际计算机视觉会议论文集，第15338–15347页。

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。2020年。语言模型是少样本学习器。神经信息处理系统进展，33:1877–1901。

Yingshan Chang、Mridu Narang、Hisami Suzuki、Gui- hong Cao、Jianfeng Gao 和 Yonatan Bisk。2022。WebQA：多跳与多模态问答。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 16495–16504 页。

Chang Che、Qunwei Lin、Xinyu Zhao、Jiaxin Huang 和 Liqiang Yu。2023年。通过基于 CLIP 的图像到文本转换增强多模态理解。载于*2023*年第六届国际大数据技术会议论文集，第 414–418 页。

Jiawei Chen、Hongyu Lin、Xianpei Han 和 Le Sun。2024年。对检索增强生成中的大语言模型进行基准测试。载于*AAAI*人工智能会议论文集，第 38 卷，第 17754–17762 页。

Xin Cheng、Di Luo、Xiuying Chen、Lemao Liu、Dongyan Zhao 和 Rui Yan。2023年。提升自我：借助自记忆实现检索增强文本生成。神经信息处理系统进展，36：43780–43799。

Hyung Won Chung、Le Hou、Shayne Longpre、Barret Zoph、Yi Tay、William Fedus、Yunxuan Li、Xuezhi Wang、MostafaDehghani、Siddhartha Brahma 等。

2024年。扩展指令微调语言模型。机器学习研究期刊，25(70):1–53。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：面向语言理解的深层双向变换器预训练。载于计算语言学协会北美分会*2019*年会议论文集，第4171–4186页。

Yunfan Gao、Yun Xiong、Xinyu Gao、Kangxiang Jia、Jinliu Pan、Yuxi Bi、Yi Dai、Jiawei Sun、Meng Wang 和 Haofen Wang。2023年。面向大语言模型的检索增强生成：综述。*arXiv*预印本 *arXiv:2312.10997*。

Jiayi Kuang、Ying Shen、Jingyou Xie、Haohao Luo、Zhe Xu、Ronghao Li、Yinghui Li、Xianfeng Cheng、Xika Lin 和 Yu Han。2025年。视觉问答中的多模态大语言模型自然语言理解与推理：综述。*ACM*计算调查，57(8)：1–36。

Mike Lewis、Yinhan Liu、Naman Goyal、Marjan Ghazvininejad、Abdelrahman Mohamed、Omer Levy、Veselin Stoyanov 和 Luke Zettlemoyer。2020a。BART：面向自然语言生成、翻译和理解的去噪序列到序列预训练。载于第*58*届计算语言学协会年会论文集，第7871–7880页。

Patrick Lewis、Ethan Perez、Aleksandra Piktus、Fabio Petroni、Vladimir Karpukhin、Naman Goyal、Heinrich Küttler、Mike Lewis、Wen-tau Yih、Tim Rocktäschel等。2020b。面向知识密集型自然语言处理任务的检索增强生成。神经信息处理系统进展，33：9459–9474。

Junnan Li、Dongxu Li、Silvio Savarese 和 Steven Hoi。2023。BLIP-2：利用冻结图像编码器和大型语言模型自举语言-图像预训练。载于国际机器学习会议，第19730–19742页。

Wei Li、Hehe Fan、Yongkang Wong、Yi Yang 和 Mo-han Kankanhalli。2024年。通过多模态组合学习改善多模态大语言模型中的上下文理解。在国际机器学习会议，第27732–27751页。

Zhang Li、Biao Yang、Qiang Liu、Zhiyin Ma、Shuo Zhang、Jingxu Yang、Yabo Sun、Yuliang Liu 和 Xiang Bai。2024年。Monkey：图像分辨率和文本标签对于大型多模态模型十分重要。在*IEEE/CVF*计算机视觉与模式识别会议论文集，第26763–26773页。

