Statistics and Application 统计学与应用, 2025, 14(6), 29-41
Published Online June 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.146145

Hans汉斯

ProMahaVQA：基于原型学习与对比损失的零样本视觉问答性能提升研究

闫婧昕

北京建筑大学理学院，北京

收稿日期：2025年5月13日；录用日期：2025年6月6日；发布日期：2025年6月18日

摘　要

视觉问答(VQA)是一项复杂的人工智能任务，要求模型理解图像内容与自然语言问题，实现跨模态语义融合。然而，现有方法在处理视觉与语言深度交互方面存在明显不足，尤其在零样本场景中泛化能力有限。为此，本文提出ProMahaVQA模型，引入跨模态原型矩阵、原型查询机制与基于马氏距离的多标签对比损失，有效提升了特征判别能力与模型鲁棒性。模型首次将原型学习机制应用于零样本VQA任务，并通过记忆矩阵支持对未见答案的识别。实验结果表明，ProMahaVQA在F-VQA、TZSL和GZSL等设置下均显著优于现有方法，展现出卓越的泛化性能与跨模态推理能力。

关键词

视觉问答，原型学习，马氏距离，零样本学习，跨模态融合

ProMahaVQA: Enhancing Zero-Shot Visual Question Answering with Prototype Learning and Contrastive Loss

Jingxin Yan

School of Science, Beijing University of Civil Engineering and Architecture, Beijing

Received: May 13th, 2025; accepted: Jun. 6th, 2025; published: Jun. 18th, 2025

Abstract

Visual Question Answering (VQA) is a challenging artificial intelligence task that requires models to comprehend image content and natural language questions through cross-modal semantic integration. However, existing methods often struggle with deep visual-language interactions, particularly

文章引用: 闫婧昕. ProMahaVQA: 基于原型学习与对比损失的零样本视觉问答性能提升研究[J]. 统计学与应用, 2025, 14(6): 29-41. DOI: 10.12677/sa.2025.146145

in zero-shot scenarios where generalization is limited. To address these challenges, we propose ProMahaVQA, a novel model that incorporates a cross-modal prototype matrix, a prototype query mechanism, and a Mahalanobis distance-based multi-label contrastive loss. These innovations significantly enhance feature discrimination and model robustness. Notably, this work is the first to integrate prototype learning into zero-shot VQA, enabling the model to recognize unseen answers via a memory matrix. Experimental results on F-VQA, TZSL, and GZSL benchmarks demonstrate that ProMahaVQA substantially outperforms existing approaches, exhibiting superior generalization and cross-modal reasoning capabilities.

Keywords

Visual Question Answering, Prototype Learning, Mahalanobis Distance, Zero-Shot Learning, Cross-Modal Fusion

1. 基于原型学习的零样本视觉问答模型

1.1. 任务定义

零样本视觉问答[1]旨在预测问题 $q \in Q$ 关于图像 $I \in I$ 的答案 $a \in A$，其中答案 a 在训练过程中未曾出现(即属于未见类别) [2]。该任务可形式化定义为一个学习函数：

$$f : I \times Q \to A \tag{1}$$

目标是求解最可能的答案：

$$a^* = arg\max_{a \in A} P\left(a \middle| I, q\right) \tag{2}$$

其中，$P\left(a \middle| I, q\right)$ 表示在给定图像 I 和问题 q 的条件下[3]，答案 a 出现的概率。为了有效建模视觉–语言联合表示，首先计算图像特征 v_I 和问题特征 v_I，并通过融合函数 Φ 进行跨模态信息整合，得到联合表示 $v_{I,q}$：

$$v_{I,q} = \Phi\left(v_I, v_q\right) \tag{3}$$

在零样本学习(ZSL) [4]设置下，为了实现对未见类别的泛化能力，模型需将 $v_{I,q}$ 对齐到原型空间中的某个原型 $p_k \in P$，其 P 由未见类别的特征中心组成。最终，答案的概率通过距离度量函数 d 计算为：

$$P\left(a \middle| I, q\right) \propto exp\left(-d\left(v_{I,q}, p_k\right)\right) \tag{4}$$

在本节中，首先介绍所提出的框架的整体流程，该模型整体名为 ProMahaVQA，由图像特征提取、支持知识提取和问题特征提取三部分组成，基于 OpenCLIP [5]编码器生成多模态特征后，通过跨模态原型矩阵与原型查询模块实现高效语义匹配与特征对齐，并引入基于马氏距离的多标签对比损失进行优化，以提升模型在零样本 VQA 任务中的泛化能力与鲁棒性。随后，本节对模型在多个数据集上的表现进行了对比实验、消融分析与可视化验证，系统评估了各模块在提升准确性与跨模态推理能力方面的贡献。

1.2. 模型概述与整体框架

本章提出的整体 ProMahaVQA 框架如图 1 所展示，该模型主要由图像特征提取、支持知识提取、问

题特征提取三个部分组成，并采用 OpenCLIP 模型(OpenCLIP 是对 OpenAI 提出的 CLIP 模型的开源实现与扩展，由 LAION 社区开发，支持在大规模开放数据集上训练更大规模的模型)。作为图像和文本编码器[6]。图像特征与支持知识特征首先通过一个全连接层进行融合，以形成多模态特征表示。

