基于探针引导的视觉语言多模态解释方法

索 伟[1)] 吕家齐[1)] 孙梦阳[2)] 刘 乐[1)] 王 鹏[1)]

[1)](西北工业大学计算机学院 西安 710129)

[2)](西北工业大学网络空间安全学院 西安 710129)

摘 要 现有的视觉语言模型大部分通过具有“黑盒”结构的深度神经网络实现跨模态推理,然而网络内部的执行过程难以被人类直观理解。因此,本文侧重于研究面向视觉问答(Visual Question Answering,VQA)任务的自然语言解释(Natural Language Explanation,NLE)方法,旨在通过生成的自然语言语句来解释模型的推理过程。虽然现有方法已经取得了一定的进展,但仍面临以下挑战:(1)答案的预测过程和解释的生成过程相互干扰,弱化了解释的忠实性。(2)现有的方法仅能生成单模态的解释,存在由于指代模糊导致的语义歧义问题。为此,我们新提出了一种面向视觉问答推理过程的多模态解释方法(Probe-based Multi-modal Explanation method,PME),该方法能从推理过程的每个隐藏层状态提取信息且不影响原推理路径,确保了解释过程对原有推理过程的忠实性。另外,我们使用伪标签方法融合了VQA-X数据集与GQA数据集,在保证忠实性的前提下实现了多模态解释,缓解了单一模态文本解释中对目标的指代语义模糊问题。本文在视觉问答数据集VQA-X和A-OKVQA上将PME和其他最新最优的模型进行了性能比较,实验结果表明PME方法在相应测试集上获得了更高的解释评估分数。我们期待我们工作能够为网络模型的内部理解提供一个新的研究基础。代码位于:https://github.com/LouisJacky/LAVIS_PME。

关键词 视觉问答;自然语言解释;跨模态推理;伪标签;预训练模型

中图法分类号 TP182; **DOI号** 10.11897/SP.J.1016.2025.01478

Probe-Based Multi-Modal Explanation Method for Visual Question Answering

SUO Wei[1)] LV Jia-Qi[1)] SUN Meng-Yang[2)] LIU Le[1)] WANG Peng[1)]

[1)](School of Computer Science, Northwestern Polytechnical University, Xi'an 710129)

[2)](School of Cybersecurity, Northwestern Polytechnical University, Xi'an 710129)

Abstract Recently, visual-language models have achieved remarkable performance on various complex tasks such as Visual Question Answering (VQA), Image Captioning, and Referring Expression Comprehension (REC). However, these models mostly adopt a "black-box" structure with deep neural networks, making their inference processes difficult to understand intuitively. To explore the internal mechanisms of neural networks based models, researchers have proposed various interpretation methods, including gradient activation, saliency maps, and natural language explanations. Among these approaches, Natural Language Explanation (NLE) has gained significant attention in the visual-language community, particularly for VQA tasks (VQA-NLE), as it provides human-interpretable explanations that help understand model reasoning and develop

收稿日期:2024-07-01;在线发布日期:2025-03-19。本课题得到国家自然科学基金面上项目(No.62472357)和青年科学基金项目(C类)[原青年科学基金项目](No.62102323)资助。**索 伟**,博士研究生,中国计算机学会(CCF)会员,主要研究领域为计算机视觉和视觉语言理解。E-mail:suowei1994@mail.nwpu.edu.cn。**吕家齐**,硕士研究生,主要研究领域为计算机视觉。**孙梦阳**,博士研究生,中国计算机学会(CCF)会员,主要研究领域为计算机视觉和跨模态对齐。**刘 乐**(通信作者),博士,副教授,中国计算机学会(CCF)会员,主要研究领域为计算机视觉和智能信息处理。E-mail:lel@nwpu.edu.cn。**王 鹏**,博士,教授,长江学者,中国计算机学会(CCF)会员,主要研究领域为计算机视觉、自然语言处理和机器学习。

more trustworthy deep learning systems. Existing VQA-NLE approaches can be categorized into two paradigms: post-hoc explanation methods and self-rationalization methods. Post- hoc methods typically employ separate decision and explanation models, where the decision model first predicts answer, and then the explanation model uses the predicted answer along with question to infer the reasoning process. However, this sequential and separate processing leads to a lack of direct logical connection between explanation and reasoning processes, compromising explanation reliability. In contrast, self-rationalization methods unify answer prediction and explanation generation into a single task. Despite offering simpler implementation and more reliable logical relationship modeling, such methods inevitably allow mutual interference between answer generation and explanation processes, weakening explanation faithfulness. Through extensive experiments on the VQA-X dataset, we validate such interference - with 8.38% of test samples showing different answer predictions when explanation loss is introduced during joint training, thus weakening the faithfulness of explanations. To address these challenges, we propose a novel Probe-based Multi-modal Explanation method (PME). Our method introduces a learnable probe structure that can extract information from each encoder hidden layer state during forward propagation without interfering with the original reasoning process. Experimental results confirm that our probe- based approach maintains identical answer predictions as the original model while generating faithful explanations. Additionally, to address the semantic ambiguity inherent in single-modal text explanations, we develop a cross-generation pseudo-labeling approach that enables simultaneous generation of natural language explanations and object detection boxes. This multi-modal approach significantly improves explanation clarity by providing explicit visual grounding for textual references, improving clarity scores from 67.5% to 79.2% on VQA-X and from 58.9% to 66.2% on A-OKVQA datasets. Extensive experiments demonstrate that our PME method outperforms state-of-the-art models on both VQA-X and A-OKVQA benchmarks, achieving 5.5% and 2.1% improvements in CIDEr scores respectively compared to the previous best method S3C. Both quantitative and qualitative analyses validate that our method more accurately reflects the model′s reasoning process while maintaining explanation faithfulness. As a model-agnostic strategy, our approach can be readily applied to other visual-language models, providing a new paradigm for developing more reliable and comprehensive model explanation methods. Code is available at:https://github.com/LouisJacky/LAVIS_PME.

Keywords　visual question answering; natural language explanation; cross-modal reasoning; pseudo labeling; pre-trained models

1　引　言

随着深度神经网络的发展，视觉语言大模型在各种复杂任务上已经取得了良好的性能，如视觉问答（Visual Question Answering，VQA）、图像描述（Image Captioning）和指代理解（Referring Expression Comprehension，REC）[1-4]。然而，这些模型大多采用具有“黑盒”结构的深度网络去执行复杂计算，其推理的过程难以被人类直观理解。为了探究深度网络的运行机制和相应的内部机理，研究者们提出了一系列模型解释方法，如梯度激活[5]、显著性图[6]和生成式的自然语言解释[7-8]等。其中，由于基于文本描述的解释方法（Natural Language Explanation，NLE）能够方便地被人类所理解，因此在视觉语言社区引起了广泛的关注，特别是对于视觉问答任务的解释（VQA-NLE）[7-12]。在该任务中，模型需要通过生成自然语言来解释决策模型在回答一个视觉问题时的推理过程，针对该问题的研究不仅有助于我们进一步探索推理网络的执行规则，也

有助于研究人员开发更加可信的深度模型。

现有的针对VQA-NLE的研究主要包括两种范式，一类采用后验解释方法(Post-hoc Explanation Method)[10-13]，另一类采用自解释方法(Self-Rationalization Method)。前者通常包括一个视觉语言决策模型和一个解释模型。决策模型用来预测给定问题的答案，而解释模型用来推断该决策的推理过程。在这里，模型的推理和解释过程被分割成两个独立的部分，导致解释的生成过程和决策模型的推理过程并没有建立直接的逻辑联系，从而使得生成的解释缺乏可靠性。相反，另一类自解释的VQA-NLE模型[7-8,14-15]，该方法将答案的预测和文本解释转化为一个统一的生成任务，使得决策模型在预测答案的同时也产生相应的解释。由于具有更加简单的实现过程和更加可靠的逻辑关系建模，该范式成为一种主流的模型解释手段而被广泛研究。

虽然上述两种方法已经极大地推动了VQA-NLE的发展，但是现有的这些技术路线仍面临以下挑战：(1)如图1(a)所示，使用自解释方法的模型[7-8,14-15]虽然能够同时生成答案和相应的解释，缓解了解释的生成过程和决策过程之间缺乏直接逻辑联系建模的问题，但此类方法不可避免地会导致答案的生成过程和解释过程相互干扰，从而弱化了其解释的忠实性。本文将在4.5节中对这一问题进行更加充分的讨论和说明。(2)现有的VQA-NLE模型只生成单一模态的文本解释而缺乏与图像之间的明确关联，使得解释由于指代模糊而存在语义歧义的问题。如图1(a)所示，仅进行文本解释“它包含太多的卡路里”将遭受严重的指代歧义，因为图中有多个食物，无法明确知道模型的推理来源。同时，通过更加详细的语言标注来解决语义模糊问题将会产生额外的标注成本。相反，一图胜千言，当补充以视觉解释后，我们不仅清楚地知道模型预测答案“否”的文本依据，也能了解模型是通过图像中的“薯条”部分进行的预测。本文将在4.6中对这一问题进行更加充分的讨论和说明。