Siqi Liu、Weixi Feng、Tsu-Jui Fu、Wenhu Chen 和 William Wang。2023a。EDIS：面向多模态网络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。

Zhenghao Liu、Chenyan Xiong、Yuanhuiyi Lv、Zhiyuan Liu 和 Ge Yu。2023b。通用视觉-语言稠密检索：为多模态检索学习统一表示空间。载于国际学习表征会议。Ilya Loshchilov 和 Frank Hutter。2019年。解耦权重衰减正则化。载于国际学习表征会议。Pan Lu、Baolin Peng、Hao Cheng、Michel Galley、Kai-Wei Chang、Ying Nian Wu、Song-Chun Zhu 和 Jianfeng Gao。2023年。Chameleon：利用大语言模型进行即插即用的组合推理。神经信息处理系统进展，36：43447–43478。

Anas Mahmoud、Mostafa Elhoushi、Amro Abbas、Yu Yang、Newsha Ardalani、Hugh Leather 和 Ari S Morcos。2024年。Sieve：使用图像描述模型进行多模态数据集剪枝。在*IEEE/CVF*计算机视觉与模式识别会议论文集，第22423–22432页。

Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sas- try、Amanda Askell、Pamela Mishkin、Jack Clark 等。2021年。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。

Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等。2019年。语言模型是无监督多任务学习器。*OpenAI* 博客，1(8):9。

Colin Raffel、NoamShazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J Liu。2020年。探索统一文本到文本转换器迁移学习的局限。机器学习研究期刊，21(140):1–67。

Stephen Robertson、Hugo Zaragoza 等。2009年。《概率相关性框架：BM25 及其后续发展》。信息检索基础与趋势，3(4):333–389。

Weijia Shi、Sewon Min、Michihiro Yasunaga、Min-joon Seo、Richard James、Mike Lewis、Luke Zettlemoyer 和 Wen-tau Yih。2024年。《Replug：检索增强型黑盒语言模型》。载于*2024*年北美计算语言学协会分会会议论文集，第8364–8377页。

Xi’ao Su、Ran Wang 和 Xinyu Dai。2022。《用于多标签文本分类的对比学习增强最近邻机制》。载于计算语言学协会第*60*届年会论文集，第672–679页。

Michael Taylor、John Guiver、Stephen Robertson 和 Tom Minka。2008 年。S oftrank：优化非光滑排序指标。载于*2008* 年国际网络搜索与数据挖掘会议论文集，第 77–86 页。

Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timothée Lacroix、Baptiste Rozière、Naman Goyal、Eric Hambro、Faisal Azhar 等。2023年。Llama：开放且高效的基础语言模型。*arXiv* 预印本 *arXiv:2302.13971*。

Jianfeng Wang、Zhengyuan Yang、Xiaowei Hu、Linjie Li、Kevin Lin、Zhe Gan、Zicheng Liu、Ce Liu 和 Li- juan Wang。2022年。GIT：面向视觉与语言的生成式图像到文本转换器。《机器学习研究汇刊》。

Xiaohua Wang、Zhenghua Wang、Xuan Gao、Feiran Zhang、Yixin Wu、Zhibo Xu、Tianyuan Shi、Zhengyuan Wang、Shizheng Li、Qi Qian 等。2024年。探索检索增强生成的最佳实践。载于*2024*年自然语言处理经验方法会议论文集，第 17716–17736 页。

Zirui Wang、Jiahui Yu、Adams Wei Yu、Zihang Dai、Yulia Tsvetkov 和 Yuan Cao。2022年。SimVLM：采用弱监督的简易视觉语言模型预训练。载于国际学习表征会议。

Ziyang Wang、Heba Elfardy、Markus Dreyer、Kevin Small 和 Mohit Bansal。2024年。通过冻结的大语言模型实现多模态检索的统一嵌入。在计算语言学协会论文集精选，第1537–1547页。