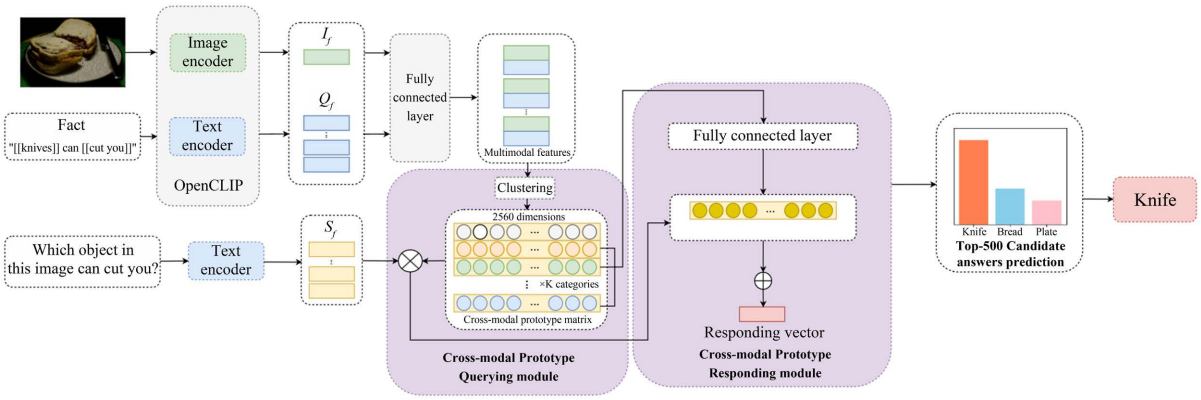

Figure 1. Overall framework of the ProMahaVQA model
图 1. ProMahaVQA 模型的整体框架

问题特征则用于跨模态原型查询模块，以与跨模态原型矩阵进行匹配。该原型矩阵通过视觉和文本特征聚类生成，并包含 2560 维的跨模态特征。模型基于余弦相似度选择最相关的原型，并将其输入跨模态原型响应模块。最终，该模块生成响应向量，并通过基于马氏距离的多标签对比损失进行优化，以提升模型的区分能力。在推理阶段，模型计算 500 个候选答案的概率分布，并选择概率最高的答案作为最终输出。例如，在图 1 所示的案例中，模型最终选择“Knife”作为答案。

1.3. 基于 OpenCLIP 的特征提取模块

本研究采用 OpenCLIP 模型提取 F-VQA 数据集中的图像特征、支持知识特征和问题特征，具体包括以下步骤：

1) 图像特征提取(Image Feature Extraction)

首先，对输入图像 I 进行预处理：调整大小至 224×224 像素并归一化，以匹配模型的输入要求。接着，将图像划分为固定大小的 N 个 Patch，每个 Patch 的尺寸为 16×16 像素。Patch 数量 N 计算如下：

$$N = \frac{P^2}{H \times W} \tag{5}$$

其中 H 和 W 分别为图像的高度和宽度，P 为 Patch 的尺寸。每个 Patch 通过线性投影层投影到固定大小的向量空间，形成 Patch 嵌入：

$$e_i = W_p \cdot x_i + b_p \tag{6}$$

其中 W_p 和 b_p 为线性投影层的参数。为了保留图像 Patch 之间的相对位置信息，加入位置编码：

$$z_i = e_i + pos_i \tag{7}$$

这些位置编码后的 Patch 嵌入输入 Transformer 编码器，该编码器由多个自注意力层和前馈神经网络层组成。自注意力机制计算如下：

$$Attention(Q,K,V) = Soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{8}$$

其中，Q、K 和 V 分别表示查询(Query)、键(Key)和值(Value)矩阵，d_k 为 Key 的维度。经过 L 层 Transformer 计算后，每个 Patch 的最终特征表示为：

$$z_i^{(l)} = \mathrm{FFN}\left(\mathrm{Attention}\left(Q,K,V\right)\right) \tag{9}$$

最终，Transformer 编码器输出的全局分类 Token (CLS Token) z_0^L 作为整幅图像的全局特征表示：

$$I_f = z_0^L \tag{10}$$

2) 文本特征提取(Textual Feature Extraction)

文本特征提取模块对问题文本和支持知识进行编码，以构建统一的语义表示[7]。使用 OPENCLIP 模型的文本编码器处理文本数据，包括问题文本和知识文本。假设问题文本由 n 个词构成，表示为序列 $\left[w_1,w_2,\cdots,w_n\right]$，支持知识文本包含 m 个词，表示为$\left[s_1,s_2,\cdots,s_m\right]$。每个词通过嵌入矩阵 E 映射到向量空间：

$$e_i = E\cdot q_i \tag{11}$$

$$e_i = E\cdot s_i \tag{12}$$

这些词向量输入 LSTM 网络，以捕捉文本的上下文信息：

$$h_t = \mathrm{LSTM}\left(e_t,h_{t-1}\right) \tag{13}$$

最终，LSTM 网络的最后一个隐藏状态(Final Hidden State)作为文本的全局表示：

$$Q_f = h_T \tag{14}$$

$$S_f = h_T \tag{15}$$

1.4. 跨模态原型矩阵生成(Cross-Modal Prototype Matrix Generation)

在视觉问答任务中，跨模态特征的有效整合与交互是生成准确答案的关键。为提升 VQA 任务中的跨模态推理能力，我们提出了一种基于跨模态原型的网络结构，该方法通过学习和存储跨模态原型来增强模型性能。本模型的核心思想是对视觉特征和文本特征进行聚类，生成原型特征，最终形成跨模态原型矩阵[8]。该原型矩阵的初始化方式不同于传统的随机初始化，而是采用视觉特征与文本特征的聚类中心进行初始化，从而确保初始原型具备有意义的语义信息。具体而言，为了优化跨模态特征的融合，我们首先从 F-VQA 数据集提取图像和文本特征，这些特征由 OpenCLIP 模型的图像和文本编码器生成。图像特征由预训练 ResNet-101 模型提取，确保高质量的视觉表示。文本全局特征由预训练 BERT 模型提取，以增强语义理解能力。随后，我们对提取的图像特征和文本特征进行拼接，构成跨模态特征向量。然后，使用 K-Means 聚类算法对跨模态特征向量进行聚类[9]，并以每个簇的均值作为初始的跨模态原型。这些原型被存储在共享的跨模态原型矩阵中，该矩阵充当视觉特征与文本特征之间的中间表示[10]。在模型训练过程中，跨模态原型矩阵不断更新和优化，确保模型能够动态调整跨模态特征的匹配精度。跨模态信息通过类别相关的原型查询与响应模块被明确嵌入到单模态特征中。此外，我们使用改进的多标签对比损失，增强模型的区分能力，提高其鲁棒性和预测精度。具体而言，图像特征和文本特征的拼接过程如下：

$$f_{cm} = \mathrm{Concat}\left(f_{im}\left(u\right),f_{text}\left(u\right)\right) \tag{16}$$