为解决上述挑战，我们提出了一种新的基于探针引导的视觉语言多模态解释模型(Probe-based Multi-modal Explanation method，PME)。不同于之前的方法，PME在不干扰模型推理过程的情况下可以同时生成两种模态的解释，即自然语言解释和相应的目标检测框，确保了解释过程对原有推理过程的可靠性和忠实性，同时缓解了文本解释中目标

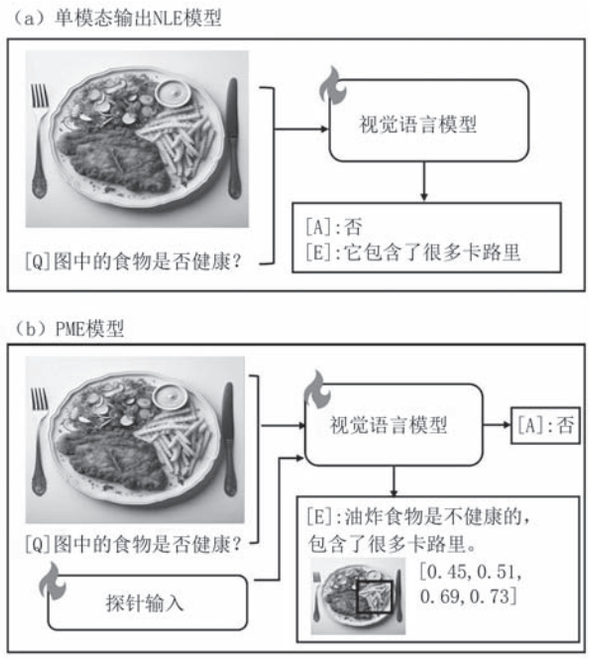


图1 现有技术路线面临的挑战((a)在使用自解释方法的模型中，视觉语言模型被整体微调并用于同时生成答案和文本解释，由于解释过程和推理过程相互影响，从而弱化了解释的忠实性。同时单一模态解释由于指代不明导致了语义歧义。(b)本文提出的基于探针引导的视觉语言多模态解释方法，通过引入探针结构捕获模型推理的隐状态信息，从而不干扰其推理过程。此外我们的方法可以同时生成基于视觉和文本的解释，直观地指出与决策相关的“薯条”图像区域，有效地缓解了指代语义歧义的问题。)

指代模糊的问题，使得解释表述更清楚直观且易于理解。具体而言，为了实现面向视觉问答的自然语言解释任务，PME使用[16-19]结构模块化程度高的通用视觉语言模型BLIP2[20]作为主要研究对象。事实上，作为一个模型不可知的策略，本文所提方法也能被方便地应用于其他视觉语言模型，我们在本文的4.2节中对该问题进行了详细的讨论。在本方法的具体实现中，为了不影响原模型的推理过程和获得更加忠实的解释，我们设计了一种可学习的探针结构，该方法可以方便地从模型推理的前向传播过程中提取每个编码器隐藏层的信息状态。伴随着新引入的注意力掩码技术，该策略确保了解释和推理过程的互不干扰，从而使得生成的解释完全忠实于原推理过程。另一方面，为了缓解单一模态解释存在的语义模糊问题，我们提出了一种新的多模态解释生成方法，该方法利用交叉生成的伪标签技术将目标检测框嵌入到生成的语言解释中，建立了文本解

释与图像目标对象的对应关系，从而弥合了单模态解释中固有的语义歧义鸿沟。经过大量的实验验证，本文提出的PME模型在VQA-X[11]和A-OKVQA[21]数据集上相较于最先进的方法S3C[8]分别提高了5.5%和2.1%的CIDEr分数，同时，本文通过定量研究和定性研究分析了多模态解释的优势，为后续进一步的研究建立了基础。

本文主要贡献总结如下：

(1)针对现有VQA-NLE模型的解释过程和原有推理过程相互干扰的问题，本文提出了一种新的基于探针引导的视觉语言多模态解释方法，并通过引入一种新颖的探针结构，提高了模型解释对原有推理过程的忠实性和可靠性，为VQA-NLE模型提供了新范式。

(2)对于单一模态的文本解释遭受的指代语义模糊，我们引入了一种新的多模态解释方法，伴随着交叉生成的伪标签技术，本文所提方法实现了自然语言描述解释和图像目标对象检测框的同步生成，确保了解释过程的语义明确性和合理性。

(3)受益于上述策略，我们的方法在VQA-X和A-OKVQA数据集上均达到了最优异的性能。同时大量定性和定量的实验结果表明，我们的方法更加准确地反映了模型的推理过程。

本文第2节简要介绍视觉语言模型和自然语言解释任务的相关工作；第3节具体描述了本文提出的方法的关键实现技术；第4节提供本文的实验描述、实验结果和结果分析；第5节和第6节总结了本文的工作，分析了本工作的局限性和未来工作的方向。

2　相关工作

2.1　视觉语言模型

视觉语言模型是一种结合了视觉和语言模态的模型，很大程度上受到语言模型发展的影响。早期的语言模型工作主要基于卷积和循环神经网络[22-24]，而随着Transformers[25]的引入，语言模型的大量工作[26-29]取得了突破，提高了少样本学习能力、迁移学习能力、文本生成能力以及复杂序列处理能力，进而促进了视觉语言模型的发展。受到语言模型BERT[27]启发的大量的视觉语言工作在图像文本特征对齐方面进行了研究[30-32]，显著提高了视觉语言模型在需要复杂推理任务中的表现[33-38]，例如视觉问答、视觉常识推理。最近的一个热门研究方向是将大语言模型扩展到多模态的大视觉语言模型[39-44]，使用文本生成任务解决复杂视觉语言问题，例如图像描述和视觉问答。Alayrac等人[45]提出了一种名为Flamingo的视觉语言模型，将视觉适应层(例如感知机)整合到大语言模型结构中，在大规模多模态语料库上进行训练，实现了同时处理混合的文本、图像、视频输入。OpenFlamingo[46]复现了Flamingo[45]的性能并开源了代码。随后，MM-GPT[47]和Otter[48]基于OpenFlamingo[46]的架构，将指令调优方法引入到多模态模型的微调中，实现了更友好的用户交互。另外，BLIP-2[20]提出了一种将视觉特征映射到与文本特征相同的特征空间中的对齐方法，使用一组可学习的查询词嵌入结构和视觉特征输入解码器(Q-Former)，将Q-Former输出的查询词嵌入通过一个全连接层后直接与文本特征拼接输入冻结的大语言模型，综合考虑图像文本匹配(Image-TextMatching)、基于图像的文本生成(Image-groundedTextGeneration)、图像文本对比学习(Image-TextContrastiveLearning)三个多模态任务构造损失函数用于训练，实现了图像特征与文本特征的对齐。Mini-GPT4[16]，mPLUG-OWL[17]，VPGTrans[18]，InstructBLIP[19]在BLIP-2[20]的基础上保留了Q-Former，改用了参数量更多的大语言模型，在精心收集的指令数据上进行了微调。相比于使用Q-Former对齐图像与文本特征，FROMAGe[49]和LLaVA[50]使用了一种更简单直接的方法，将图像特征通过一层全连接层映射后直接输入大语言模型。但是这种做法本质上是将原本Q-Former和大语言模型的功能耦合在了一个模型之中，模块化程度的降低给模型的可解释性带来了更大的挑战。最近，基于LLaVA[50]的结构，Shikra[51]实现了理解和生成自然语言形式表示的目标检测框。此外，未开源的模型GPT-4[52]也表现出了强大的图像-文本理解能力。

上述视觉语言大模型虽然在多种复杂推理任务上取得了优异的结果，然而这些模型的推理过程难以被人类直观理解。为此，本工作侧重于研究面向视觉语言大模型的自然语言解释任务，通过模型生成的自然语言语句来帮助人们理解“黑盒”模型的推理过程。我们的工作使用了BLIP-2[20]的Q-Former进行图像与文本特征对齐，但与之前的工作[16-19]不同的是我们没有简单替换掉BLIP2[20]的语言模型，而是在其中加入探针结构，保持原有模型的完整推理过程，从原有推理过程的每个隐藏状态中提取信