Wenhao Wu、Haipeng Luo、Bo Fang、Jingdong Wang 和 Wanli Ouyang。2023年。Cap4video：辅助描述能为文本视频检索带来什么？载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第10704–10713页。

Lee Xiong、Chenyan Xiong、Ye Li、Kwok-Fung Tang、Jialin Liu、Paul N. Bennett、Junaid Ahmed 和 Arnold Overwijk。2021年。用于稠密文本检索的近似最近邻负对比学习。载于国际学习表征会议。

Zihao Yi、Jiarui Ouyang、Yuwen Liu、Tianhao Liao、Zhe Xu 和 Ying Shen。2024年。基于大语言模型的多轮对话系统最新进展综述。*arXiv*预印本 *arXiv:2402.18013*。

Shi Yu、Zhenghao Liu、Chenyan Xiong 和 Zhiyuan Liu。2023年。Openmatch-v2：一款一体化多模态、基于预训练语言模型的信息检索工具包。载于第*46*届国际 *ACM SIGIR* 信息检索研究与开发大会论文集，第3160–3164页。

Xin Yuan、Zhe Lin、Jason Kuen、Jianming Zhang、Yilin Wang、Michael Maire、Ajinkya Kale 和 Baldo Fai- eta。2021年。用于视觉表示学习的多模态对比训练。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第6995–7004页。

Pengchuan Zhang、Xiujun Li、Xiaowei Hu、Jianwei Yang、Lei Zhang、Lijuan Wang、Yejin Choi 和 Jian- feng Gao。2021年。VinVL：重新审视视觉语言模型中的视觉表示。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第5579–5588页。

Ruochen Zhao、Hailin Chen、Weishi Wang、Fangkai Jiao、Xuan Long Do、Chengwei Qin、Bosheng Ding、Xiaobao Guo、Minzhi Li、Xingxuan Li 和 Shafiq Joty。2023年。用于增强生成的多模态信息检索：一项综述。载于计算语言学协会论文集：*EMNLP 2023*，第4736–4756页。

Tianshuo Zhou、Sen Mei、Xinze Li、Zhenghao Liu、Chenyan Xiong、Zhiyuan Liu、Yu Gu 和 Ge Yu。2024年。MARVEL：通过视觉模块插件释放密集检索的多模态能力。载于计算语言学协会第*62*届年会论文集，第14608–14624页。

Xiangrong Zhu、Yuexiang Xie、Yi Liu、Yaliang Li 和 Wei Hu。2025。知识图谱引导的检索增强生成。载于*2025* 年美洲计算语言学协会分会会议论文集，第 8912–8924 页。

A 数据集

表6：所采用的 WebQA 的统计信息。

模态	文档	查询		
		训练集	Dev	Test
图像	389,750	16,400	2,554	2,511
Text	787,697	15,366	2,446	2,455
总计	1,177,447	31,766	5,000	4,966

WebQA 是一个开放域问答数据集，其中每个查询都关联一个或多个多模态文档，以辅助生成回答。WebQA 分为三个部分：训练集、开发集和测试集，其数据统计信息如表6所示。检索语料库包含 389,750 篇包含视觉信息和图像说明的图像文档，以及 787,697 篇纯文本文档。我们划分出的 WebQA-Image 仅使用图像文档进行测试，以评估 CIEA 在多模态建模中的有效性。相比之下，WebQA-Multi 从总计 1,177,447 篇文档中进行检索，以评估模型能否准确表示文本信息。

WebQA 被许多基线方法使用，例如 MARVEL（Zhou 等人，2024年）。MARVEL 也使用了 ClueWeb22-MM。我们已向卡内基梅隆大学申请该数据集；遗憾的是，由于出口管制限制，我们的申请被拒。因此，我们也使用 EDIS 来评估我们的方法。EDIS（Liu 等人，2023a）是一个综合性数据集，由来自 Google 的 100 万组图文对组成。该数据集包含一个由 26,000 组图文对构成的训练集，以及验证集和测试集，两者各包含 3,200 组图文对。EDIS 的实体数量高达 4.03，体现了丰富多样的语义内容，使其成为图文检索任务研究中的宝贵资源。