其中，$f_{im}\left(u\right)$代表图像特征，$f_{text}\left(u\right)$代表文本特征，Concat 表示拼接操作。使用 K-Means 聚类算法，我们对跨模态特征向量进行聚类，以生成跨模态原型：

$$\left\{m_k\right\}_{k=1}^{N} = \mathrm{KMeans}\left(R_k\right) \tag{17}$$

其中，m_k 表示第 k 个簇的聚类中心(Cluster Center)，R_k 代表第 k 个簇的跨模态特征集合，N 为总聚类数。每个簇的均值被用作初始跨模态原型，计算如下：

$$PM_{k,i} = \frac{1}{N_{k,i}} \sum_{j=1}^{N_{k,i}} r_{k,i,j} \tag{18}$$

其中，$PM_{k,i}$ 第 k 类的第 i 簇的原型，$N_{k,i}$ 表示第 k 类第 i 簇的样本数，$r_{k,i,j}$ 代表该簇中第 j 个样本的跨模态特征。为了进一步优化跨模态特征的学习过程，我们引入以下公式，对跨模态原型矩阵进行动态更新：

$$PM_{k,i}^{new} = \alpha \cdot PM_{k,i}^{old} + (1-\alpha) \cdot PM_{k,i}^{update} \tag{19}$$

在该公式中 $PM_{k,i}^{update}$ 表示跨模态原型矩阵中第 k 类第 i 簇的更新原型向量，该值在模型训练时动态计算，反映了类别特征中心的调整情况[11]。该更新机制的目的是迭代优化原型，使其更准确地反映每个类别的全局分布。更新过程可采用加权平均或优化策略进行实现，加权平均方式平衡历史原型和新批次特征的影响。优化策略可通过最小化损失函数来调整原型，使其更精确地与当前数据分布对齐。该机制在跨模态学习中至关重要，有助于提高跨模态特征匹配的准确性，同时增强类别可分性，从而提升 VQA 任务的整体性能[12]。应用于视觉问答任务通过这一迭代优化过程，跨模态原型矩阵被构建并用于后续的视觉问答任务。在 VQA 任务中，模型能够查询并选择最相关的原型，进而提升分类性能，提高答案预测的准确性。本研究的实验结果表明，该跨模态原型矩阵不仅有效地提升了跨模态特征融合能力，还在零样本 VQA 任务中展现了更强的泛化能力，使模型能够在未见类别上做出准确推理[13]。

1.5. 跨模态原型查询与响应模块(Cross-Modal Prototype Query and Response Module)

在跨模态原型查询与响应模块中，我们旨在选择和利用最相关的原型，以生成更准确的回答。该过程通过跨模态原型矩阵来促进模态间的信息流动，并将原型信息嵌入单一模态特征中，从而增强模型性能。这一方法不仅有助于缓解数据偏差问题，还能提升跨模态交互能力。首先，我们计算问题特征 f_Q 与原型矩阵中每个原型的余弦相似度，以衡量它们的相关性：

$$Sim(f_Q, PM_k) = \frac{f_Q \cdot PM_k}{\|f_Q\| \|PM_k\|} \tag{20}$$

其中，$Sim(f_Q, PM_k)$ 表示问题特征 f_Q 与第 k 个原型 PM_k 之间的相似度，$\|\cdot\|$ 代表 L2 范数。基于余弦相似度得分，我们选取相似度最高的前 M 个原型，并将其表示为 $\{PM_{topi}\}_{i=1}^{M}$。为了动态调整每个原型对问题的贡献，我们采用 Softmax 函数对这些相似度得分进行归一化，生成原型的权重 W_P：

$$W_P^{(i)} = \frac{\exp\left(\mathrm{Sim}\left(f_Q, PM_{topi}\right)\right)}{\sum_{j=1}^{M} \exp\left(\mathrm{Sim}\left(f_Q, PM_{topj}\right)\right)} \tag{21}$$

其中，$W_P^{(i)}$ 代表第 i 个选定原型的权重。Softmax 归一化保证了所有权重的总和为 1，从而突出最相关的原型。随后，我们将选定的原型输入全连接(Fully Connected, FC)层，以生成融合后的原型向量：

$$f_{proto} = FC\left(\{PM_{topi}\}_{i=1}^{M}\right) \tag{22}$$

其中，FC 代表全连接层操作，f_{proto} 为融合后的原型向量。最终的响应向量通过原型向量与权重 W_P 进行加权求和计算得到：

$$f_{response} = \sum_{i=1}^{M} W_P^{(i)} \cdot f_{proto}^{(i)} \tag{23}$$

其中，$f_{response}$ 为最终的响应向量，$f_{proto}^{(i)}$ 代表第 i 个原型经过线性变换后的向量。最终，$f_{response}$ 进一步与单模态特征融合，形成丰富的跨模态表征，使模型能够更准确地进行预测。通过权重 W_P 的动态调整，该模块能够有效优化原型贡献，从而提升多模态信息的整合能力，减轻数据偏差问题，提高视觉问答任务