息用于自然语言解释任务。

2.2 视觉语言领域的自然语言解释任务

自然语言解释(NaturalLanguageExplanation,NLE)任务旨在通过生成的自然语言语句来解释“黑盒”模型的推理过程,该任务由Hendricks等人[13]首先在图像描述任务的基础上提出,之后延伸到了视觉语言领域的其他任务中。Park等人[11]提出了一种多模态解释方法,同时提供自然语言解释和注意力可视化,验证了多模态解释方法可以使视觉模态和语言模态的解释互补受益。Wu等人[12]在面向视觉问答的自然语言解释任务基础上,实现了输入图像中的重要区域与生成文本中词语的可视化对应,突破了单一模态的文本解释极易导致指代语义模糊的问题。Marasović等人[10]提出了一个全栈视觉推理的自然语言解释生成模型,集成了像素级别的视觉信息、语义框架和常识图谱,在视觉常识推理、视觉文本蕴涵和视觉问答任务上进一步发展了自然语言解释任务。虽然这些研究[10-13]在视觉语言领域的自然语言解释任务上做出了有价值并且有启发性的工作,但是Sammani等人[7]指出他们[10-13]都使用了独立的任务模型和解释模型先后生成答案和自然语言解释,导致解释的生成过程与任务模型的推理过程之间缺少直接的逻辑联系。

Sammani等人[7]进而提出了一种名NLX-GPT的视觉语言与解释器模(VLandExplainer Model),该模型将图像特征和文本特征映射到相同的特征空间,拼接后同时输入视觉语言模型中,并重新设计了视觉语言模型的文本生成任务,让生成的文本中同时包含对原任务的答案和自然语言解释,从而能够在视觉和视觉-语言任务中同时预测答案并提供解释。最近,Suo等人[8]在此基础上提出了一种基于自我批判学习的半监督视觉问答自然语言解释方法,设计了一个自我批判强化学习模块,用视觉语言与解释器模型生成的候选解释与原输入、提示词拼接后重新输入模型,用模型再输出正确答案的概率计算奖励分数进行强化学习,进一步提升了模型的逻辑一致性。还有一些最近的工作[14-15]将视觉语言与解释器模型与多模态解释方法结合,在原任务上提升了模型的性能。这些工作[7-8,14-15]因为使用了视觉语言与解释器模型代替了分立的任务模型-解释模型结构,所以不存在解释生成过程与推理过程脱离的问题,但是同时也导致了解释生成过程会对推理过程产生影响。这种影响让解释过程本质上成为了推理过程的一部分而不是对原有推理过程的解释,从而弱化了解释的忠实性。

随着大模型Prompt技术的发展,最近一些研究在预测答案时通过输出中间逐步的推理解释来增强模型推理能力。具体来说,Wei等人[53]首次提出了一种思维链策略以捕捉模型的推理过程,该方法已在许多少样本学习任务中取得成功[54-56]。在此基础上,也有研究工作[57]以零样本方式扩展了这个想法,并在每个问题前添加一个简单的提示“让我们一步一步思考”。这种零样本CoT策略优于以前的零样本方法,并为NLP推理任务生成合理的推理路径。此外,Zhao等人[58]提出了一种逻辑思维链,要求模型逐步思考和验证推理过程,以提高模型的答案预测能力。

不同于上述方法,我们新提出了一种新的基于探针引导的视觉语言多模态解释策略。相较于后验解释方法及自解释方法[10-13],PME模型能够同时生成预测答案和相应的解释。此外,不同于现有的自解释方法[10-13]及思维链生成方法[53,57-58],我们的框架使用探针结构从推理中提取信息而不改变模型对答案的预测过程,避免了解释的生成对原有推理路径的影响,保证了其忠实性。另外,针对单一模态的文本解释存在的指代语义歧义问题,我们的多模态解释方法实现了生成自然语言解释的同时生成目标指代框,从而提高了解释的可靠性。

3 基于探针的多模态解释方法

正如在第一节中讨论的,在面向视觉问(VQA)的自然语言解释(NLE)任务中,现有的VQA-NLE方法难以建模网络内部的推理过程和保证生成解释的忠实性。为了基于原有VQA推理模型生成相应的解释,我们提出了一种新的基于探针引导的视觉语言多模态解释方法。该方法主要包含一个探针结构和一个多模态解释生成器。具体而言,新引入的探针结构能够从推理模型的每个隐藏层状态中提取信息,并将该信息解码生成自然语言解释。伴随着注意力掩码机制,这些探针可以有效避免解释过程对推理决策模型的干扰。此外,为了进一步缓解单一模态下的语义歧义问题,我们利用交叉生成伪标签技术,在生成的文本解释中嵌入目标检测框,构建了跨模态的图像和文本同步解释框架,保证了解释过程对推理模型的忠实性和明确性,接下来我们将详细介绍它们。

3.1　视觉语言模型的推理过程

BLIP2[20]因其优秀的推理和泛化能力在之前的工作中被广泛使用[16-19]。在本文中，我们将该模型作为主要研究对象。BLIP2由视觉编码器、Q-Former模块和语言模型组成。在完成VQA任务时，BLIP2[20]首先映射输入的自然语言问题为文本特征$Q_0=\{q_t\}_{t=1}^{T}, q_t\in R^C$。同时，对于输入图像$I\in R^{W*H*3}$，一个视觉编码器(Vision_Encoder)被用来提取视觉特征。然后，BLIP2[20]中的Q-Former模块将这些特征映射到文本空间表示为$V_0=\{v_s\}_{s=1}^{S}, v_s\in R^C$。上述计算可以表示如下：

$$V_0=Q\_Former\left(\text{Vision\_Encoder}(I)\right) \tag{1}$$

然后，视觉特征V_0与文本特征Q_0拼接，得到语言模型编码器(LLM Encoder)的初始输入H_0：

$$H_0=\{v_1, v_2, \cdots, v_S, q_1, q_2, \cdots, q_T\} \tag{2}$$

其中，S代表视觉特征向量数量，T代表样本中问题文本的序列长度。在本工作中S取值32。该语言模型编码器由24层相同结构的编码器块(Block)组成。其中，每一个Block包含一个自注意力层(Self-Attention)和一个前馈神经网络(Feed Forward Network，FFN)。我们将拼接后的特征输入编码器中得到最后一层的输出，它将被一个语言模型解码器(Decoder)进一步地解析为该VQA任务的答案。在上述操作中，编码模块主要负责多模态的对齐和推理过程，因此，本文后续的讨论将主要建立在对该模块的研究中。

3.2　探针结构设计

对于每一个VQA样本，为了提取相应的解释且不对原始推理链路造成干扰，我们引入了一种新的探针结构。同时，考虑到模型的最后的隐状态并不能充分地汇总模型的推理过程[59]，我们将探针作为一组独立的输入添加到模型中。具体而言，我们首先冻结了原有的BLIP2模型，设计了一组可学习的探针$E_0=\{e_n\}_{n=1}^{N}, e_n\in R^C$，然后将$E_0$与原编码器的初始输入$H_0$拼接，我们得到了新的初始输入$H_0'$：

$$H_0'=\{v_1, v_2, \cdots, v_S, q_1, q_2, \cdots, q_T, e_1, e_2, \cdots, e_N, \} \tag{3}$$

其中N代表可学习的探针数量。我们使用P=N+S+T表示融合了探针结构的输入序列的长度。为了让推理模型不受探针的干扰，我们进一步地在每个编码器块中添加了自注意力掩码机制。如图2所示，该注意力掩码机制首先构造一个掩码矩阵$M_{P\times P}$，并对其中每个元素按如下规则赋值：

$$M(j,k)=\begin{cases}-\infty, & j<N\text{同时}k>N\\ 0, & \text{否则}\end{cases} \tag{4}$$