B 实现细节

在我们的实验中，我们使用 T5-ANCE（Yu 等人，2023年）作为基线语言模型，并使用 CLIP 作为视觉理解模块来构建 CIEA 模型，从而增强 T5-ANCE 的图像理解能力。为便于比较，我们使用 MARVEL 的预训练投影器初始化投影器。视觉编码器则使用 OpenAI 提供的 clip-vit-base-patch32 检查点进行初始化[2]。我们将文本输入长度截断为 128 个词元。

在训练过程中，我们使用 AdamW（Loshchilov 和 Hutter，2019年）优化器，将最大训练轮数设为 40，批次大小设为 64，学习率设为 5e-6，温度超参数 τ 设为 0.01。我们遵循 UniVL-DR（Loshchilov 和 Hutter，2019年）的设置，采用 ANCE 采样方法进行训练。以使用批内负例微调的 CIEA-DPR 模型为基础，我们继续训练 CIEA-DPR，以实现平衡的模态，并引入困难负例。在评估阶段进行负例采样时，我们将训练集打乱，并在同一批次内选择其他样本作为 DPR 中的负例；而在 ANCE 方法中，我们使用 CIEA-DPR 检索结果中相似度最高的前 100 个样本作为困难负例。所有模型每 500 步评估一次，并将早停步数设为 5。对于参数λ，我们进行网格搜索，并选择在验证集上损失最低的参数，将其分别设为 WebQA-Multi 的 0.0011、WebQA-Image 的 0.019 和 EDIS 的 0.001。WebQA-Multi 数据集的各种基线方法结果见 MARVEL（Zhou 等人，2024年），而其他基线方法的实验结果则使用原论文提供的开源代码复现。所有实验均在单张 NVIDIA A100 上进行。

C 统计显著性

为避免结果只是偶然的潜在疑虑，我们使用不同的随机种子重复实验，并在表8中报告平均值和标准差。结果表明，与 MARVEL 相比，所提出的 CIEA 实现了显著且稳定的提升。此外，不同模块均对 ICEA 的整体性能做出了积极贡献。

D 语言模型的影响

基于投影的方法以语言模型作为骨干网络，因此语言模型的性能是影响检索效果的关键因素之一。为验证我们的方法能否在以不同语言模型为骨干网络时有效运行，我们进

[2]https://huggingface.co/sentence-transformers/clip-ViT-B-32

表 7：不同语言模型在 DPR 训练框架上的结果。

语言模型	设置	MRR@10	NDCG@10	MRR@20	NDCG@20	REC@20	REC@100
T5-ANCE	仅文本	52.67	45.69	53.08	48.48	67.70	82.07
	基于项目	56.91	54.26	57.31	56.88	77.20	89.26
	CIEA	**57.49**	**54.80**	**57.86**	**57.24**	**77.21**	**89.41**
BERT	仅文本	39.50	36.23	40.06	38.69	55.37	72.72
	基于项目	52.65	49.41	53.21	51.97	71.16	85.40
	CIEA	**53.40**	**49.96**	**53.83**	**62.56**	**71.95**	**85.72**
BART	仅文本	24.52	21.79	25.04	23.98	36.91	54.66
	基于项目	50.46	47.72	50.94	50.04	69.40	83.69
	CIEA	**50.97**	**48.06**	**51.43**	**50.72**	**70.44**	**84.81**
GPT2	仅文本	20.38	18.62	20.86	20.19	29.97	44.33
	基于项目	**45.94**	42.45	46.46	45.08	63.16	78.34
	CIEA	45.82	**42.60**	**46.51**	**45.13**	**63.46**	**78.91**
GPT-2-LARGE	仅文本	37.25	34.52	37.67	37.11	54.37	73.08
	基于项目	51.44	48.98	51.89	51.64	72.06	86.61
	CIEA	**54.28**	**51.81**	**54.68**	**54.39**	**74.94**	**88.52**

