arXiv:2305.04183v1 [cs.CL] 7 May 2023

OpenViVQA：越南语视觉问答任务、数据集与多模态融合模型

Nghia Hieu Nguyen[a,b], Duong T. D. Vo[a,b], Kiet Van Nguyen[a,b], Ngan Luu-Thuy Nguyen[a,b]

[a]信息技术大学信息科学与工程学院，越南胡志明市

[b]越南国家大学，越南胡志明市

摘要

近年来，视觉问答（VQA）因其极具潜力的应用（如智能汽车上的虚拟助理、为盲人提供的辅助设备，或使用自然语言作为查询从文档图像中进行信息检索）和挑战性而受到研究界的关注。视觉问答任务要求方法具备融合问题和图像中的信息并生成恰当答案的能力。神经视觉问答模型在大规模数据集上取得了长足发展，而这些数据集大多针对英语等资源丰富的语言。然而，现有数据集将视觉问答任务局限为答案选择任务或答案分类任务。我们认为，这种形式的视觉问答远不及人类能力，并且只通过选择答案而非生成答案，消除了视觉问答任务中回答环节的挑战性。本文介绍了 OpenViVQA（**开放**领域**越南**语**视觉问答**）数据集，这是首个用于越南语开放式答案视觉问答的大规模数据集，包含与**11,000+** 张图像相关联的**37,000+** 个问答对（问答对）。此外，我们提出了 FST、QuMLAG 和 MLPAG，它们融合图像和答案中的信息，然后使用这些融合特征，像人类一样通过迭代构建答案。我们提出的方法取得了可与 SAAA、MCAN、LORA 和 M4C 等最先进模型相媲美的结果。该数据集[1] 可供研究界使用，以推动研究进展。

[1]https://github.com/hieunghia-pat/OpenViVQA-dataset

提交至爱思唯尔的预印本 *2023年5月9日*

为低资源语言（如越南语）开发更通用的算法，包括 Transformer。

关键词：视觉问答、视觉语言理解、低资源语言、信息融合、多模态表示

---

1. 引言

视觉问答（VQA）是一项信息融合任务，以图像和问题作为输入，要求计算机根据图像中的信息为该问题生成答案 [3]。作为最具挑战性的任务之一，视觉问答近年来吸引了计算机视觉（CV）和自然语言处理（NLP）研究领域的广泛关注。这项任务要求相关方法能够理解给定问题中的语言学概念，然后利用相应图像中的视觉概念来回答问题。视觉问答任务的动机源于从多个来源检索信息的需求，其中包括使用自然语言问题查询视频或文档图像中的信息。

尽管视觉问答自 2015 [3] 以来便已存在，并且针对英语等高资源语言得到了广泛研究，但针对越南语这一低资源语言的视觉问答任务，相关研究仍然很少。2021 年，Tran 等人 [48] 构建了 ViVQA 数据集，这是首个用于越南语视觉问答任务研究的数据集。ViVQA 数据集仅涵盖 VQAv2 数据集的一小部分，因为它是通过半自动方法 [48]构建的。我们认为，由于缺乏验证，且机器翻译结果存在不足， [48] 提出的半自动方法在将英语视觉问答数据集翻译成越南语时，无法确保达到人类水平。因此，ViVQA 数据集无法作为探索和设计实验以及评估视觉问答系统的可靠基准数据集；我们需要一种新颖的人工标注视觉问答数据集，作为越南语视觉问答任务研究的基准。

另一方面，由于大多数视觉问答数据集将视觉问答任务视为分类任务 [15], ，我们认为这种方法并不合理，且与人类的能力相去甚远，因为人类通常可以使用各种形式的自然语言来回答问题，例如词语、短语或完整句子。

为克服上述由数据集和任务所导致的局限性

我们定义了一种不同形式的视觉问答任务，即开放式视觉问答，其中包含开放式问题和开放式答案（第 3.1 节）。随后，基于这一新颖定义，我们介绍了 OpenViVQA（面向越南语的开放域视觉问答）数据集，其中包含 **11,199** 张图像以及由人工标注的 **37,914** 个问答对（问答）。特别是，OpenViVQA 数据集中的图像拍摄于越南，因此对于研究界而言，这些图像是研究和探索在越南拍摄的图像与在越南以外其他地区拍摄的图像之间差异的宝贵资源。这组具有地区特征的图像可以推动开发适合我们数据集的恰当预训练模型。

此外，通过实验，我们表明，以往针对英文视觉问答数据集的方法在处理开放式视觉问答任务时表现不佳，尤其是在我们的新数据集上。因此，我们提出了三种多模态答案生成方法，并展示了它们如何取得更好的结果。这些所提方法可作为初步基线，用于进一步研究我们的数据集，尤其是开放式视觉问答任务。

本文结构安排如下：第 2 节介绍与本研究相关的工作。在第 3 节中，我们描述数据的创建与验证，并展示 OpenViVQA 数据集面临的挑战。随后，我们在第 4 节介绍所提出的三种方法。第 5 节说明我们如何设计实验，以及如何评估基线方法和所提方法的结果。第 6 节解释 OpenViVQA 带来的挑战为何会影响这些方法的结果。最后，我们在第 7 节总结本研究，并提出未来的研究方向。

2. 相关工作

在本章节中，我们简要回顾构建视觉问答数据集的重要研究，包括英语和越南语视觉问答数据集（图1）。随后，我们回顾了对多模态方法设计思路产生影响的研究。

2.1. 视觉问答数据集

2.1.1. 以往的视觉问答数据集

Antol 等人 [3] 推出了 VQAv1 数据集，其中包含 254,721 张图像、764,163 个问题和 4,598,610 个答案，定义了英语视觉问答任务。VQAv1 数据集使用了 MS COCO [31]数据集中的 204,721 张图像，以及 50,000 张额外的抽象场景图像。随着……的发布

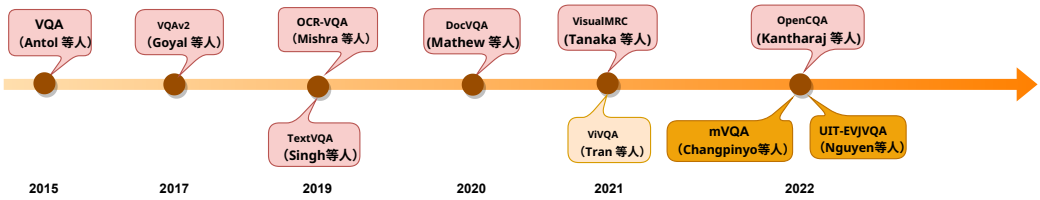


图1：VQA数据集时间线

VQAv1 数据集受到了大量关注，并提出了许多方法[25, 34, 54] 来应对视觉问答任务。

尽管 Antol 等人 [3]旨在通过 VQAv1 数据集引入开放式视觉问答任务，但 Teney等人 [47] 在进行大量估算和统计分析后指出，VQAv1 数据集中的答案可以视为一个类别。因此，他们提出将视觉问答任务作为分类任务来处理：训练机器对答案进行分类，换言之，从一组预先定义的答案中选择答案，而非生成答案。

然而，与其他分类任务一样，类别不平衡现象十分显著，通常会导致机器学习方法的表现过度自信。Goyal 等人 [15] 在观察并对视觉问答方法的结果进行统计分析时发现了这一现象。具体而言，这种现象类似于分类任务中的类别不平衡：模型倾向于针对特定类型的问题给出最常见的答案。例如，当模型收到一个以“有多少”开头的问题时，会立即选择“两个”作为相应的答案，因为在询问数量的问题中，这是出现频率最高的答案。Goyal 等人 [15] 将这一现象称为语言先验，以强调模型在提供答案时对问题的依赖，同时完全忽略图像。