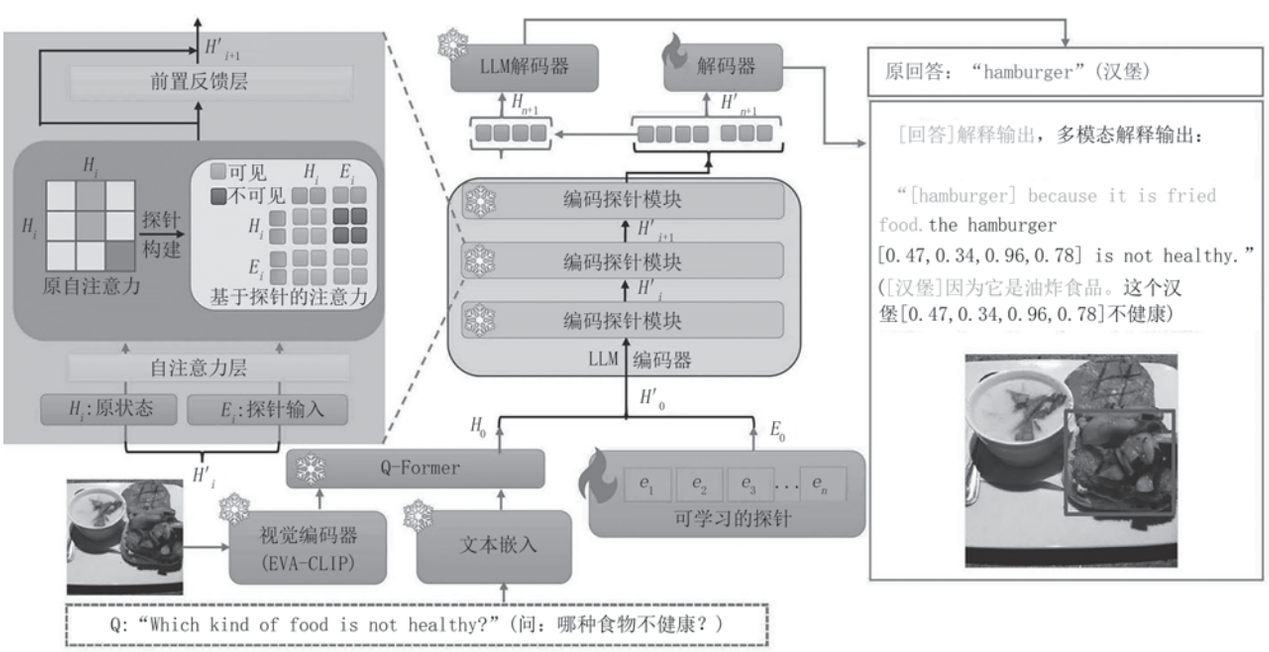


图2　基于BLIP2[20]的PME模型结构图(首先，PME模型在BLIP2[20]的基础上加入了可学习的探针，并将原编码器块修改为编码探针模块，通过使用一个注意力掩码机制使得探针在提取原状态信息时不干扰模型的推理过程。其次，我们应用交叉伪标签技术构造了一个新的多模态解释数据集，训练模型同时生成自然语言解释和目标检测框。最后我们整合上述两种方法提高了推理模型解释的忠实性和可靠性。)

这里j,k分别表示掩码矩阵的第j行和第k列。其中,每一个编码块中的自注意力权重和相应的输出被按照如下方式计算:

$$\begin{cases} Att_i = Q_i K_i^T / \sqrt{D} \\ Score_i = Softmax(M + Att_i) \\ h_i' = Score_i V_i \end{cases} \tag{5}$$

在这里,Q_i,K_i和V_i分别表示第i个编码块自注意力操作中的查询矩阵(***Query***),键矩阵(***Key***)和值矩阵(***Value***)。h_i'表示第i个编码块中自注意力的输出。对于在每一个编码块中的前馈神经网络,我们将h_i'直接作为其输入获得H_i'。通过对24层编码器中的每一层都堆叠类似的操作,我们获得了最终的编码器输出H_{24}':

$$H_{24}' = \{v_1, v_2, \cdots, v_S, q_1, q_2, \cdots, q_T, e_1, e_2, \cdots, e_N, \} \tag{6}$$

受益于上述操作,我们可以方便地分离出原始BLIP-2解码器的输入H_{24}。而H_{24}'将被送入一个独立的解码器以得到模型的推理解释。

3.3 多模态解释数据集构建

通过引入探针和注意力掩码机制,我们可以生成忠实于VQA推理过程的文本解释。然而,如图1所示,当图像中出现多个目标对象时,单模态的自然语言解释仍存在指代语义模糊的问题。为此,本文使用目标检测框作为辅助解释来弥合语义歧义鸿沟。

由于现有的数据集只有单模态的自然语言解释标注。因此,我们进一步地引入了具有视觉问答和视觉解释标注的GQA数据集,并使用半监督学习中的伪标签(Pseudo-labeling)方法建立了一个新的训练集。具体而言,我们首先使用文本解释数据集VQA-X和输出模版{<ans>,Ans,<exp>,Exp}微调BLIP2得到文本解释模型($Model_{exp}$)。同时,用GQA数据集和输出模版{<ans>,Ans,<exp>,Pos}微调BLIP2得到视觉解释模型($Model_{pos}$)。在上述模板中,我们遵循之前的工作[7-8],使用<ans>表示字符串"the answer is",<exp>表示字符串"because",Ans是标注的VQA任务答案,Exp是标注的对当前样本的自然语言解释。而Pos代表由字符串"$[x_1, y_1, x_2, y_2]$"表示的一个目标检测框,它通过计算该指代框的左上角和右下角的像素坐标与图像的宽高比值得到。因此,$x_1, y_1, x_2, y_2 \in [0, 1]$。

通过训练上述模型,我们可以分别得到一个文本解释模型$Model_{exp}$和视觉解释模型$Model_{pos}$。值得注意的是,由于构造跨模态解释数据集的步骤不需要维护原模型的推理过程,BLIP2全部的参数被更新以获得更好的性能。然后,我们通过交叉标注获得每一个样本的跨模态解释伪标签。该策略可以对每一个具有文本注释的VQA-X样本得到一个由$Model_{pos}$产生的伪标注框。反之亦然,每一个具有视觉解释的GQA样本得到一个由$Model_{exp}$产生的伪文本解释。最后,为了减少由伪标签引入的标注噪声,我们利用预测的答案Ans作为基准进行样本筛选并最终得到具有跨模态解释的训练数据集S_{train}(约45万个样本)。在该训练数据集中,每一个训练样本st都具有下列组织形式:

$$s_t = \{<ans>, Ans, <exp>, Exp, Pos\} \tag{7}$$

3.4 多模态解释训练

基于3.2节的公式6得到的H_{24}',我们引入了一个新的解码器D用来生成多模态解释,它的结构及初始参数与BLIP2中的解码器相同。最后,我们通过交叉熵(Cross Entropy Loss)对文本解释和视觉解释s_t计算损失并更新该解码器D和探针E_0的参数:

$$L = -\frac{1}{K}\sum_{k=1}^{K} \log P(y_k | \{y_i\}_{i=1}^{k-1}; \theta) \tag{8}$$

其中,y_k表示当前多模态训练样本st中的第k个词(即文本解释和视觉框解释框)。θ表示探针及解码器的参数,K表示标注的多模态解释的长度。值得注意的是,PME模型在训练时只对输出的Exp和Pos部分计算损失函数,而不影响BLIP2模型的答案预测。通过联合探针结构和多模态解释,我们的方法能够在不破坏原始BLIP2推理过程的基础上实现基于多模态的推理解析。

4 实验与分析

4.1 实验设置

4.1.1 数据集

为了验证我们提出的解释方法的有效性,我们主要在广泛使用的视觉推理问答数据集VQA-X和A-OKVQA上构造了实验。同时为了生成多模态解释,我们也应用GQA数据集训练模型来获取目标检测框,接下来我们将分别介绍它们。

VQA-X是一个视觉语言数据集,它的标注包括文本形式的问题、答案和能证明答案合理性的解释。VQA-X从VQA数据集[60]中收集了问题和答

案，使用MSCOCO[61]中的图像，包括28K图像和33K问题答案对，按照29K/1.4K/1.9K划分为训练集、验证集和测试集。

A-OKVQA包括24 903个问题/答案/基本原理三元组，从COCO2017[62]数据集中收集并过滤得到23.7K唯一图像，按照17.1K/1.1K/6.7K划分为训练集、验证集和测试集。与VQA-X相比，A-OKVQA有更丰富的问题，需要更广泛的知识进行推理。

GQA[63]是一个强调逻辑推理的视觉语言数据集，共包含约1.1万张图像，每张图像都有与之关联的问题和答案。这些问题涵盖了各种类型的推理，从基本的视觉识别到更复杂的组合式推理。每个答案都有与之对应的目标检测框标注，指向图像中的相关区域。

4.1.2　评价标准

为了公平比较，我们使用了和最先进的NLX-GPT[7]完全相同的评价指标。如表1至6所示，我们分别使用B1-B4，M，R，S和C表示BLEU[64]，METEOR[65]，ROUGE-L[66]，SPICE[67]和CIDEr[68]去评测生成的解释。此外，我们使用与S3C[8]相同的人工评估过程对生成的文本解释进行了评价，即对于每个解释，三位人工评估者决定该解释是否能证明答案，并选择一个选项（包括“是(yes)，弱是(weakyes)，弱否(weakno)和否(no)”）。这些选择将被映射为分数(1,和0)。最终的人类评估分数将通过对所有测试样本进行平均计算得出。