表 8：WebQA-Multi 上的标准差值。

模型	MRR@10 NDCG@10	REC@100
CIEA	**66.03±0.012 63.70±0.030**	**92.63±0.008**
无图像查询	65.66±0.008 63.44±0.072	92.54±0.003
无注意力	65.90±0.098 63.39±0.003	92.60±0.001
MARVEL	65.30±0.017 63.19±0.005	92.40±0.012

行了以下实验：使用 T5-ANCE（Yu 等人，2023年）、BART（Lewis 等人，2020a）、BERT（Devlin 等人，2019年）、GPT-2 和 GPT-2-LARGE（Radford 等人，2019年），涵盖编码器-解码器、仅编码器和仅解码器架构。对于编码器-解码器模型，我们遵循 MARVEL 的方法（Zhou 等人，2024年），将多模态文档输入编码到编码器中，同时在解码器部分输入一个“\s”字符，并使用其最后一层隐藏层表示作为文档表示。对于仅编码器模型，我们使用第一个字符的表示；对于仅解码器模型，我们使用最后一个词元的最后一层隐藏层表示，因为每个词元只会对前面的词元计算注意力。由于 GPT-2 LARGE 的参数数量庞大，为避免内存问题，我们将其批次大小设为 16，其他模型的批次大小则设为 64。除表 3中展示的使用 ANCE 负采样方法训练的结果外，我们还在表 7中列出了使用 DPR 的批内负采样方法训练的结果。显然，尽管 ANCE 的表现始终优于 DPR，CIEA 在这两种负采样方法下的表现均优于简单的基于投影的方法和纯文本方法，这表明所提出的该方法与图像信息的对齐效果更好。

E 失败案例分析 s

我们在表 9中提供了失败案例分析，以便更好地理解所提出的方法并促进后续研究。每个案例都展示了真实图像-描述对，以及 CIEA 检索到的排名前三的候选项。由于排名靠前的结果中均没有与真实图像-描述对匹配的项，因此所有实例都被归类为检索失败。前两个案例展示了模型在捕捉颜色属性方面取得的部分成功（例如，“橙色”和“绿色”），但模型可能难以建立精确的颜色-物体映射（例如，“天花板”和“瓷砖”），尤其是在多物体环境中。第三个案例揭示了动作解释方面更为根本的局限：检索结果与目标动作“放松”几乎无关。这些失败凸显了视觉-语义对齐中持续存在的挑战，尤其是在物体属性的上下文绑定和对动作表示的动态解释方面。

问题： 新加坡乌节路上的哪家餐馆有层叠式橙色天花板?			
金色	**检索（前 3 名）**		
			
咖啡俱乐部，乌节道路 (8171776629) 位于威洛克广场，新加坡。	乌节路12号 ，圣诞节，12月6日 乌节 Road，圣诞节亮灯活动 2006，新加坡。	乌节路 ，Xmas，12月6日 乌节路 ，圣诞亮灯活动 2006，新加坡。	乌节路15号 ，圣诞节，12月6日乌节 Road，圣诞节2006年圣诞亮灯活动，新加坡。
问题： 哪尊佛像前面铺有绿色瓷砖?			
金色	**检索（前3名）**		
			
1号佛像 佛像位于侯赛因萨加尔湖	佛陀伊甸园（34）	佛陀 1251876	建造佛像 – Panoramio
问题： 人们可以在哪里埃菲尔铁塔旁的草地上休息?			
金色	**检索（前 3 名）**		
			
埃菲尔铁塔，巴黎 17 2010年9月。	FW 埃菲尔铁塔。	柏林 ，米特区，亚历山大广场，电视塔 - Panoramio。	埃菲尔铁塔.

表 9：失败案例分析。