为克服这一现象，Goyal 等人 [15] 重新标注了 VQAv1 数据集，为同一个问题提供针对不同图像的不同答案，从而平衡了特定类型问题中各答案的出现频率，促使模型利用图像中的信息来给出答案。随后，这个新型平衡数据集以 VQAv2 之名发布。事实证明，当时许多最先进的视觉问答方法都受到这种语言先验现象的影响，因此它们在 VQAv1测试数据集上的良好表现并不可靠。之后，基于分类方法以及

注意力机制的指导 [49, 4, 35], 大多数最先进的方法都基于协同注意力策略设计，尝试以融合的方式对问题和图像中的信息施加注意力，然后进行答案选择。

2.1.2. 具有阅读理解能力的视觉问答数据集

尽管视觉问答领域已研究多年，但研究界主要关注解决这类视觉问答任务：问题查询的是图像中物体和上下文的细节。2019年，Singh等人 [43] 指出，视觉问答方法不具备阅读能力，因此推出了一个新数据集——TextVQA 数据集，其中每个问题都利用图像中出现的场景文本。这种对视觉问答任务的新定义吸引了大量研究 [43]。与此同时，还发布了许多类似的数据集，例如 OCR-VQA [37], 、DocVQA [36]和 VisualMRC [46]。

然而，图14、图15和图8中的统计数据及示例表明，这些英文视觉问答数据集的挑战性不如 OpenViVQA 数据集。OpenViVQA 数据集中的答案会将图像中的场景文本作为组成部分，并以类似人类的描述风格提供更详细、具体的信息。相反，英语视觉问答数据集中的答案只是图像中现成的场景文本。

2.1.3. 开放式视觉问答数据集

随着开放式视觉问答研究趋势的发展，也出现了一些类似的数据集，其答案的标注形式与 OpenViVQA 数据集中的答案类似，均采用开放式形式，例如 VisualMRC 数据集 [46] 或 OpenCQA [24]。谷歌发布的另一个数据集是 mVQA[8] ，它是 VQAv2 数据集的多语言版本，其中的答案在翻译成非英语语言后采用开放式形式。

构建开放式视觉问答数据集有诸多优势。开放式视觉问答数据集消除了语言先验现象，这是分类任务中另一种类别不平衡现象。此外，这些数据集中的开放式答案引导研究界研究和探索如何自然、流畅地使用人类语言作答，而不是从预先定义的答案集合中选择答案，从而弥合人类与机器之间的差距。

2.1.4. 越南语视觉问答数据集

尽管已有许多用于研究英语视觉问答任务的基准数据集，但低资源语言，尤其是越南语，尚无可用资源。2019年，Tran 等人 [48] 发布了 ViVQA 数据集，这是首个越南语视觉问答数据集。他们利用机器翻译，将 VQAv2 数据集的一部分问题和答案译成越南语，随后人工验证译文中问题和答案的准确性与流畅度。然而，我们在 8 中列举的示例指出，这种半自动方法无法保证译文句子的质量，因此该数据集不适合作为研究和评估越南语视觉问答模型的可靠基准数据集。为此，我们构建并向 NLP 研究界推出首个大规模人工标注的视觉问答数据集，为后续研究提供全新的基准数据集，并用于验证预训练视觉语言模型在越南语视觉问答任务中的表现。

2.2. 视觉问答方法

视觉问答方法包括三个主要组成部分：外部信息嵌入模块、多源信息融合模块，以及答案分类器或答案生成器模块。目前，视觉问答方法的发展主要集中于改进外部信息嵌入模块和多源信息融合模块。

早期方法使用 ResNet [16] 等预训练图像模型从图像中提取特征 [25, 34, 54]。Anderson 等人 [2] 提出了用于视觉问答任务的自底向上—自顶向下机制，通过 Faster R-CNN[42] 从图像中提取区域特征。这种特征提取方式十分有效，因为它减少了图像中与给定问题无关区域所引入的噪声。Jiang 等人 [23] 开展了实验，证明对于基于注意力的深度学习方法，使用网格特征作为图像的输入特征时，其性能与自底向上—自顶向下注意力机制大致相同。近年来，出现了一些利用信息提取进行视觉问答任务的预训练模型，例如 Oscar [30]或 VinVL [57], ；此类模型显著提升了视觉问答方法在各种数据集上的结果 [30, 57]。

以往的方法使用 FastText [6]或 GloVe [41] 等预训练词嵌入，并结合长短期记忆网络 [17] 提取语言特征。近期研究 [55, 19] 使用 BERT [12]等大型语言模型（LLM）来增强语言特征提取，并取得了积极成果。

除了从多个来源提取信息的方式之外，信息融合也是输出特征中至关重要的一部分，它能够选择或构建适当的答案。大多数视觉问答方法都侧重于利用注意力机制 [35, 4, 49] 来实现信息融合，换言之，它们通过执行注意力机制来确定多个信息来源之间的相关性。根据综述研究 [56], ，信息融合的注意力策略可根据注意力层的数量分为两类：单跳注意力和多跳注意力。以往的研究采用了单跳注意力，例如 [54, 34, 25]，但这些方法并未取得理想的结果，而采用多跳注意力的方法，例如 [55, 58, 60, 59] ，则取得了更好的效果。这些结果表明，单层注意力模块无法恰当地建模视觉问答任务所需的推理能力。近期研究通过采用具有多个 Transformer 层的协同注意力机制 [49] 进一步强化了这一观点，例如 ViLBERT [32], 、VisualBERT [29], 、LXMERT [45],、VL-BERT [44], 、UniCoder-VL [28], 、UNITER [10], 、X-LXMERT [11], 、Pixel-BERT[21],、VLMo [33], 和 SimVLM [52]。

3. OpenViVQA 数据集

OpenViVQA 数据集经过专门设计，体现了问题和答案的开放性，以促进视觉问答任务中答案生成方面的研究。此外，我们的数据集必须利用越南的视觉场景，将其与越南语特性相结合，并清晰地展现我们对越南研究界的贡献。这样，我们还可以用越南语询问这些图像中的场景文本所包含的信息。本节首先详细定义开放式视觉问答任务。随后，我们介绍 OpenViVQA 数据集的创建流程，从图像收集到问题和答案的生成。最后，我们展示数据集的总体统计数据和分析结果。

3.1. 开放式视觉问答任务

先前研究 [3, 15, 24, 46] 使用了“开放式问题”这一术语，却未对其作出明确界定。本文基于语言学研究，对视觉问答任务中的开放式问题和开放式答案作出定义，并详细说明如何确保 OpenViVQA 数据集中的问答均为开放式。

根据 Worley 等人 [53], 的观点，问题的开放性与封闭性取决于两个方面：概念和语法。从概念上说，开放式问题“包含或引发问题本身所含概念中的紧张、冲突或争议” [53]；否则则为封闭式问题。从语法上说，开放式问题要求答案不止一个词，或要求答案为短语。

在概念层面，开放式问题涉及开放的主题，其答案可以来自不同的知识领域和观点（在概念上引发张力、冲突或争议 [53]）。然而，在视觉问答任务中，答案仅限于图像中可获取的上下文或内容。因此，按照概念层面对开放式问题进行定义并不合适。

从语法层面对开放式问题进行定义时，Worley 等人 [53] 从语法角度将视觉问答数据集中的开放式问题定义为需要用一个以上的词语或短语作答的问题 [53]。VQAv2 [15] 等视觉问答数据集被称为开放式问题视觉问答数据集。然而，根据表4，一个词语作答占总答案数的 89,41%，这意味着 VQAv2 中的大多数问题都是封闭式问题，这与先前研究的说法相矛盾 [15]。由此可见，按照答案的语法特征来判断问题是否开放，并不符合先前研究的初衷。后续的视觉问答数据集（如 OCR-VQA 数据集 [37] 和 TextVQA 数据集 [43] ）也存在同样的情况，其中封闭式问题占多数。