4.1.3　实现细节

我们使用了开源深度学习库LAVIS[69]作为开发框架，将所有图像在预处理时缩放为224*224大小，使用冻结的EVA-CLIP_g[70]作为视觉编码器。在我们的模型中可学习的部分包括探针prompt和一个独立的解码器D。初始学习率为1e-05，权重衰减为0.05，最大训练轮数为3，训练批大小为8。我们设置可学习探针数量为48，并在单张NVIDIA GeForceRTX3090上进行训练。

表1　经过基于答案的过滤后，在VQA-X测试集上与最优方法的比较

	VQA-X					
	B4	M	R	S	C	Human
RVT[10]	17.4	19.2	42.1	15.8	52.5	60.5
PJ-X[11]	22.7	19.7	46.0	17.1	82.7	69.3
FME[12]	23.1	20.4	47.1	18.4	87.0	-
QA-only[9]	17.3	18.6	41.9	14.9	49.9	-
e-UG[9]	23.2	22.1	45.7	20.1	74.1	71.4
NLX-GPT[7]	28.5	23.1	51.5	22.1	110.6	73.7
S3C[8]	29.1	23.4	51.9	22.7	112.1	75.9
PME-BLIP2	30.5	23.5	52.2	23.5	117.7	77.7
PME-QwenVL	31.9	25.1	54.5	25.1	121.1	80.7
PME-LLaVA1.5	**32.4**	**24.9**	**54.6**	**25.5**	**123.5**	**81.9**

注：其中，B4，M，R，S和C表示BLEU[64]，METEOR[65]，ROUGE-L[66]，SPICE[67]和CIDEr[68]分数。Human表示人工评价分数。

表2　经过基于答案的过滤后，在A-OKVQA测试集上与最新最优方法的比较

	A-OKVQA					
	B4	M	R	S	C	Human
e-UG[9]	18.9	19.5	47.0	19.7	69.4	51.2
NLX-GPT[7]	26.5	20.3	51.2	20.4	95.0	55.9
S3C[8]	25.4	20.0	51.4	21.0	100.6	57.2
PME-BLIP2	25.5	20.4	51.4	21.3	102.7	61.5
PME-QwenVL	27.0	22.9	52.5	22.4	104.5	66.9
PME-LLaVA1.5	**27.5**	**23.6**	**53.0**	**23.9**	**107.9**	**68.1**

注：其中，B4，M，R，S和C表示BLEU[64]，METEOR[65]，ROUGE-L[66]，SPICE[67]和CIDEr[68]分数。Human表示人工评价分数。

表3　不经过基于答案的过滤，在VQA-X测试集上与最优方法的比较

	VQA-X					
	B4	M	R	S	C	Human
CAPS[11]	5.9	12.6	26.3	11.9	35.2	-
PJ-X[11]	19.5	18.2	43.4	15.1	71.3	65.4
FME[12]	24.4	19.5	47.7	17.9	88.8	-
NLX-GPT[7]	25.6	21.5	48.7	20.2	97.2	70.2
S3C[8]	26.5	22.0	49.0	20.9	100.5	72.7
PME-BLIP2	29.9	23.2	51.6	22.8	115.8	75.6
PME-QwenVL	32.4	25.9	**53.9**	26.0	118.9	79.0
PME-LLaVA1.5	**32.8**	**26.2**	53.6	**27.1**	**119.4**	**79.9**

注：其中，B4，M，R，S和C表示BLEU[64]，METEOR[65]，ROUGE-L[66]，SPICE[67]和CIDEr[68]分数。Human表示人工评价分数。

表4　不经过基于的答案的过滤，在A-OKVQA测试集上与最优方法的比较

	A-OKVQA					
	B4	M	R	S	C	Human
e-UG[9]	15.1	18.1	42.4	14.9	51.5	44.1
NLX-GPT[7]	20.1	17.0	46.3	15.8	65.4	46.9
S3C[8]	21.8	17.9	47.3	17.3	70.6	50.0
PME-BLIP2	24.2	19.1	48.9	18.6	83.4	55.4
PME-QWenVL	26.9	22.3	51.1	20.9	90.1	59.2
PME-LLaVA1.5	**27.1**	**22.9**	**51.5**	**21.6**	**92.5**	**59.7**

注：其中，B4，M，R，S和C表示BLEU[64]，METEOR[65]，ROUGE-L[66]，SPICE[67]和CIDEr[68]分数。Human表示人工评价分数。

表5 在VQA-X测试集上的消融实验

	探针结构	解释微调	伪标签	目标检测框	B1	B2	B3	B4	M	R	S	C	AnsDiff/%↓
1	√	×	×	×	55.1	43.7	34.3	26.8	20.9	49.9	19.7	105.1	**0**
2	×	√	×	×	55.6	44.0	34.5	27.1	20.8	49.8	19.6	104.1	8.38
3	×	√	√	×	57.9	45.8	35.9	28.3	21.9	51.4	21.3	111.9	13.31
4	×	√	√	√	63.0	49.4	38.6	30.4	23.5	52.2	23.2	115.8	12.60
5	√	×	√	√	**63.8**	**50.0**	**39.0**	**30.5**	**23.5**	**52.2**	**23.5**	**117.7**	**0**

注：其中，B1-B4，M，R，S和C表示BLEU[64]，METEOR[65]，ROUGE-L[66]，SPICE[67]和CIDEr[68]分数。AnsDif(%)表示与BLIP-2基准答案相比预测答案有差异的样本在总测试样本中所占百分比。

表6 PME方法使用不同数量的可学习探针时，在VQA-X数据集上的性能比较

可学习词嵌入的数量	B1	B2	B3	B4	M	R	S	C
32	63.4	49.6	38.7	30.3	23.4	52.2	23.3	116.6
48	63.8	50.0	39.0	30.5	23.5	52.2	23.5	117.7
64	64.1	50.3	39.3	30.9	23.7	52.7	23.7	118.4

注：其中，B1-B4，M，R，S和C表示BLEU[64]，METEOR[65]，ROUGE-L[66]，SPICE[67]和CIDEr[68]分数。

4.2 性能比较实验

为了验证我们提出的方法在视觉问答任务上的解释能力，如表1至4所示，我们在VQA-X和A-OKVQA数据集上将PME模型与最新最优的方法进行了比较。同时，本文在更多的多模态大模型(即QwenVL[71]和LLaVA[72])上构建了实验以证明本方法的有效性和鲁棒性。同时，由于这些模型没有使用独立的编码器结构，而是仅依赖解码器完成推理。因此，为了不干扰原有的推理过程，我们将探针及对应的掩码机制添加到了输入序列中，并将其与原始模型的输出一起送入一个具有6层Transformer结构的解码器中进行微调。与BLIP2类似，交叉熵损失仅被用于更新探针及新引入的解码器，而原始模型的参数将保持冻结。在表1和表2中我们只评估那些答案预测正确的样本解释，而在表3和表4中展示了不经过答案过滤直接对生成解释的评测。

从表1至4中的实验结果可知，无论是否过滤正确的答案样本，我们提出的PME模型的解释分数在两个数据集上都显著地高于现有的后验解释方法[9-12]。例如在表1中，相较于最先进的自解释模型NLX-GPT和S3C[7,8]，我们提出的多模态解释策略分别实现了在VQA-X数据集上CIDEr分数7.1%和5.6%的绝对提升。更重要的是，与现有的最优方法S3C相比，无论是自动化评价还是人工评价，本方法均能获得一致的性能提升，同时保证了解释的可靠性和忠实性。

4.3 消融实验

为了证明PME方法的有效性，如表5所示，我们在VQA-X测试集上结合BLIP-2模型进行了消融实验。其中，AnsDif表示与BLIP-2基准答案相比预测答案有差异的样本在总测试样本中所占百分比。具体而言，在第1行中，我们使用探针结构对模型的推理过程进行解释，可以发现AnsDif的值为0，该结果表明本文所提结构并不会影响原始模型的推理过程。相反，在第2行中，我们使用解释标注来微调BLIP2模型的解码器。通过分析实验结果我们发现，简单地使用解释标注进行微调会干扰原始模型的预测，并使得8.38%的答案预测发生变化。在第3-4行，我们进一步引入伪标签与目标检测框的多模态解释策略来提升模型的解释能力，可以发现这些策略使得解释的CIDEr分数提高了7.4%。最后，当我们使用探针结构替换简单的解释微调方法，可以发现CIDEr指标达到117.7，而答案预测的忠实性也得到了保证。