的表现。

1.6. 答案生成模块

在答案生成模块中，我们首先将生成的响应向量 f_{response} 和问题特征 f_Q 共同输入基于马氏距离的改进多标签对比损失函数，以计算损失并优化模型。在保证计算效率的同时，我们选取初始得分最高的 500 个候选答案，以平衡计算资源与答案空间的复杂性。实验结果表明，这一阈值能够涵盖大部分正确答案，而不会显著影响模型性能。优化的损失函数定义如下：

$$L=-\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C}\left[y_{i,k}\cdot log\sigma\left(V_{\text{response}}\cdot f_Q\right)+\left(1-y_{i,k}\right)\cdot log\left(1-\sigma(V_{\text{response}}\cdot f_Q)\right)\right] \tag{24}$$

其中，$y_i=\left[y_{i,1},y_{i,2},\cdots,y_{i,C}\right]$ 代表样本 i 的多标签真实值，$V_{\text{response}}\cdot f_Q$ 为响应向量与问题特征的内积，$V_{\text{response}}\cdot f_Q$ 为 sigmoid 激活函数。接下来，我们计算每个候选答案的概率分布：

$$P\left(a_j\right)=\frac{\exp\left(\left(V_{\text{response}}\cdot f_Q\right)\right)}{\sum_{k=1}^{500}\exp\left(\left(V_{\text{response}}\cdot f_Q\right)\right)} \tag{25}$$

其中，$P\left(a_j\right)$ 表示第 j 个候选答案的概率。最终，我们选择概率最高的候选答案作为最终答案输出。这一方法确保了计算资源聚焦于最相关的候选答案，同时通过马氏距离优化损失函数，从而提升预测的准确性。在零样本视觉问任务中，传统的对比损失函数面临诸多挑战。由于这些方法主要针对单标签任务设计，难以处理高维视觉特征和文本特征之间复杂的多标签关系[14]。此外，传统方法依赖精确的模态对齐，在涉及未见类别时表现不佳，且在捕捉深层跨模态交互方面存在局限性。为了解决这些问题，我们提出了一种基于马氏距离的多标签对比损失。该方法利用协方差矩阵建模跨模态的语义依赖关系，以更精准地捕获模态间的深层次特征交互。对于正样本(共享标签的样本对)，损失函数最小化马氏距离，以增强其特征表示的相似性：

$$L_{\text{pos}}=\frac{1}{N_{\text{pos}}}\sum_{(i,j)\in\text{Pos}}\left(F_i-F_j\right)^{\top}S^{-1}\left(F_i-F_j\right) \tag{26}$$

其中，F_i 和 F_j 为样本特征表示，S 为协方差矩阵。对于负样本(无共享标签的样本对)，引入间隔参数 m 确保特征表示具有足够的区分度：

$$L_{neg}=\frac{1}{N_{\text{neg}}}\sum_{(i,j)\in\text{Neg}}\max\left(0,m-\left(F_i-F_j\right)^{\top}S^{-1}\left(F_i-F_j\right)\right) \tag{27}$$

最终损失函数结合正负样本的损失项：

$$L_{\text{contrastive}}=\lambda_{\text{pos}}L_{\text{pos}}+\lambda_{\text{neg}}L_{\text{neg}} \tag{28}$$

其中，λ_{pos} 和 λ_{neg} 控制正负损失的相对权重。相比于基于欧几里得距离的方法，该方法能够更有效地建模高维特征空间中的复杂类别关系，从而提高零样本视觉问答任务的泛化能力。马氏距离的引入使得模型能够学习到更丰富的跨模态特征表达，尤其适用于零样本视觉问答任务中的未见类别问题[15]。

2. 实验结果与分析

在本节中，我们围绕 TZSL 和 GZSL 两种设置开展了系统实验，重点分析了模型在 F-VQA 和 ZS-F-VQA 数据集上的性能表现，以全面评估其在零样本视觉问答任务中的适应性与有效性。首先，通过与当前主流方法的对比实验，验证了所提框架在多个评估指标上的领先优势。其次，开展消融实验，系统评估各关键组件对整体性能的影响，凸显模块设计的合理性与必要性。最后，结合典型案例的可视化分析，从直观层面展示模型在泛化能力和语义对齐方面的优势。

2.1. 定量结果与分析

如表 1 实验结果所示，我们的框架在多个基准方法(包括 HieQ + I、MLP、Up-Down、SAN、Hie-Q + I + Pre 和 BAN)上均取得了显著优势。具体而言，我们的方法在 HIT@K 指标上全面超越现有方法，在 HIT@1 达到 60.43%，HIT@3 达到 81.69%，HIT@10 进一步提升至 89.96%，这表明我们的模型能够有效提取有意义的表示，以实现精准的视觉问答任务。

Table 1. Overall performance on the standard F-VQA dataset (TOP-500), reported as Hit@K percentages. † indicates that the model employs a mapping-based configuration, where the answer prediction is performed by directly computing the similarity between the fused feature representation and candidate answers, rather than using a traditional classifier layer (%)
表 1. 以标准 F-VQA 数据集(TOP-500)上的整体性能，以 Hit@K 百分比的形式报告。†表示该模型采用的是基于映射(mapping-based)的配置方式，其中答案的预测是通过直接计算融合特征表示与候选答案之间的相似度来完成的，而不是使用传统的分类器层(%)

Method	HIT@1	HIT@3	HIT@10	MRR	MR
Hie-Q+I	33.70	50.00	64.08	-	-
MLP	34.12	52.26	69.11	-	-
Up-Down	34.81	50.13	64.08	-	-
Up-Down †	40.91	57.47	72.74	-	-
SAN	41.69	58.17	72.69	-	-
Hie-Q+I+Pre	43.14	59.92	71.34	-	-
BAN	44.02	58.29	70.66	-	-
BAN†	45.95	63.36	78.12	-	-
MLP†	47.55	66.76	81.55	-	-
SAN†	49.27	67.30	81.79	0.605	14.75
Ours	**60.43**	**81.69**	**89.96**	**0.708**	**12.74**