为了全面借鉴先前研究中对开放式问题的定义，我们根据问题的语法而非答案的语法来定义视觉问答（VQA）数据集中问题的开放式特征。具体而言，如果某个 VQA 数据集中的所有问题均不具有可由启发式算法轻易识别的共同模式，则该数据集便属于开放式问题 VQA 数据集。基于此，VQAv1 数据集、VQAv2 数据集、TextVQA 数据集或 OpenViVQA 数据集中的问题均属于开放式问题，因为这些问题是通过众包构建的，具有多样化的长度以及复杂的语义依存关系（第3.3.2节）。相比之下，OCR-VQA 中的问题不属于开放式问题，因为它们是按照特定的问题模式定义的（8）。

如果满足以下任一条件，我们将视觉问答数据集定义为封闭式答案视觉问答数据集：

- 所有答案均为词语或短语（确定答案何时属于词语或短语的方法详见第3.3.2节）。
- 从定义好的集合中采样答案（或进行答案选择）的视觉问答方法，比从定义好的集合中依次采样词元来构造答案（或生成答案）的视觉问答方法取得了更好的结果。

不属于封闭式答案视觉问答数据集的视觉问答数据集，就是开放式答案视觉问答数据集。

第一项条件与 Worley 等人[53]研究中对封闭式问题的定义相符。第二项定义则推广了问答（QA）任务中答案的分类方式（例如多项选择问答，其中问答方法必须从 A、B、C 或 D 中选出最准确的答案）。请注意，对于某些问答任务，我们没有提及其提取出的答案，因为这些任务的本质是信息提取，而非问答任务。

为什么要设置第二项条件？根据视觉问答方法的行为来判断答案是否开放，这样合理吗？实际上，我们提出第二项条件是受 Teney等人 [47]的研究启发。在这项研究中，Teney等人以及后续的多项研究 [15, 55, 34, 25] 提出了适用于 VQAv1 [3] 和 VQAv2 数据集 [15] 的视觉问答方法，这些方法从定义好的集合中采样答案。也就是说，例如，如果我们在 VQAv2 数据集上定义概率测度：

$$\mu(a) = \mu(\{a\}) = \frac{\#a}{\#\omega} \tag{1}$$

其中， ω 是 VQAv2 数据集中所有答案的集合， $a \in \omega$ 是一个答案，#a 表示 a 在 ω中出现的总次数，而 #$A = \sum_{a \in A}$(#a) ($A \subset \omega$)；因此，存在答案选择型视觉问答方法，意味着对于大多数 $a \in \omega$， μ(a)非常显著，以至于使用视觉问答方法来近似答案分布，比从单个词元生成答案更有效。这种方法与多项选择问答任务完全相同，在这类任务中，答案是从一个定义好的集合中抽取的。因此，使用视觉问答方法的行为来表示答案的开放性是合理的。

根据我们对开放式答案的定义，正如 Teney等人 [47]所指出的，VQAv1 [3]和 VQAv2 [15] 中的答案属于封闭式答案。另一方面，

我们的实验表明，答案选择视觉问答方法在处理 OpenViVQA 数据集时表现不佳，因此不满足我们对封闭式答案定义中的第二个条件。此外，从表 3可以看出，OpenViVQA 数据集中的答案具有多样的语言层级，这违反了封闭式答案的第一个条件。因此，可以确定 OpenViVQA 数据集中的答案属于开放式答案。

最后，我们定义了由开放式问题及其答案组成的开放式视觉问答数据集。开放式视觉问答任务是由开放式视觉问答数据集定义的视觉问答任务。

3.2. 数据集创建

OpenViVQA 数据集的创建过程可由图2中的以下示意图描述。首先，我们收集图像，然后将其分配到多个子集。接着，我们设计指南，并以此为基础对参与的众包工作者进行训练。训练过程结束后，问答对的创建阶段正式开始。我们为每张图像构成问题和答案对，并将其分为文本问答或非文本问答。下一步是数据集验证和调整，在此过程中，我们检查拼写错误，并确保其符合预期质量的一致性。之后，OpenViVQA 数据集完成，并划分为训练集、开发集和测试集，用于实验。

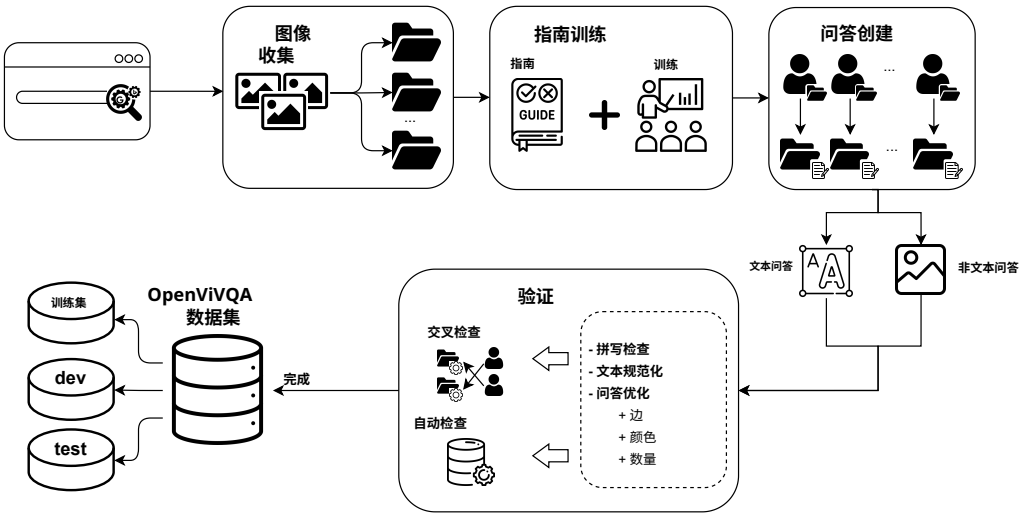


图2：OpenViVQA 数据集的创建总体流程。

3.2.1. 图像收集

在我们的研究中，我们扩充了图像集，使其区别于其他使用英语的相关研究，因为这些研究只利用西方地点的视觉场景 [31, 3], ，无法真正呈现越南景观的鲜明特色。具体而言，源自越南场景的图像通常描绘了熙熙攘攘的街道、摩托车和摊贩，以及人们独具特色的越南生活方式。这些图像有助于使用越南语词汇来提问并描述文化特定概念。此外，由于我们的研究也需要使用越南场景文本，在越南拍摄的图像最适合作为 OpenViVQA 数据集的基础。

我们首先准备一组关键词，以涵盖广泛的概念，例如人类生活方式与活动、交通工具、室内空间、市场、街道和文化遗址。为了增加地理和文化背景的多样性，我们还在部分关键词中添加了越南的地名，例如河内或西贡。例如，部分关键词可以是“河内的餐馆”（quán ăn ở Hà Nội）、“西贡的街道”（đường phố ở Sài Gòn）、“越南市场”（chợ Việt Nam）、“商店招牌”（bảng hiệu cửa hàng）或“学生郊游”（sinh viên đi dã ngoại）。随后，我们将这些关键词输入 Google 图片，并结合 Octoparse [2] 抓取工具和 Python 代码，获取与这些关键词相对应的多个图像收集。

收集完所有图像后，我们开始进行筛选。由于提问时需要保留大部分可见细节，我们只保留分辨率为 500x400 像素或更高的图像。JPEG 或 PNG 以外文件类型的图像也会被排除。此外，我们还会人工检查并剔除损坏或过于模糊的图像，以确保细节清晰可见。最后，所有图像被分配到多个子集中，每个子集包含 100 张图像。