上述结果表明我们的多模态解释方法在保证忠实度的前提下有效地提升了模型的解释能力。

4.4 对比实验

为了分析探针数量对PME模型性能的影响，我们分别设置了32、48、64个可学习探针，并在VQA-X测试集上进行评估。实验结果如表6所示。当可学习探针的数量从32个增加到48个时，模型解释的CIDEr分数提升了0.94%；当可学习探针数量进一步提升时，模型只获得了少量的CIDEr分数增益。因此为了平衡性能和参数量，在本工作中我们将可学习探针的数量设置为48。

4.5　答案与解释的干扰性分析

为了证明传统自解释方法中的解释过程对答案预测存在不可避免的影响，以及我们提出的探针结构能够有效地避免这种干扰，我们在VQA-X数据集上分别对BLIP2模型进行了三组微调实验。如图3所示，在第一组实验中，我们仅使用答案标注对BLIP2模型的解码器进行微(BLIP2-A)。在第二组实验中，我们使用答案和解释融合标签对BLIP2模型的解码器进行微调(BLIP2-A&E)，并使用微调后的模型同时输出答案与解释。实验结果表明，当加入针对解释的损失后，第二组与第一组相比有8.38%的测试样本对相同问题的预测答案发生了改变。这证明了现有的联合训练方法将会干扰模型的预测，从而无法保证解释过程的忠实性。在第三组实验中，我们使用探针结构进行微调(BLIP2-A&P)。受益于该设计，编码器的输出将分为两部分，其中一部分通过与第一组相同的解码器生成答案，而探针输出通过额外的解码器生成解释。实验结果表明，加入探针结构后的第三组(BLIP2-A&P)能够保持与第一组(BLIP-A)完全相同的预测答案，避免了解释过程对答案预测的干扰。

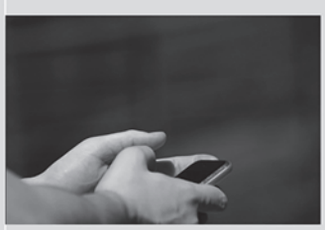
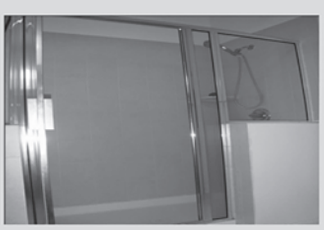
数据集 VQA-X			
问题	What is man doing?	What is this?	What is the gender of the players?
GT 基准真实标注	答：[texting]	答：[shower]	答：[male]
BLIP2-A	答：[reading]	答：[shower]	答：[male]
BLIP2-A&E	答：[texting] 解释：he is holding a cell phone	答：[bathroom] 解释：there is a shower and a toilet	答：[men] 解释：they are wearing men's clothing
BLIP2-A&P	答：[reading] 解释：he is holding a phone	答：[shower] 解释：there is a shower head	答：[male] 解释：they have short hair and wear men's clothing

图3　三组微调实验(图中“BLIP2-A”行表示单独使用答案标签微调后的BLIP2模型的预测答案，“BLIP2-A&E”行表示使用答案和解释共同作为标签微调后的BLIP2模型的预测答案。“BLIP2-A&P”行表示使用探针结构对BLIP2模型进行微调。)

4.6　多模态解释必要性分析

在引言中我们提到仅使用单一文本解释会产生严重的语义歧义。为了进一步说明该问题，如图4所示，即使现有数据集中的语言注释是通过严格的标注说明和数据清洗过程获得的，但仅就文本解释依然不能有效地缓解语义模糊问题。因此本文主要研究多模态解释方法，一方面在很多情况下，一图胜千言，视觉解释比文本解释更有洞察力。例如，对于“leaning[学习]”这个动作，用语言描述比较困难，但其在视觉上却很容易识别。另一方面视觉解释可以辅助文本解释，提供决策的视觉证据，从而提升模型整体的解释能力。

为了进一步证明多模态解释可以有效地缓解语义歧义问题。我们在表7中对单模态和多模态两种类型解释进行人工评测。具体而言，该过程由三位人工评估者决定其是否存在歧义，并选择一个选项(包括“否(解释完全清晰，没有歧义)，弱否(内容大致清晰，但仍有轻微的歧义)，弱是(部分内容存在歧义，但其他部分可以理解)和是(存在严重的歧义，几乎无法理解)”)。这些选择将被映射为分数(1，$\frac{2}{3}$，$\frac{1}{3}$和0)。由于本节重点关注于解释的歧义性分析，因此本实验仅在基于答案过滤后的样本上进行。最终的人类评估分数将通过对所有测试样本进行平均计算得出。通过观察表7中的结果，我们发现使用多模态解释可以获得更高的人工评分，从而证明多模态解释具有更加明确的语义表达。

数据集 VQA-X			
问题	Does this photo show train tracks?	What room of the house is this?	What game is being played?
GT 基准真实标注	答案：yes 解释：there are railroad tracks layed in parallel next to the train	答案：kitchen 解释：there are appliances and pots and pans	答案：soccer 解释：they are kicking around a ball
指代模糊	无法判断解释中的”tracks”指代了图像中的哪部分或者全部铁轨	无法判断解释中的”appliances and pots and pans”指代了图像中的哪部分或者全部的电器和锅	无法判断解释中的”they”指代了图像中的哪些或者全部的人

图4 单模态解释遭受指代模糊问题的示例分析

表7 单模态解释与多模态解释中基于人工的语义歧义评价

	VQA-X	A-OKVQA
PME(单模态解释)	67.5	58.9
PME(多模态解释)	79.2	66.2

4.7 定性分析

如图5所示，我们展示了最优的NLX-GPT[7]方法和我们的PME模型在VQA-X数据集上的定性结果。实验结果表明，我们的PME方法生成了与答案更加逻辑一致的解释，同时缓解了语义指代模糊的问题。具体地，在样本(a)(b)中，NLX-GPT[7]虽然能识别出图像中的重要特征(例如a中的领带和b中的冲浪板)，但预测的答案却与解释相矛盾。相反地，我们的PME方法则能够得到逻辑一致的答案与解释，这表明我们的方法可以生成忠实于推理过程的解释。此外，在样本(d)中，NLX-GPT[7]生成的单一文本解释＜它是棕色的，有颗粒感＞中的“它”具有指代语义模糊的问题，因为图中的桌子、立柜和橱柜都是木质的，都满足棕色的且有颗粒感的条件，进而导致我们无法明确模型是否能够正确地捕获相关语义。相反，我们的PME模型通过多模态解释“它是棕色的并且是用木头做的[0.00，0.58，1.00，1.00]”从图像中明确标出了指代对象，缓解了单一文本解释的指代语义模糊问题。

为了更好地分析本方法的局限性，我们在图6展示了本方法失败的几个示例。在这些示例中，我们发现本方法对于微小物体(示例g)、背景杂乱(示例h)和语义众多(示例i)的场景具有较差可解释性，这可能是由于伪标签的生成过程中将不可避免地引入一些噪声，进而损害模型对于这些复杂样本的解释能力。在未来的研究中我们将构造更加丰富多样的数据样本提高模型对于复杂场景的泛化能力。

5 局限性和未来研究方向

在本文中，我们对多模态解释方法进行了初步探索，并有效地缓解了现有解释模型中存在的解释不忠实和语义歧义的问题。然而我们的方法也存在下列局限：(1)为了不干扰原模型的推理过程，我们冻结了BLIP2的所有参数，而BLIP2是建立在预训练的语言大模型上，因此我们的方法直接继承了它们的弱点，例如多模态生成任务中的幻觉问题和对于视觉问答任务中对于语言模型的过度依赖问题。(2)作为一个初步的探索，本工作只针对视觉问答任务中的一般推理和事实推理进行了自然语言解释任务的研究，而没有涉及视觉语言领域其他类似的推理任务，例如视觉语言导航和视觉文本对话等，我们将这些任务考虑作为未来的研究工作。(3)我们观

			
问题	(a) Is this man dressed formal? 这个男人穿得正式吗？	(b) Is he in a swimming pool? 他在游泳池里吗？	(c) Is the man right-handed? 这个人是右撇子吗
GT 基准真实标注	[yes]<he's wearing a crisp dress shirt and a tie> [是的]<他穿着一件挺括的正装衬衫，打着领带>	[no]<the man is riding an ocean wave> [不]<那个男人正在乘着海浪>	[yes]<he is holding a racket with his right hand> [是的]<他右手拿着球拍>
NLX	[no]<he is wearing a tie> [不]<他系着领带>	[yes]<he is on a surfboard> [是的]<他在冲浪板上>	[no]<he is holding a tennis racket> [不是]<他拿着网球拍>
PME(Ours)	[yes]<he is wearing a suit and tie.> the man [0.19, 0.08, 0.88, 1.00] is dressed formal [是的]<他穿着西装打着领带。> 该男子[0.19, 0.08, 0.88, 1.00]穿着正式	[no]<he is surfing in the ocean.> the man [0.30, 0.29, 0.60, 0.79] is in the ocean [不]<他正在海里冲浪。> 这个人[0.30, 0.29, 0.60, 0.79]在海里	[yes]<he is holding the racket with his right hand.> the man [0.30, 0.39, 0.60, 0.92] is right-handed [是的]<他用右手握着球拍> 这个人[0.30, 0.39, 0.60, 0.92]是右撇子
			