进一步地，如表 2 所示，我们在 ZS-F-VQA 数据集的 GZSL (广义零样本学习)和 TZSL (传统零样本学习)设定下，我们的模型在所有指标上均取得最高分数，特别是在 HIT@K 和 MRR (平均倒数排名)指标上相较于基线方法有显著提升。这些结果表明，我们的方法在零样本视觉问答任务中具有优越的泛化能力，能够更有效地适应未见类别的推理需求。

Table 2. Overall results on the ZS-F-VQA dataset under TZSL/GZSL settings
表 2. 在 TZSL/GZSL 设定下 ZS-F-VQA 数据集(%)上的整体实验结果表

Method	GZSL					TZSL				
	HIT@1	HIT@5	HIT@10	MRR	MR	HIT@1	HIT@5	HIT@10	MRR	MR
Up-Down †	0.00	2.67	16.48	-	-	13.88	25.87	45.15	-	-
BAN†	0.22	4.18	18.55	-	-	13.14	26.92	46.90	-	-
MLP†	0.07	4.07	27.40	-	-	18.14	37.85	59.88	-	-
SAN†	0.11	6.27	31.66	0.093	48.18	20.41	37.20	62.24	0.38	19.14
Ours	34.45	57.68	77.81	0.378	16.16	38.94	73.87	88.60	0.48	6.05

在 ImNet-A 和 ImNet-O 数据集上的评估进一步验证了我们模型的强大性能。如表 3 和表 4 所示，在 ImNet-A 数据集中，我们的模型在 HIT@1 和 HIT@5 上分别达到了 66.93%和 88.94%，展现出其在识别正确答案上的精准性和鲁棒性。同时，模型在 MRR (均值倒数排名)上达到了 0.87，MR (平均排名)为 22.54，反映了其在排序相关答案方面的高效性，进一步凸显了模型在跨模态交互理解中的优势。在 ImNet-O 数据集上，类似的性能优势同样明显，HIT@1 达到 53.21%，并在 HIT@3 和 HIT@10 维度上保持稳定提升。

Table 3. Overall results on ImNet-A dataset (%)
表 3. ImNet-A 数据集上的整体(%)实验结果表

Method	HIT@1	HIT@3	HIT@10	MRR	MR
DGP	45.30	56.20	70.50	0.64	36.97
KG-GAN	51.47	59.52	71.20	0.52	29.98
Ours	**66.93**	**75.82**	**88.94**	**0.87**	**22.54**

Table 4. Overall results on ImNet-O dataset (%)
表 4. ImNet-O 数据集上的整体(%)实验结果表

Method	HIT@1	HIT@3	HIT@10	MRR	MR
DGP	41.50	52.30	66.20	0.59	47.93
KG-GAN	44.92	57.13	68.50	0.48	33.56
Ours	**53.21**	**68.90**	**82.05**	**0.71**	**29.55**

这些实验结果表明，我们的跨模态原型学习框架和基于马氏距离优化的方法在零样本 VQA 任务中至关重要。通过高效融合视觉和文本模态，我们的模型能够在未见场景中实现高效泛化，在答案预测的鲁棒性和精确排名方面树立了新标杆。

2.2. 消融研究

在消融实验中，我们以完整模型框架作为基线，评估了不同组件对整体性能的贡献。如表 5 所示，用 OpenCLIP 替换标准特征提取器显著提高了图像和文本的特征提取能力，而当去除 OpenCLIP 时，HIT@1 下降至 44.76%。此外，去除聚类方法导致 HIT@10 急剧下降至 68.33%，表明该方法在增强类别区分能力方面具有关键作用。同样，移除基于马氏距离的损失函数使得 HIT@1 降至 44.24%，HIT@10 降至 70.99%，强调了该损失在泛化至未见类别时的重要性。最终，当所有核心组件都被去除时，HIT@1 下降至 21.04%，这凸显了聚类机制、马氏损失和高级特征提取的协同作用对于零样本 VQA 任务的至关重要性。

Table 5. Ablation results (%) evaluating the impact of different answer embedding methods on model performance on the standard F-VQA dataset (TOP-500)
表 5. 不同答案嵌入方式对标准 F-VQA 数据集(TOP-500)上模型性能消融研究(%)实验结果表

Method	HIT@1	HIT@3	HIT@10
Ours (w/o OPENCLIP)	**44.76**	46.09	60.80
Ours (w/o MAHA)	44.24	**66.74**	**70.99**
Ours (w/o clustering)	40.82	55.78	68.33
Ours (w/o all)	21.04	27.82	30.67

1) 聚类后的原型影响分析

如实验结果如图 2 所示，每个聚类点对应一个特定类别(如“狗”、“猫”)，每个类别的中心(黑色小圆点)表示该类别的原型。这些原型在零样本场景下尤为关键，因为它们能够作为该类别典型特征的泛化表示。原型周围数据点的分布展示了该类别的多样性，而数据点与原型的接近程度表明了模型在捕获类别核心特征方面的能力。ProMahaVQA 利用原型来动态适应新数据，而无需进行大规模的额外训练，从而提高泛化性，尤其是在零样本任务中的表现。

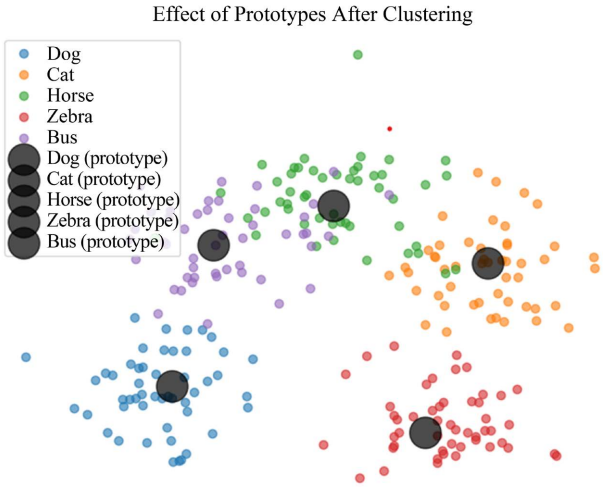

Figure 2. Cross-modal feature clustering and prototype representation
图 2. 跨模态特征聚类与原型表示