3.2.2. 问答对创建

在这一阶段，我们首先聘用了一定数量且具备足够语言能力的越南语众包工作者。借助众包方式，我们期望获得足够的样本数量和必要的多样性，以确保通过构建真实问题与答案，实现语言表达的多样性以及词汇表的广泛性。

[2] https://www.octoparse.com/

指南具有重要意义，旨在监控数据集的质量。我们要求众包工作者遵循表1中的以下指南。为了确保问题和答案具有开放式特征，我们要求问题更多地采用提取式提问，而不是采用二元或选择题等验证形式。答案也应长于单个词语。此外，我们决定对数量、颜色和方向相关的提问进行控制，因为这些因素在指示和区分物体方面可能发挥重要作用，但在众包过程中，由于语言变异和不一致性，它们也很容易被误判。用于展示这些标准的示例如9所示。

表1：OpenViVQA 数据集问题与答案的创建指南

标准	规则
问答对数量	每张图像至少创建 3 个问答对
答案复杂度	鼓励使用短语或句子形式的答案，而非单个词语。
问题类型	不得是是非题或选择题。
数量	• 请用字母字符而非数字字符书写数量。 • 数量不大于 10。
颜色	• 只能从以下列表中选择：黑色、白色、红色、橙色、黄色、绿色、蓝色、天蓝色、紫色、粉色、棕色和灰色。 • 如果上述颜色无法准确表示物体的真实颜色，请忽略问答中的颜色属性 物体的颜色。
方向 （左侧和右侧）	如果表示方向的词语后面跟着一个物体，则根据该物体来定义方向 ；否则，以观察者的视角来定义方向。

众包工作者需要使用预制工具，为分配的子集中的每张图像创建问题和答案。他们应尽可能为更多图像创建问答对。如果某张图像中的细节较少，导致该图像的问答对数量低于指南规定的最低数量，众包工作者可以创建与现有问答对含义相近的问答对。不过，如果图像过于模糊，且没有任何明确细节，则可以跳过。问题和答案的创建阶段将持续进行，直到所有分配的子集均已完成。

3.2.3. 问答类型分类

为了更好地对 OpenViVQA 数据集中的场景文本属性进行分析和实验，我们决定将数据集中的每个问答对划分为“文本问答”或“非文本问答”。具体而言，非文本问答是指侧重于利用物体信息的问答对，包括物体的属性及其与其他物体之间的关系。例如，任何询问物体或概念指示、物体颜色、数量或位置的问题

都标记为非文本问答。另一方面，如果某个问答对利用了图中出现的场景文本信息本身，或以类似非文本问答的方式利用场景文本询问其他特定物体，我们就将其归类为文本问答。

图 3 展示了非文本问答和文本问答的一些典型示例。对于最左侧的图像（图 3a），众包工作者只询问了一个男孩正在进行的活动，因此这是一个常见的非文本问答案例。另一方面，图 3b 中的问答对则体现了利用图像中的场景文本的情况，具体有两种。第一种情况中，问题借助场景文本（摊位编号）明确指出了所问对象的位置。与此同时，第二种情况中的问题则直接聚焦于出现的场景文本内容，并预期从中提取答案（招牌上显示了虾饺的数量及相应价格）。



问题：穿橙色上衣的男孩在做什么？
答案：正在遛狗
翻译后的问题：穿橙色上衣的男孩在做什么？
翻译后的答案：遛狗

（a）非文本问答



问题：有哪些人正在站在在371号摊位？
答案：有一名男性和一名女性
问题译文：正站在371号摊位？
答案译文：一名男性和一名女性



问题：20,000 越南盾可以买到多少个虾饺？
答案：五个虾饺
翻译问题：用 20,000 越南盾可以买多少个虾饺？
翻译答案：五个虾饺

（b）文本问答

图 3：非文本问答和文本问答的典型示例

问答类型分类阶段需要一组众包工作者积极参与。之后，我们选出两个子集，以便细致地监测 14 名众包工作者之间的分类一致性。这样做是为了量化他们在判断每个问答对属于文本问答还是非文本问答时的相互一致程度。对于这些子集中的所有问答对，我们采用传统的百分比一致率，其计算方法是：将所有标注员均判定为相同类型的问答对数量除以问答对总数。我们还采用 Fleiss' Kappa [13] 来量化分类一致性方面的一致程度，并考虑可能由偶然因素造成的一致。Kappa 值的确定方式如下：

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{2}$$

其中，$1 - P$ 是高于随机水平可达到的一致程度，而 $P - P_e$表示实际达到的高于随机水平的一致程度。该公式源自 P_i 项，即标注员对第 i 个问答的一致程度，以及 p_j 项，即所有分类中属于第 j 种问答类型的比例。

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1) \tag{3}$$

$$p_j = \frac{1}{Nn} (\sum_{i=1}^{N} n_{ij}) \tag{4}$$

其中，N 是问答数量（本例中为 491），n 是标注员数量（14 名标注员），k 是分类类别数量（2 种问答类型）。n_{ij} 是将第 j 种类型分配给第 i 个问答的标注员数量。据此，我们计算出 $\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i$ 和 $\bar{P}_e = \sum_{j=1}^{k} p_j^2$，进而得到 κ 值。

因此，我们得到 Fleiss' Kappa 值为 0.8975，一致率（Percent Agreement）得分为 87.37%。这些结果表明，众包工作者之间具有较高的一致性，且这一一致性显著高于随机水平，同时文本问答与非文本问答之间存在明显区别。

3.2.4. 数据集验证

为确保数据集的质量和一致性，我们会对其进行验证流程，执行全面检查和优化，如图2所示，这是流水线中的步骤之一。

首先，我们为若干众包工作者分配了数据集中的随机子集，并请他们修正所发现的任何拼写或语法错误。为了便于训练过程，我们对数据集中的问答对进行预处理：将文本转换为小写，在词语与标点符号之间添加空格，并规范化场景文本中的价格。对于现实中提问时至关重要、但书写形式多种多样的价格，我们会自动将其转换为如下形式：“#,### 越南盾”，其中 # 是数字字符，逗号是越南广泛使用的千位分隔符，而“越南盾”是越南货币。

3.3. 数据集分析

3.3.1. 初步统计

	图像	文本	非文本
训练集	9,129	13,104	17,729
Dev	1,070	1,733	1,772
Test	1,000	1,766	1,770
总计	11,199	16,643	21,271

表2：图像和问答对的统计。

OpenViVQA 数据集由 11,199 张图像组成，涉及 37,914 个文本问答或非文本问答问答对。我们的数据集的详细信息列于表2中。

如第3.2.3节所述，OpenViVQA 数据集中的问答对分为文本问答和非文本问答。这些文本问答在统计上数量显著，这对视觉问答模型提出了挑战：模型必须具备场景文本阅读能力，才能充分理解图像内容，并找到正确的信息来作答。

此外，正如我们在第3.3.2节中所定义的，OpenViVQA 引入了开放式视觉问答任务，因此其中的问题和答案都是开放式的。我们认为，这类复杂答案会对在 VQAv1 [3]和 VQAv2[15]数据集上取得最先进水平的视觉问答模型构成挑战。为进一步了解细节，我们根据每个答案所含词元的数量，对问题长度（图4）和答案长度（图5）进行了统计。

如图4所示，OpenViVQA 数据集中的问题长度各异，其长度分布比其他视觉问答数据集更平缓。此外，答案长度的分布体现了 OpenViVQA 数据集中答案的多样性和复杂性（图5）。大多数答案的长度在 2 到 10 之间，这表明当前视觉问答方法采用的分类方法不适用于 OpenViVQA 数据集；我们应提出具备答案生成能力的方法，以取得更好的结果。关于这一结论的细节将在我们的实验中展示。