问题	(d) What is the table made of? 桌子是用什么做的	(e) Is the girl having fun with her kite? 这个女孩玩她的风筝很开心吗	(f) What is the elephant doing? 大象在做什么？
GT 基准真实标注	[wood]<the table is brown and has a wood finish> [木材]<这张桌子是棕色的，表面是木质的>	[yes]<the girl is smiling happily> [是的]<女孩开心地笑着>	[drinking]<his trunk is in the water> [喝水]<他的鼻子在水里>
NLX	[wood]<it is brown and grainy> [木材]<它是棕色的，有颗粒感>	[yes]<she is smiling and holding it up> [是的]<她笑着把它举了起来>	[drinking]<it is in the water> [喝水]<它在水中>
PME(Ours)	[wood]<it has a brown color.> the table [0.00, 0.58, 1.00, 1.00] is made of wood [木材]<它是棕色的。> 这个桌子[0.00, 0.58, 1.00, 1.00]是木头做的	[yes]<she is smiling while holding the kite. the girl [0.00, 0.09, 0.78, 1.00] is having fun with her kite [0.00, 0.00, 0.50, 0.33] [是的]<她拿着风筝微笑着。> 女孩[0.00, 0.09, 0.78, 1.00]正在玩她的风筝[0.00, 0.00, 0.50, 0.33]	[drinking]<he is reaching his trunk into the water.> the elephant [0.09, 0.09, 0.92, 0.92] is drinking [喝水]<他正把鼻子伸进水里。> 大象[0.09, 0.09, 0.92, 0.92]正在喝水

图5　本文方法与现有最优方法NLX-GPT在VQA-X数据集上的定性比较（[]和<>分别表示答案和解释。我们展示了NLX-GPT[7]、我们的PME方法的结果以及真实标注；它们分别简写为NLX、PME和GT。）

察到在表5的第二行中引入探针会带来BLEU分数轻微下降而CIDEr分数上升的现象。由于两种评估指标在侧重点上的差异，BLEU更注重在词元(n-gram)上的精确匹配，而CIDEr更能反映语义的一致性。由于文本解释本身具有多样性的特点，上述自动化的评测指标并不能完全地反应解释质量，进

	(g)	(h)	(i)
问题	(g)What kind of animal is this? 这是什么动物?	(h)What kind of animals are on the man's lap? 男人腿上的是什么动物?	(i)Are the people having a party? 人们在开派对吗?
GT 基准真实标注	[bird]<it has two wings and is flying> [鸟]<它有两个翅膀并且在飞>	[cats]<they are furry and have whiskers> [猫]<它们毛茸茸的，有胡须>	[yes]<they are standing in a kitchen with drinks in hand and laughing> [是的]<他们站在厨房里，手里拿着饮料，笑着>
PME(Ours)	[bird]<it has wings and is flying in the air> this is a bird [0.41, 0.39, 0.51, 0.59] [鸟]<它有翅膀并且在空中飞> 这是一只鸟[0.41, 0.39, 0.51, 0.59]	[cats]<they are black and white cats.> the animals [0.39, 0.39, 1.00, 1.00] are cats [0.39, 0.39, 1.00, 1.00] [猫]<它们是黑色和白色的猫。> 这些动物[0.39, 0.39, 1.00, 1.00]是猫[0.39, 0.39, 1.00, 1.00]	[yes]<because they are drinking in a kitchen.> they are having a party [是的]<因为他们在厨房里喝饮料。> 他们正在开派对

图6 本文方法在一些样本上表现失败的示例（[]和<>分别表示答案和解释。我们展示了我们的PME方法的结果以及真实标注；它们分别简写为PME和GT。）

而我们也引入人工评测对我们的方法进行评价。在未来的工作中我们将探索更多的结构和更加可靠的评测手段，对模型的可解释性进行进一步的讨论。

可以预见的是，随着视觉语言模型的发展，未来会有更多的工作致力于帮助人们理解视觉语言模型的推理过程，进而推动自然语言解释任务的研究。希望本文可以为相关工作提供一个新的研究思路作为参考，进而对后续的研究有所启发。

6 结 论

本文提出一种新的基于探针引导的多模态解释方法，该方法基于通用的VQA模型（BLIP2），并引入一种新的探针结构，实现了在不干扰推理过程情况下的模型解释，保证了解释过程的忠实性。另外，本文利用VQA-X数据集和GQA数据集来交叉生成伪标签，构造了一个同时具有视觉问答、自然语言解释和指代框的训练集。通过设计一种多模态解释方法，缓解了单一文本模态下的指代语义模糊问题。本文在公开的数据集VQA-X和A-OKVQA上将PME方法与其他最新最优方法比较，我们的PME策略均取得了更好的性能，这验证了本文提出的框架能更有效地解释模型的推理过程。另外，本文设计了完整的消融实验，验证了同时生成目标检测框对自然语言解释的辅助作用，在保证忠实性的前提下生成了更加明确可靠的多模态解释。

参 考 文 献

[1] Cornia M, Stefanini M, Baraldi L, et al. Meshed-memory transformer for image captioning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 10578-10587

[2] Rennie S J, Marcheret E, Mroueh Y, et al. Self-critical sequence training for image captioning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 7008-7024

[3] Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 6077-6086

[4] Pan Y, Yao T, Li Y, et al. X-linear attention networks for image captioning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 10971-10980

[5] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 618-626

[6] Patro B N, Lunayach M, Patel S, et al. U-cam: Visual explanation using uncertainty based class activation maps//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 7444-7453

[7] Sammani F, Mukherjee T, Deligiannis N. Nlx-gpt: A model for natural language explanations in vision and vision language tasks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022:

8322-8332

[8] Suo W, Sun M, Liu W, et al. S3c: Semi-supervised vqa natural language explanation via self-critical learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 2646-2656

[9] Kayser M, Camburu O M, Salewski L, et al. E-vil: A dataset and benchmark for natural language explanations in vision language tasks//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 1244-1254

[10] Marasović A, Bhagavatula C, Park J S, et al. Natural language rationales with full-stack visual reasoning: from pixels to semantic frames to commonsense graphs//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Online, 2020: 2810-2829

[11] Park D H, Hendricks L A, Akata Z, et al. Multimodal explanations: Justifying decisions and pointing to the evidence//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 8779-8788

[12] Wu J, Mooney R. Faithful multimodal explanation for visual question answering//Proceedings of the ACL Workshop on BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Florence, Italy, 2019: 103-112

[13] Hendricks L A, Akata Z, Rohrbach M, et al. Generating visual explanations//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 3-19

[14] Zhu H, Togo R, Ogawa T, et al. Interpretable visual question answering referring to outside knowledge//Proceedings of the IEEE International Conference on Image Processing. Kuala Lumpur, Malaysia, 2023: 2140-2144

[15] Yang Q, Li Y, Hu B, et al. Chunk-aware alignment and lexical constraint for visual entailment with natural language explanations//Proceedings of the ACM International Conference on Multimedia. Lisboa, Portugal, 2022: 3587-3597

[16] Zhu D, Chen J, Shen X, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models//Proceedings of the International Conference on Learning Representations. Vienna, Austria, 2024

[17] Ye Q, Xu H, Xu G, et al. Mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178, 2023

[18] Zhang A, Fei H, Yao Y, et al. Vpgtrans: Transfer visual prompt generator across llms//Proceedings of the Annual Conference on Neural Information Processing Systems. New Orleans, USA, 2023: 20299-20319

[19] Dai W, Li J, Li D, et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500, 2023

[20] Li J, Li D, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models//Proceedings of the International Conference on Machine Learning. Honolulu, USA, 2023: 19730-19742

[21] Schwenk D, Khandelwal A, Clark C, et al. A-okvqa: A benchmark for visual question answering using world knowledge//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 146-162

[22] Graves A. Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850, 2013

[23] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model//Proceedings of the Annual Conference of the International Speech Communication Association. Makuhari, Japan, 2010: 1045-1048

[24] Jozefowicz R, Vinyals O, Schuster M, et al. Exploring the limits of language modeling. arXiv preprint arXiv: 1602.02410, 2016