2) 支持实体与关系的重要性

支持实体和关系的数量显著影响了模型的性能，特别是在零样本学习环境下。支持实体提供来自外部知识库的重要背景信息，帮助模型建立视觉内容与文本问题之间的联系；而关系则表示这些实体之间的语义关联，对 VQA 任务的推理至关重要。如实验结果如图 3 所示，适量增加支持实体能够提供更丰富的上下文信息，提高模型处理复杂问题的能力。然而，过多的支持实体可能会引入噪声，导致性能下降或稳定在一定水平。同样，适量增加关系数量有助于提升模型的推理能力，但过多的关系会降低模型的聚焦能力，增加不必要的复杂度。因此，在支持实体和关系数量之间取得平衡至关重要。

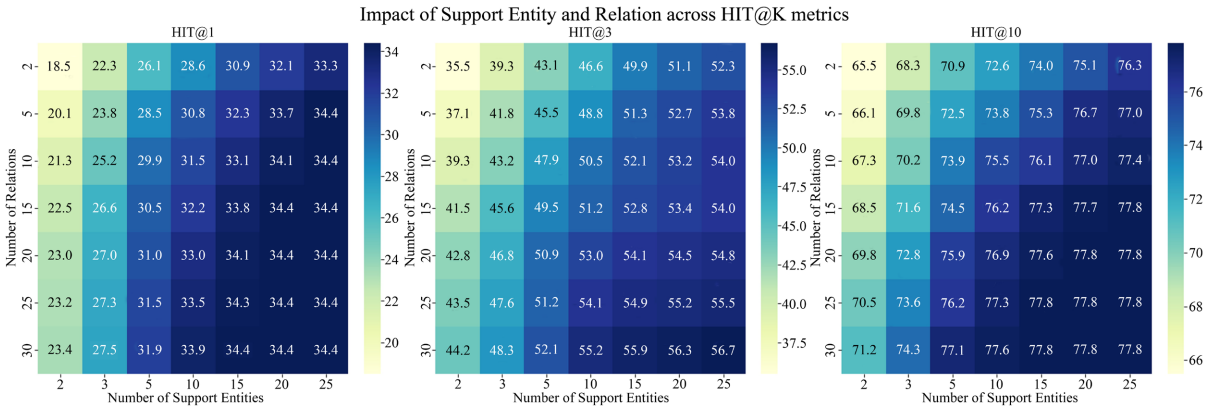

Figure 3. Impact of support entity and relation across HIT@K metrics
图 3. 支持实体与关系数量对各项 HIT@K 指标的影响

3) 马氏距离的影响

如实验结果如图 4 所示，马氏距离缩放因子(Scaling Factor)在决定模型泛化能力和类别区分能力方面起着关键作用，尤其是在零样本学习场景下。具体而言，该因子与对比损失中的参数 λ_{pos} 和 λ_{neg} 相关：

$$L_{contrastive} = \lambda_{pos} \sum_{(i,j)\in P} d_{Maha}\left(z_i, z_j\right) - \lambda_{neg} \sum_{(i,j)\in N} d_{Maha}\left(z_i, z_j\right) \tag{29}$$

实验结果显示，适当调整缩放因子可显著改善模型性能。例如，当缩放因子设为 2.0 时，模型的 HIT@1 达到 34.0%，HIT@3 提升至 57.6%，HIT@10 最高达到 77.8%，表明此时模型在类别区分与泛化能力之间达到了最佳平衡。然而，当缩放因子超过 2.0 后，模型的性能开始下降，表明过大的缩放因子可能会导致模型对类别间微小差异过于敏感，从而影响泛化能力。

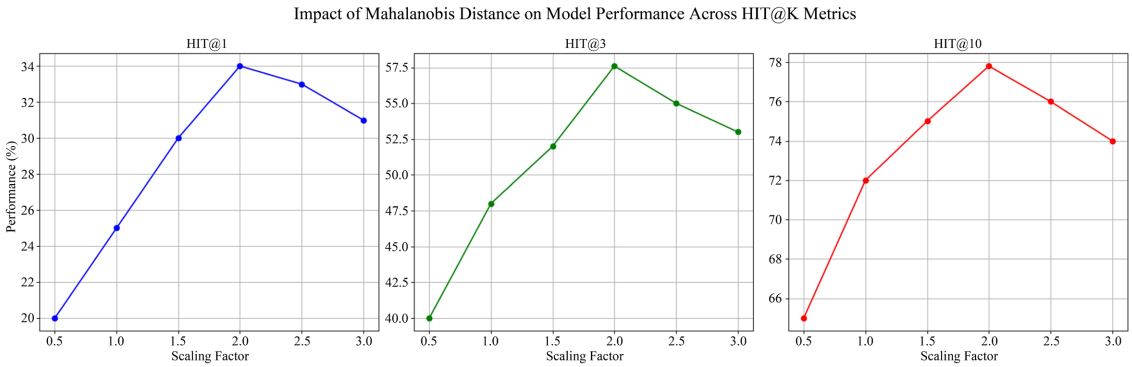

Figure 4. Impact of Mahalanobis distance on model performance across HIT@K metrics
图 4. 马氏距离对模型在各项 HIT@K 指标上性能的影响

4) 原型数量、特征维度和聚类距离的影响

如实验结果如图 5 所示，随着原型数量和特征维度的增加，模型性能显著提升，特别是在较高维度(如 2560 维)时，表明更丰富的原型表征有助于更精确地捕获类别内和类别间的差异，提高泛化能力。然而，当原型数量超过一定阈值时，性能提升趋于平稳，表明增加更多原型已无法进一步提升模型的区分能力。此外，合理的聚类距离对于保持模型的类内一致性和类间可分性至关重要，过少的聚类数可能导致类别区分不清，而过多的聚类数则会引入噪声，导致过拟合。因此，合理调整聚类参数能够有效提升跨模态相似性建模能力，使模型在零样本 VQA 任务中表现更优。