3.3.2. 与其他视觉问答数据集的比较

我们从统计和语言学方面，对 OpenViVQA 数据集与其他类似的英语和越南语视觉问答数据集进行了比较分析

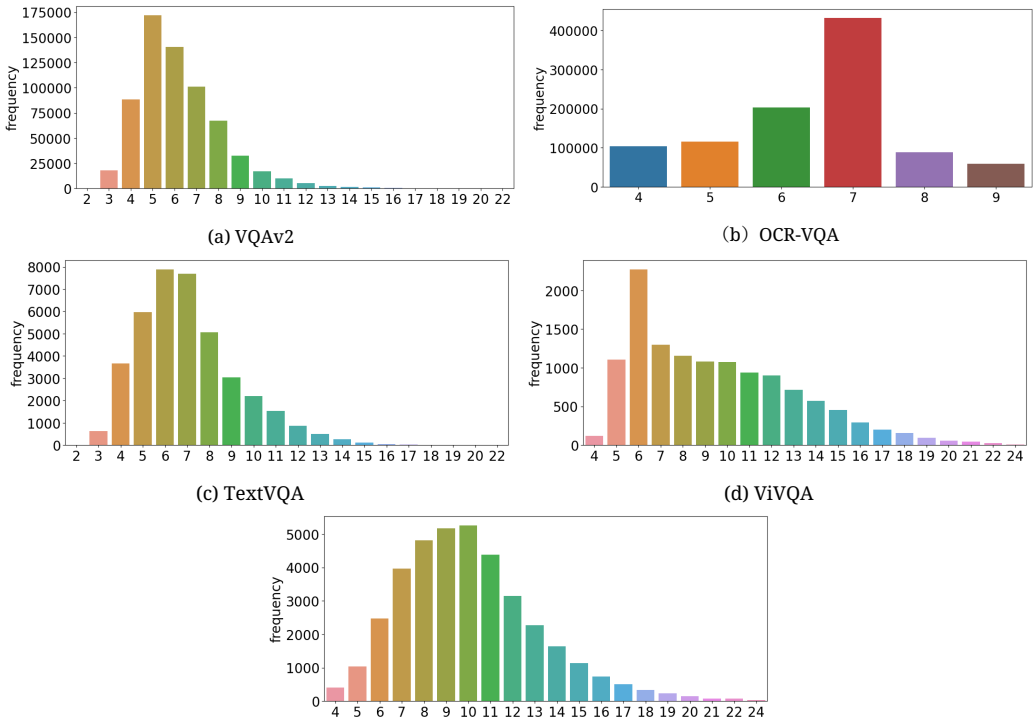
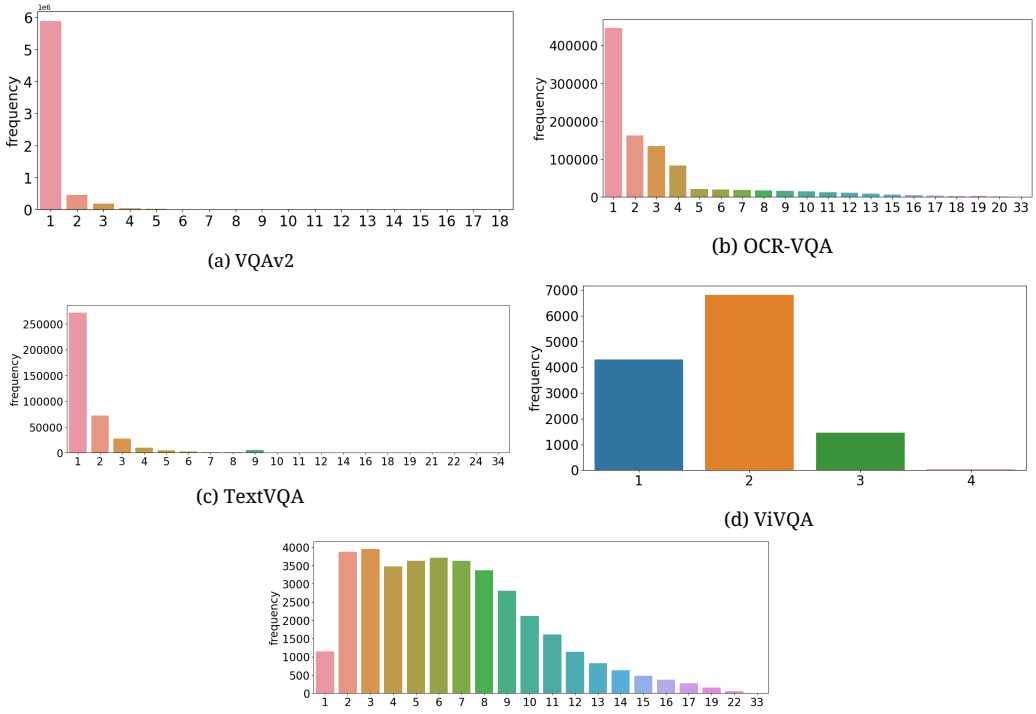


图4：视觉问答数据集中问题长度的比较。

（例如问题长度和答案长度的分布或图像总数）。我们将视觉问答数据集的语言学方面定义为问题和答案的数量、问题或答案中的语义依存关系数量，以及根据语义依存关系构建的语义树的高度。

为了确定给定句子的语言复杂度，我们提出了语言复杂度规范（LCS）算法。首先，LCS 根据各语言相应的依存句法分析器的分析结果，统计给定句子中词元之间依存关系的数量。随后，LCS 利用依存句法分析结果构建语义树，并确定其高度。依存关系总数越多，语义树的高度越高，给定句子就越复杂（图 6）。我们使用 LCS 指出，与其他英语视觉问答数据集相比，我们的数据集中答案的复杂度如何，从而强调了英语和越南语在视觉问答任务中的语言差异，也凸显了专门研究越南语视觉问答的这一新数据集的必要性



(a) VQAv2

(b) OCR-VQA

(c) TextVQA

(d) ViVQA

(e) OpenViVQA

图5：视觉问答数据集中答案长度的比较。

因此，有必要专门构建这一新数据集，以研究越南语视觉问答任务。

此外，我们还依赖依存句法分析器构建了一种算法，用于判定给定文本属于词语、短语还是句子。本文将该算法称为语言层级判定算法（LLS）。该算法基于以下假设：只包含一个词元的文本（越南语中经过分词的词元，或英语中按空格拆分的词元）属于词语；包含一个作为动词 的根词元，并包含一个作为该动词主语的主语 词元的文本属于句子；否则，文本属于短语（图7）。我们使用 LLS 展示人类在回答给定问题时倾向于采用何种句子语言层级，以此强调我们数据集中答案的本质及其开放式特征。