[25] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the Annual Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 5998-6008

[26] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners//Proceedings of the Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2020: 1877-1901

[27] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA, 2019: 4171-4186

[28] Howard J, Ruder S. Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146, 2018

[29] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 2020, 21(140): 1-67

[30] Li X, Yin X, Li C, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 121-137

[31] Singh A, Hu R, Goswami V, et al. Flava: A foundational language and vision alignment model//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 15638-15650

[32] Su W, Zhu X, Cao Y, et al. Vl-bert: Pre-training of generic visual-linguistic representations//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019

[33] Lu J, Batra D, Parikh D, et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks//Proceedings of the Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2019: 13-23

[34] Gan Z, Chen Y C, Li L, et al. Large-scale adversarial training for vision-and-language representation learning//Proceedings of the Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2020: 6616-6628

[35] Tan H, Bansal M. Lxmert: Learning cross-modality encoder representations from transformers//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural

Language Processing. Hong Kong, China, 2019: 5100-5111
[36] Wang J, Hu X, Gan Z, et al. Ufo: A unified transformer for vision-language representation learning. arXiv preprint arXiv: 2111.10023, 2021
[37] Bao H, Wang W, Dong L, et al. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts// Proceedings of the Annual Conference on Neural Information Processing Systems. New Orleans, USA, 2022: 32897-32912
[38] Zellers R, Lu X, Hessel J, et al. Merlot: Multimodal neural script knowledge models//Proceedings of the Annual Conference on Neural Information Processing Systems. Los Angeles, USA, 2021: 23634-23651
[39] Dai W, Hou L, Shang L, et al. Enabling multimodal generation on clip via vision-language knowledge distillation//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Dublin, Ireland, 2022: 2383-2395
[40] Hu X, Gan Z, Wang J, et al. Scaling up vision-language pre-training for image captioning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 17980-17989
[41] Cho J, Lei J, Tan H, et al. Unifying vision-and-language tasks via text generation//Proceedings of the International Conference on Machine Learning. Vienna, Austria, 2021: 1931-1942
[42] Li J, Li D, Xiong C, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation//Proceedings of the International Conference on Machine Learning. Baltimore, USA, 2022: 12888-12900
[43] Wang P, Yang A, Men R, et al. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework//Proceedings of the International Conference on Machine Learning. Baltimore, USA, 2022: 23318-23340
[44] Wang Z, Yu J, Yu A W, et al. Simvlm: Simple visual language model pretraining with weak supervision//Proceedings of the International Conference on Learning Representations. Vienna, Austria, 2022
[45] Alayrac J B, Donahue J, Luc P, et al. Flamingo: A visual language model for few-shot learning//Proceedings of the Annual Conference on Neural Information Processing Systems. New Orleans, USA, 2022: 23716-23736
[46] Awadalla A, Gao I, Gardner J, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390, 2023
[47] Gong T, Lyu C, Zhang S, et al. Multimodal-gpt: A vision and language model for dialogue with humans. arXiv preprint arXiv: 2305.04790, 2023
[48] Li B, Zhang Y, Chen L, et al. Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv: 2305.03726, 2023
[49] Koh J Y, Salakhutdinov R, Fried D. Grounding language models to images for multimodal generation. arXiv preprint arXiv:2301.13823, 2023
[50] Liu H, Li C, Wu Q, et al. Visual instruction tuning// Proceedings of the Annual Conference on Neural Information Processing Systems. New Orleans, USA, 2023: 34892-34916
[51] Chen K, Zhang Z, Zeng W, et al. Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195, 2023
[52] Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023
[53] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models// Proceedings of the Annual Conference on Neural Information Processing Systems. New Orleans, USA, 2022: 24824-24837
[54] Xu C, Xu Y, Wang S, et al. Small models are valuable plug-ins for large language models//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Toronto, Canada, 2024: 283-294
[55] Yang X, Wu Y, Yang M, et al. Exploring diverse in-context configurations for image captioning//Proceedings of the Annual Conference on Neural Information Processing Systems. New Orleans, USA, 2023: 40924-40943
[56] Zhang Z, Zhang A, Li M, et al. Automatic chain of thought prompting in large language models//Proceedings of the International Conference on Learning Representations. Kigali, Rwanda, 2023
[57] Kojima T, Gu S S, Reid M, et al. Large language models are zero-shot reasoners//Proceedings of the Annual Conference on Neural Information Processing Systems. New Orleans, USA, 2022: 22199-22213
[58] Zhao X, Li M, Lu W, et al. Enhancing zero-shot chain-of-thought reasoning in large language models through logic// Proceedings of the Joint International Conference on Computational Linguistics. Taipei, China, 2024: 25
[59] Fang Z, Wang J, Hu X, et al. Compressing visual-linguistic model via knowledge distillation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 1428-1438
[60] Antol S, Agrawal A, Lu J, et al. Vqa: Visual question answering//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 2425-2433
[61] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 740-755
[62] Chen X, Fang H, Lin T Y, et al. Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv: 1504.00325, 2015
[63] Hudson D A, Manning C D. Gqa: A new dataset for real-world visual reasoning and compositional question answering// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 6700-6709
[64] Papineni K, Roukos S, Ward T, et al. Bleu: A method for automatic evaluation of machine translation//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA, 2002: 311-318
[65] Denkowski M, Lavie A. Meteor universal: Language specific translation evaluation for any target language//Proceedings of the Workshop on Statistical Machine Translation. Baltimore,

USA, 2014: 376-380

[66] Lin C Y. Rouge: A package for automatic evaluation of summaries//Proceedings of the Workshop on Text Summarization Branches Out. Barcelona, Spain, 2004: 74-81

[67] Anderson P, Fernando B, Johnson M, et al. Spice: Semantic propositional image caption evaluation//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 382-398

[68] Vedantam R, Lawrence Zitnick C, Parikh D. Cider: Consensus-based image description evaluation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 4566-4575

[69] Li D, Li J, Le H, et al. Lavis: A library for language-vision intelligence//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Toronto, Canada, 2023

[70] Sun Q, Fang Y, Wu L, et al. Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv: 2303.15389, 2023

[71] Bai J, Bai S, Yang S, et al. Qwen-vl: A versatile vision-language model for understanding, localization. arXiv preprint arXiv:2308.12966, 2023

[72] Liu H, Li C, Li Y, et al. Improved baselines with visual instruction tuning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 26296-26306

SUO Wei, Ph. D. candidate. His research interests include computer vision and vision-language understanding.

LV Jia-Qi, M. S. candidate. His research interest is computer vision.

SUN Meng-Yang, Ph. D. candidate. Her research interests are computer vision and vision-language interaction.

LIU Le, Ph. D., associate professor. His research interests are computer vision and intelligent information processing.

WANG Peng, Ph. D., professor. His research interests are computer vision, natural language processing and machine learning.

Background

Visual Question Answering (VQA) is a basic vision- and-language task. It requires a multi-modal model to understand images and natural language questions and infer answers. This paper investigates the task of Natural Language Explanation for VQA ((i. e.,) VQA- NLE), which aims to explain the reasoning process of visual language models by generating natural language sentences. Currently, two main paradigms have been widely studied:

(1) Post-hoc Explanation Method: This paradigm uses a task model to obtain answers for the current task (e. g., VQA) and feeds the task model's output along with the question to an explanation model (e. g., LSTM or GPT). The explanation model then generates an explanation for the task model's decision. However, the post-hoc explanation method completely divides the explanation model and decision model, lacking a direct logical connection between the explanation generation process and the task model's inference process.

(2) Self-rationalization Methods: In this paradigm, VQA-NLE is regarded as a text generation task, which can output the answer and a natural language explanation simultaneously. However, this paradigm cannot generate a faithful explanation due to the interactive effect between the answer and the explanation. Meanwhile, existing unimodal explanation methods suffer from semantic ambiguity problem.

To address these issues, we propose a new Probe- based Multi-modal Explanation method (PME). the PME can generate natural language explanations and object detection boxes based on the original VQA model's inference process, ensuring faithful explanations. Specifically, we acquire the

internal reasoning process by introducing a learnable probe structure, which extracts information from each encoder hidden layer during the forward pass of the model's inference process without affecting it. Additionally, we use a pseudo-labeling method to integrate the VQA-X training set and the GQA training set, obtaining annotations for both question-answer explanations and object detection boxes. Based on the comparison and ablation experiments, it can be found that multi-modal explanation improves the CIDEr score by 4.4% on VQA-X dataset. our proposed PME model achieves a 5.0% and 6.4% on A-OKVQA dataset.

This work was supported by the National Natural Science Foundation of China (No. 62472357) and the National Natural Science Foundation of China (NSFC) under Grants 62102323.