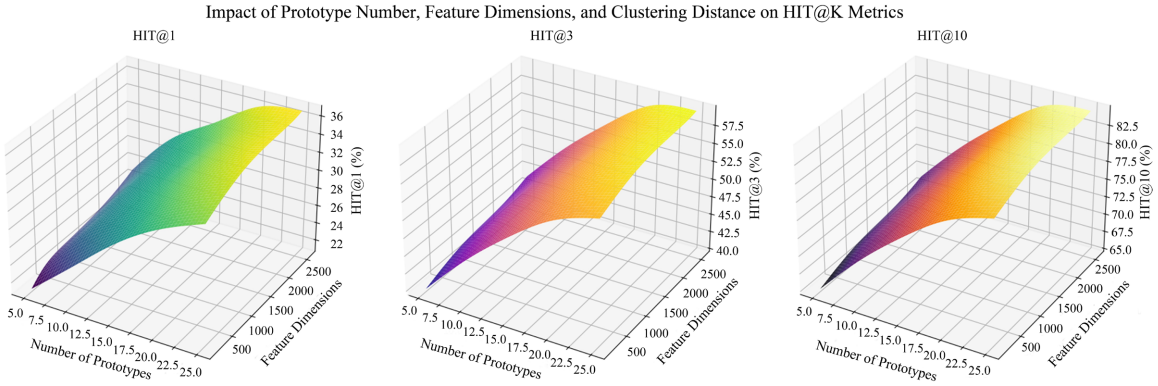

Figure 5. Impact of prototype number, prototype feature dimensions, and clustering distance on model performance across HIT@K metrics
图 5. 原型数量、原型特征维度及聚类距离对模型在各项 HIT@K 指标上性能的影响

5) 跨模态相似度阈值的影响

如实验结果如图 6 所示，跨模态相似度阈值直接影响图像－文本匹配和答案预测的准确性。适当提高相似度阈值可以提升模型的区分能力，使其更有效地筛选出相关的图像－文本对。然而，阈值过高会导致匹配标准过于严格，可能错过部分相关对，影响模型性能。实验结果表明，适中的相似度阈值能够实现最佳平衡，在 HIT@1、HIT@3 和 HIT@10 方面取得指标上取得最优平衡，显著提升模型的推理准确率与泛化能力。

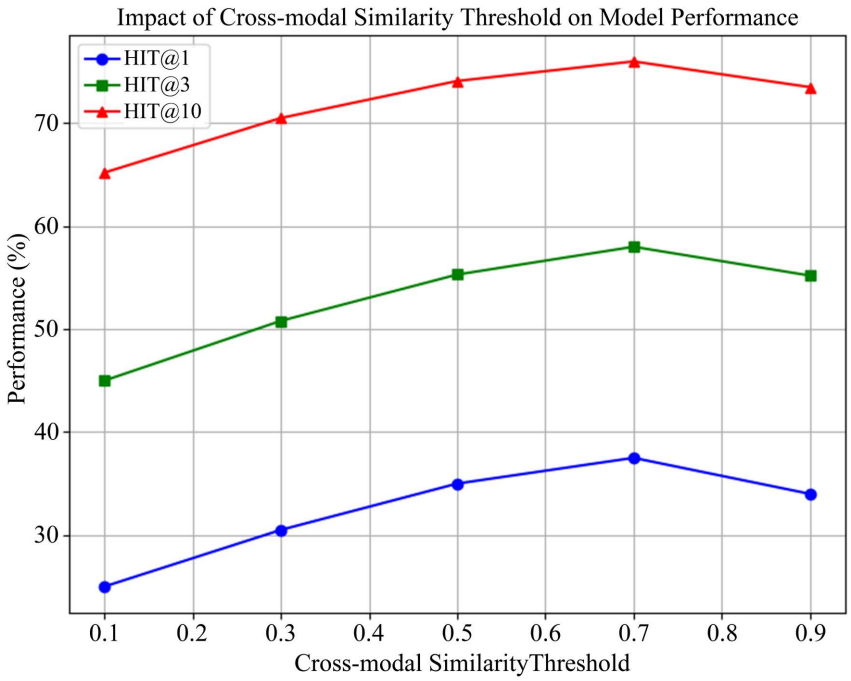

Figure 6. Impact of cross-modal similarity threshold on model performance across HIT@K metrics
图 6. 跨模态相似度阈值对模型在各项 HIT@K 指标上性能的影响

2.3. 可视化对比分析

本研究针对 GZSL (广义零样本学习)和 ZSL (零样本学习)场景进行了案例分析，进一步验证了 ProMahaVQA 模型在复杂环境下的表现。如实验结果如图 7 所示，实验涵盖多种问题类型，包括科技物品识别、类别分类和功能推理(如“这张图片中哪种水果是柑橘类？”或“我们可以用图片中的球拍做什么？”)。结果表明，在 ZSL 场景下，ProMahaVQA 能够准确识别未见类别。例如，在水果类别相关的问题中，模型不仅正确预测出“柠檬”(标准答案)，还合理地将“橙子”和“葡萄柚”等相似类别作为备选答案。在 GZSL 场景下，模型能够有效平衡已见与未见类别，避免对已见类别的偏向。例如，在“图片中哪个物体属于建筑类别？”的问题中，模型准确输出“建筑”并给出合理候选项。相比之下，传统的 SAN†模型表现出较明显的已见类别偏好。例如，在“什么是科技物品？”的问题中，SAN†更倾向于预测“笔记本电脑”和“平板电脑”，而未能识别出图像中的关键物体“手机”。同样，在功能推理问题“我们可以用图片中的球拍做什么？”中，SAN†给出的答案如“练习体育运动”过于笼统，难以与标准答案对齐。ProMahaVQA 通过引入跨模态原型学习与马氏距离优化，在识别未见类别方面展现出更强的表现，显著提升了答案排序的准确性。在 ZSL 和 GZSL 两类任务中均能稳定对齐标准答案，充分体现了其在跨模态语义融合与零样本推理方面的优势。