在将 LCS 或 LLS 算法应用于任何视觉问答数据集之前，我们对问题和答案进行了预处理。对于越南语视觉问答数据集，例如

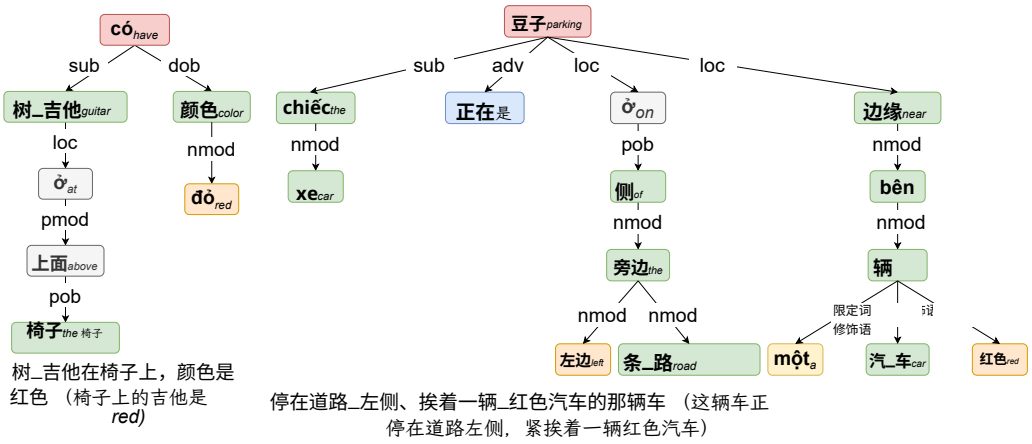


图 6：简单句（左）和复杂句（右）的语义依存关系树。简单句有 6 个依存关系，其语义树的高度为 4；复杂句有 14 个依存关系，其语义树的高度也为 4。

ViVQA [48] 或 OpenViVQA，我们首先使用 VNCoreNLP [51] 对越南语句子进行分词。这是必要的，因为越南语中的词语可能包含两个以上的音节（例如“học sinh”（学生）或“đại học”（大学）），而将词语单独拆分成音节可能会造成混淆，甚至导致对语言学概念的误解。例如，如果我们将二元语法“học sinh”视为由两个独立词语“học”和“sinh”构成，那么这个二元语法就可能被理解为“学习”（học）“生物学”（sinh）。因此，在进行任何进一步分析之前，必须首先进行词语分割。对于英文视觉问答数据集，我们只需使用空格拆分句子即可获得词元。获得经过词元预处理的句子后（越南语为经过分词的句子，英语为拆分后的句子），我们使用 PhoNLP [38] 提供的依存句法分析来构建语义依存关系及其语义树（用于越南语句子），并使用 SpaCy [18] 进行英语句子的依存句法分析。

如表 3 所示，与其他视觉问答数据集相比，OpenViVQA 的答案中依存关系最多，依存树也最深。尽管 TextVQA 和 OCR-VQA 数据集中的答案所含依存关系的最大数量高于 OpenViVQA 数据集，但 OpenViVQA 答案中依存关系的平均数量相当高，甚至是最高的。这表明 OpenViVQA 中人类答案的复杂性，同时也反映了其自然性，尤其是越南语答案的特点。

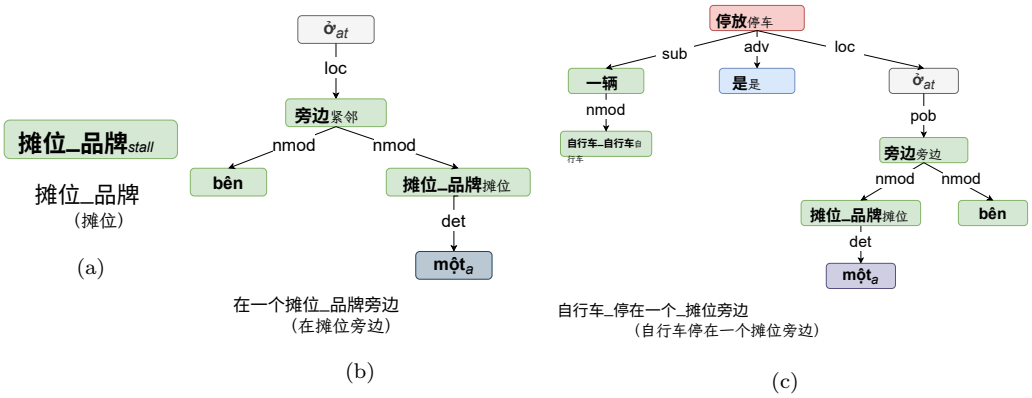


图7：文本的三个语言层级示例：（a）词语，（b）短语和（c）句子。

	数据集	Word			依赖关系			高度		
		最小值	平均值	最大值	最小值	平均值	最大值	最小值	平均值	最大值
问题	VQAv2 [15]	2	6.2	23	2	6.3	26	1	3.3	14
	TextVQA [43]	2	7.1	33	2	7.5	39	1	3.9	21
	OCR-VQA [37]	4	6.5	9	4	6.5	10	2	3.6	6
	ViVQA [48]	3	9.5	24	2	7.3	23	2	5.5	14
	OpenViVQA（我们的方法）	3	10.1	32	2	7.8	27	2	5.2	16
答案	VQAv2 [15]	1	1.2	18	0	2.8	44	1	1.0	11
	TextVQA [43]	1	1.6	85	0	1.5	103	1	1.3	40
	OCR-VQA [37]	1	3.3	74	0	2.8	100	1	1.8	38
	ViVQA [48]	1	1.8	4	0	0.5	3	1	1.5	3
	OpenViVQA（我们的）	1	6.9	56	0	4.8	52	1	4.0	22

表 3：视觉问答数据集在问题和答案方面的语言学比较。请注意，这些结果是在训练-开发集上获得的。

数据集。根据我们的实验结果，我们将表明，这类复杂答案对大多数英文视觉问答数据集上的最先进的VQA方法构成了挑战，而我们的方法（例如，通过生成答案来处理 OpenViVQA 数据集）是有效的。此外，ViVQA 数据集中的语义依存关系数量较少，这意味着该数据集中的答案较为简单。根据 ViVQA 和 OpenViVQA 的这些统计数据，我们证明了[48]中提出的 S 数据集构建方法并不有效，并建议手动构建基准数据集，而不是采用 S 方法，以确保数据质量。

此外，正如我们先前所述，OpenViVQA 数据集中的问答对根据其与图像中场景文本的相关性分为两类：文本问答和非文本问答。因此，一个视觉问答数据集可以同时包含这两类

数据集	#词语	#短语	#句子
VQA2 [15]	5,884,207	651,128	45,775
OCR-VQA [37]	3,287	302,497	15,010
TextVQA [43]	28,317	35,964	4,947
ViVQA [48]	3,276	6,321	0
OpenViVQA（我们的）	1,067	21,022	12,289

表4：视觉问答数据集之间的语言层级比较。请注意，这些结果是在训练-开发集上获得的。

数据集	图像来源	文本问答方法		非文本问答	开放式图像问题			答案
VQAv2$_{en}$ [15]	MS COCO [31]	H	✗	✓	✗	204,721	1,105,904	11,059,040
OCR-VQA$_{en}$ [37]	来自研究 [22]	S	✓	✗	✓	207,572	1,002,146	1,002,146
TextVQA$_{en}$ [43]	OpenImages [27]	H	✓	✗	✓	28,408	45,336	453,360
DocVQA$_{en}$ [36]	加州大学旧金山分校行业文献图书馆	H	✓	✗	✓	12,767	50,000	50,000
VisualMRC$_{en}$ [46]	网页	H	✓	✗	✓	10,197	30,562	30,562
OpenCQA$_{en}$ [24]	网页	H	✓	✗	✓	7,724	7,724	7,724
ViVQA$_{vi}$ [48]	VQAv1 [3]	S	✗	✓	✗	15,000	12,598	12,598
OpenViVQA$_{vi}$	谷歌搜索引擎	H	✓	✓	✓	11,199	37,914	37,914

表5：视觉问答数据集之间的图像、问题和答案数量对比。各数据集名称后的下标表示其语言（H代表人工标注方法，S代表半自动方法）。

文本问答和非文本问答（表5）。尽管 OpenViVQA 的图像-问题-答案三元组数量少于其他英文视觉问答数据集（表5），但它同时包含这两类问答对，并且是规模最大的越南语视觉问答数据集。此外，如表4所示，Open- ViVQA 数据集中约有 36% 的答案是句子；而在其他类似的视觉问答数据集中，例如 TextVQA，句子形式的答案占 7%，OCR-VQA 中则占 4%。与 VQAv2 等视觉问答数据集相比，其答案大多为词语（89.41%），短语占 9.89%，句子所占比例极低（0.7%）。对于 ViVQA 数据集，短语数量是词语的两倍，答案中没有句子。

4. 我们提出的方法

如第3.2节所分析，OpenViVQA 数据集包含长度各异的开放式答案（图5）。我们认为，这些开放式答案具有挑战性，无法采用 SAAA [25], MCAN [55] 和 LoRRA [43] 所设计的分类方法来处理。为证明这一观点，我们提出了三种答案生成方法，它们不仅保留了各自基于分类器的方法的特点，还能够像人类一样给出答案。在接下来的章节中，我们将

M4C 方法和我们提出的三种方法统称为基于生成器的方法，以便于称呼。下面将详细介绍我们提出的架构。

4.1. 通过堆叠进行融合

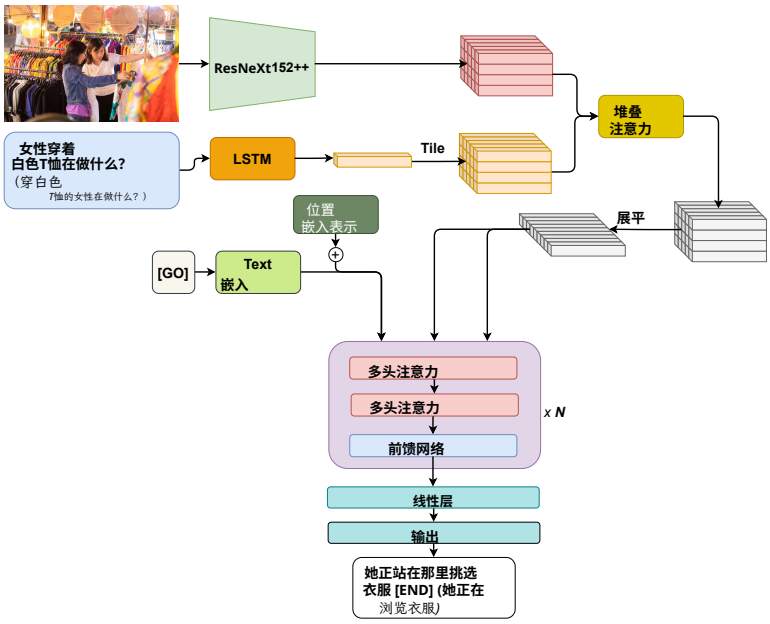


图 8：通过堆叠进行融合（FST）方法。

受 SAAA 启发，我们设计了一种新颖的方法——通过堆叠进行融合（FST），该方法使用堆叠注意力 [54] 融合图像和问题的信息，然后通过在预定义词汇表中迭代选择词元来生成答案。总的来说，FST 包含四个组件：图像嵌入模块、问题嵌入模块、多模态融合模块和答案生成器模块（图 8）。

FST 的图像嵌入使用 ResNeXt152++ [23]从图像中提取特征（尤其是网格特征）。问题嵌入由一个长短期记忆网络 [17] 组成，用于从问题中提取特征。设$x_I \in \mathbb{R}^{s \times d_I}$ 和 $x_Q \in R^{d_Q}$分别为从图像嵌入模块和问题嵌入模块中提取的信息，其中 s 是图像中的空间位置总数。

多模态融合模块由一个与 [25]类似的堆叠注意力模块组成。具体而言，重复 x_Q ，使其形状为 $\mathbb{R}^{s\times d_Q}$。然后通过以下公式计算注意力权重:

$$a = softmax(W_x(W_I x_I^T + W_Q x_Q^T))^T \in \mathbb{R}^{s\times D} \tag{5}$$

其中，为便于说明， $W_x \in \mathbb{R}^{D\times D}$、 $W_I \in \mathbb{R}^{D\times d_I}$ 和$W_Q \in \mathbb{R}^{D\times d_Q}$表示偏置。注意力向量 a可视为堆叠的注意力向量，它们将依次作用于图像特征 x_I。随后，我们得到如下融合特征 x_f :

$$x_f = \sum_{d\in\{1...D\}} (a_d \otimes x_I) \tag{6}$$

其中， $\otimes$ 表示广播乘法。

在答案生成器模块中，答案通过以融合特征 x_f 为条件生成，具体如下:

$$o_t = f(o_0, o_1, ..., o_{t-1} \mid x_f) \tag{7}$$

其中， o_t 是第 t步的输出词元。为对该函数 f建模，我们使用Transformer架构 [49] 的解码器及其掩码技术作为答案生成器模块。

4.2. 问题引导的多模态学习与答案生成

我们假设，图像特征应转换为能够保留足够视觉信息以回答给定问题的特征。随后，将这些特征与问题特征融合，得到用于生成答案的融合信息。为此，我们借鉴 [55]提出的引导注意力（GA）机制，设计了问题引导的多模态学习与答案生成（QuMLAG）模型。QuMLAG由四个组件构成:图像嵌入模块、文本嵌入模块、多模态融合模块和答案生成器模块（图9）。

ImageEmbedding 模块接收从 Faster R-CNN 提取的图像特征（区域特征[23]），然后将这些特征输入一个

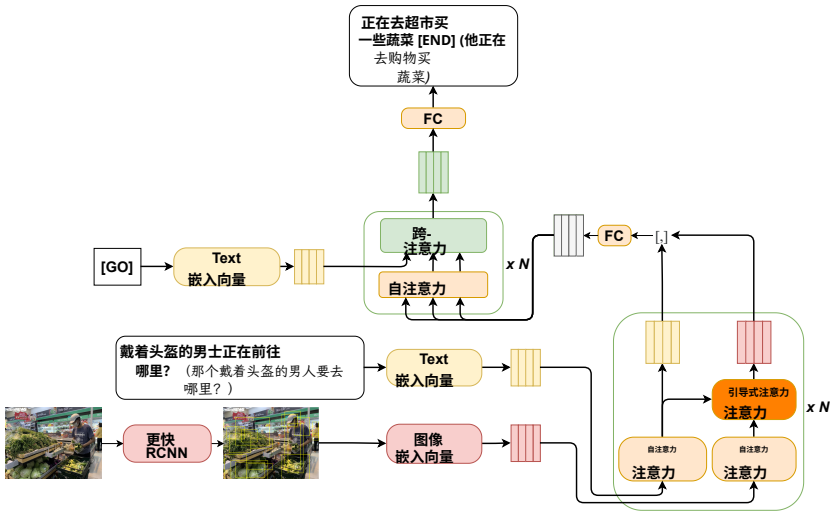


图9：问题引导的多模态学习与答案生成（QuMLAG）方法。

全连接层将其维度投影到隐藏维度 dim，从而生成图像特征 $x_I \in \mathbb{R}^{s \times dim}$，其中 s 表示图像中物体的区域总数。问题嵌入包含一个LSTM网络 [17]。首先，通过FastText将问题嵌入为高维嵌入向量 [6],，然后应用LSTM网络，得到问题的语言特征 $x_Q \in \mathbb{R}^{l \times dim}$，其中 l 表示问题的词元总数。

QuMLAG 的多模态融合模块基于 GA机制设计。具体而言，该模块由三个多头注意力模块组成。第一个多头注意力模块用于将从 Faster R-CNN 提取的图像特征投影到信息的 dim 维潜在空间中，换言之，该模块用于细化 x_I 特征。第二个多头注意力模块的作用与第一个相同，但处理的是问题特征 x_Q。最后一个多头注意力模块用于执行 GA 注意力机制，其中 x_I 充当查询，x_Q 充当键和值。使用第三个多头注意力模块应用问题引导注意力后，特征向量 x_I 会消除无助于回答给定问题的图像特征。然后，将输出特征 x_I 和 x_Q 拼接，得到融合特征 $x_f = [x_I, x_Q] \in \mathbb{R}^{(s+l) \times dim}$。