Question	What is the technology stuff?	Which object in this image belongs to the category of Constructions?	Which fruit in this image is a citrus?
Image			
Our Model	**Cell phone**, Camera, Screen	**Building**, Fence, Farmhouse	**Lemon**, Orange, Grapefruit
SAN†	Laptop, Remote control, Tablet	Roof, Tractor, Shed	Pear, Orange, Strawberry
Ground Truth	**Cell phone**	**Building**	**Lemon**
Question	What can we use the racket in the image for?	What eight-sided object is visible in this image?	What object is used for sitting?
Image			
Our Model	**Playing tennis**, Hitting balls, Sports	**Octagon**, Stop sign, Table	**Chair**, Stool, Bench
SAN†	Practicing sports, Playing badminton, Swinging	Board, Street sign, Warning sign	Couch, Stool, Floor
Ground Truth	**Tennis**	**Stop sign**	**Chair**

Figure 7. Cases under GZSLVQA (Up) and generalized VQA (Down) setting
图 7. GZSL 视觉问答(上)与广义视觉问答(下)设定下的案例展示

3. 总结

本章围绕零样本视觉问答中的核心挑战，提出并系统论证了基于原型学习与马氏距离对比损失优化的 ProMahaVQA 模型。首先，从现有方法的不足出发，明确指出传统 VQA 模型在泛化能力、语义对齐及多标签处理方面的局限。随后，详细介绍了 ProMahaVQA 的整体架构，包括跨模态原型矩阵构建、查询响应机制以及马氏距离驱动的多标签对比学习策略，展示了其在未见类别识别和跨模态特征对齐方面的优势。在实验部分，本文在多个主流零样本 VQA 数据集(F-VQA、ZS-F-VQA、ImNet-A、ImNet-O)上进行了系统的定量评估与对比实验，结果显示本模型在 HIT@K、MRR 和 MR 等指标上均显著优于现有方法。通过消融实验验证了各模块在性能提升中的关键作用，并进一步通过案例可视化分析，展示了模型在 ZSL 与 GZSL 场景下的推理能力和语义泛化能力。综上所述，ProMahaVQA 有效缓解了零样本 VQA 中存在的类别偏差、模态不对齐与泛化能力弱等问题，为多模态推理任务提供了更具适应性与解释性的解决方案。下一章将继续扩展该模型在医学 VQA 领域的应用，并探讨其在复杂场景下的可迁移性与实用价值。

参考文献

[1] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., *et al*. (2015) VQA: Visual Question Answering. 2015 *IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 2425-2433. https://doi.org/10.1109/iccv.2015.279

[2] Lu, J., Yang, J., Batra, D., *et al*. (2016) Hierarchical Question-Image Co-Attention for Visual Question Answering. *Advances in Neural Information Processing Systems*, 29.

[3] Chen, Z., Chen, J., Geng, Y., Pan, J.Z., Yuan, Z. and Chen, H. (2021) Zero-Shot Visual Question Answering Using Knowledge Graph. In: Hotho, A., *et al*., Eds., *Lecture Notes in Computer Science*, Springer International Publishing, 146-162. https://doi.org/10.1007/978-3-030-88361-4_9

[4] Zhang, X., Wu, C., Zhao, Z., *et al*. (2023) PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering. arXiv:2305.10415.

[5] Abacha, A.B., Shivade, C., Hasan, S.A., *et al*. (2019) VQA-Med: Overview of the Medical Visual Question Answering Task at Image CLEF 2019. CEUR 2019 *Working Notes*, Lugano, 9-12 September 2019, 9-12.

[6] Jin, D., Pan, E., Oufattole, N., Weng, W., Fang, H. and Szolovits, P. (2021) What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences*, **11**, Article 6421. https://doi.org/10.3390/app11146421

[7] Dao, S.D., Zhao, E., Phung, D., *et al*. (2021) Multi-Label Image Classification with Contrastive Learning. arXiv:2107.11626.

[8] Sahoo, S. and Maiti, J. (2025) Variance-Adjusted Cosine Distance as Similarity Metric. arXiv:2502.02233.

[9] Xian, Y., Lampert, C.H., Schiele, B. and Akata, Z. (2019) Zero-Shot Learning—A Comprehensive Evaluation of the Good, the Bad and the Ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **41**, 2251-2265. https://doi.org/10.1109/tpami.2018.2857768

[10] Liu, H. and Singh, P. (2004) ConceptNet—A Practical Commonsense Reasoning Tool-Kit. *BT Technology Journal*, **22**, 211-226. https://doi.org/10.1023/b:bttj.0000047600.45421.6d

[11] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., *et al*. (2015) Dbpedia—A Large-Scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web*, **6**, 167-195. https://doi.org/10.3233/sw-140134

[12] Yang, Z., He, X., Gao, J., Deng, L. and Smola, A. (2016) Stacked Attention Networks for Image Question Answering. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 21-29. https://doi.org/10.1109/cvpr.2016.10

[13] Kim, J.H., Jun, J. and Zhang, B.T. (2018) Bilinear Attention Networks. arXiv:1805.07932.

[14] Snell, J., Swersky, K. and Zemel, R.S. (2017) Prototypical Networks for Few-Shot Learning. *Advances in Neural Information Processing Systems*, 30.

[15] Zhu, L., She, Q., Chen, Q., Meng, X., Geng, M., Jin, L., *et al*. (2023) Background-Aware Classification Activation Map for Weakly Supervised Object Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 14175-14191. https://doi.org/10.1109/tpami.2023.3309621