随后将融合特征向量 x_f 输入答案生成器模块。

与 FST 一样，答案生成器模块旨在实现以下公式：

$$o_t = f(o_0, o_1, ..., o_{t-1} \mid x_f) \tag{8}$$

我们使用了 Transformer架构的解码器模块 [49] 来实现该功能 f 以及 FST。

4.3. 多模态学习与指针增强答案生成器

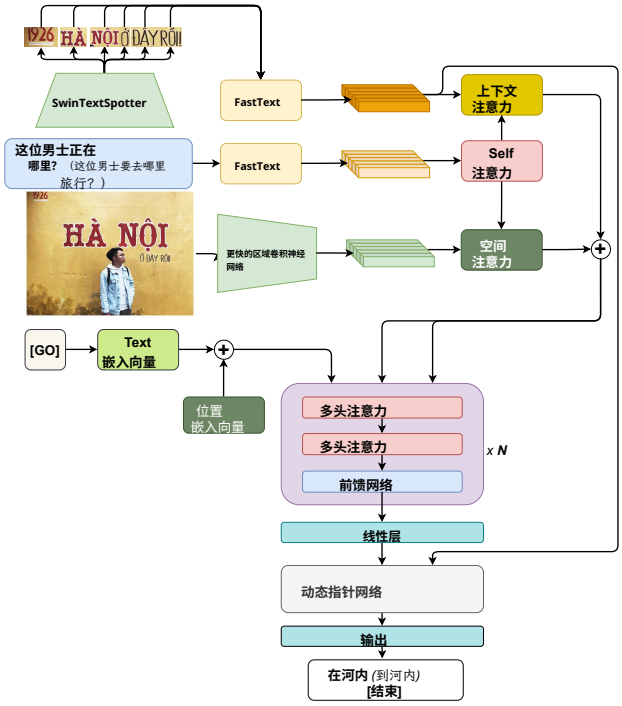


图10：多模态学习与指针增强答案生成器（MLPAG）方法。

我们的目标是设计一种新颖的方法，其设计理念源自 LoRRA，并且能够在迭代构建答案的过程中，动态地从预定义词汇表中选择词元，或从图像中的场景文本中选择内容。该方法称为多模态学习与指针增强答案生成器（MLPAG），采用了与 LoRRA 相同的多模态融合机制 [43]。此外，鉴于 LoRRA 能够读取场景文本并将其用于

答案生成，MLPAG 的设计也经过了精心考量，以延续 LoRRA 的设计理念。具体而言，MLPAG 包含五个组件：图像嵌入模块、场景文本嵌入模块、问题嵌入模块、多模态融合模块和答案生成器模块（图10）。

图像嵌入模块由一个全连接层组成，该层将从 Faster R-CNN[42] 中提取的图像特征（或区域特征 [23]）投影到隐藏维度 dim ，然后得到图像特征向量 $x_I \in \mathbb{R}^{s \times dim}$ ，其中 s 是图像中的区域总数。问题嵌入模块包含一个使用 FastText[6], 的词元嵌入模块和一个全连接层，用于将嵌入式问题特征投影为嵌入特征 $x_Q \in \mathbb{R}^{l \times dim}$ ，其中 l 是问题中的词元总数。场景文本嵌入模块包含一个使用 FastText [6] 将图像中检测到的场景文本嵌入的嵌入模块，以及一个将 FastText 嵌入的特征向量投影到隐藏维度 dim的全连接层，随后得到特征向量 $x_S \in \mathbb{R}^{n \times dim}$ ，其中 n 是检测到的场景文本总数。

MLPAG 的多模态融合模块包含三个主要组成部分：上下文注意力模块、空间注意力模块和自注意力模块。自注意力模块是一个多头注意力 [49] 模块，用于对 x_Q 自身执行自注意力，得到注意力特征 $a_Q \in \mathbb{R}^{l \times dim}$。上下文注意力模块是一个多头注意力模块 [49] ，用于对 x_S 执行关于 a_Q的交叉注意力。空间注意力模块是一个多头注意力模块，用于对 x_I 执行关于 a_Q的交叉注意力。两个向量 x_I 和x_S 分别经过空间注意力和上下文注意力后，消除了与问题无关的视觉信息。随后，融合信息特征由以下方式确定：

$$x_f = x_S \oplus x_I \tag{9}$$

其中， $\oplus$ 是逐元素求和运算符。

随后，我们设计答案生成器模块，以迭代方式生成答案。我们使用 Transformer解码器 [49] 对以下函数进行建模：

$$o_t = f(o_0, o_1, ..., o_{t-1} \mid x_f) \tag{10}$$

此外，为了使 MLPAG 能够将图像中的场景文本复制到答案中，我们为答案生成器模块提供了 DynamicPointerNet-

工作 [19]。具体而言，设 $h \in \mathbb{R}^{l \times dim}$ 为用于生成输出词元 $o \in \mathbb{R}^{l}$的特征向量，动态指针网络按如下方式考虑 h 和 x_S：

$$ocr = (W_h h + b_h)(W_S x_S + b_S)^T \in \mathbb{R}^{l \times n} \tag{11}$$

其中 $W_h \in \mathbb{R}^{dim \times dim}$、$W_S \in \mathbb{R}^{dim \times dim}$、$b_h \in \mathbb{R}^{dim}$ 和 $b_S \in \mathbb{R}^{dim}$。

另一方面，隐藏特征 h经过一个全连接层，投影到词汇表空间 $h' \in \mathbb{R}^{l \times v}$，其中 v 是定义的词汇表大小。然后，最终输出 $o \in \mathbb{R}^{l}$ 由以下方式确定：

$$o = max([h', ocr]) \tag{12}$$

拼接运算符 $[,]$ 沿最后一个维度执行。

5. 实验与结果

5.1. 基线模型

我们将提出的模型与以下介绍的几种强大基线方法进行比较。

•**SAAA**：这是谷歌为 [25] VQAv1 数据集 [3], 提出的一种强基线。SAAA 遵循堆叠注意力机制 [54]，将图像特征与问题特征结合起来，再利用这种组合从预先定义的答案集合中选出答案，因此这是一种用于视觉问答任务的分类方法。

• **MCAN**：MCAN 是针对 VQAv2 数据集 [55], 提出的，旨在利用图像特征引导问题特征，以生成组合特征；随后 [55] 通过降维模块对这些组合特征进行降维，并据此选出合适的答案。

• **LoRRA**：与 TextVQA 数据集 [43], 一同提出。LoRRA 旨在利用图像中可用的场景文本，并从这些场景文本与预定义的答案集合中选择合适的答案。尽管 LoRRA 能够读取图像中的文本，但它本质上是一种用于视觉问答任务的分类方法。

• M4C：由 Facebook 在 TextVQA 数据集 [19] 上提出，该方法是首个将视觉问答任务视为生成任务来处理的方法。M4C 采用 BERT [12] 进行问题嵌入，并使用 BERT [12] 的编码器模块作为整个编码器，即多模态编码器，用于嵌入所有类型的特征，包括图像 [2], 中物体的对象特征或区域特征、来自 BERT 模型的问题特征、图像中可获取的场景文本的视觉特征，以及答案中先前词元的特征。此外，由于 M4C 将视觉问答任务视为生成任务，而非分类任务，因此采用了另一种方式，将图像中的场景文本复制到答案中。具体而言，Hu 等人[19] 设计了动态指针网络，不同于 LoRRA 将图像中的场景文本直接选作答案，该网络将词汇表中所有词元的嵌入特征与场景文本的嵌入特征进行融合，然后由 M4C 在答案生成过程中，于步骤 t 选择应当输出来自词汇表的词元还是场景文本。

5.2. 评估指标

受机器翻译任务和图像描述任务的启发，我们衡量机器生成的答案（候选文本 - hypo）与人工提供的答案（参考文本 - ref）之间的距离。具体而言，我们使用 BLEU（BLEU@1、BLEU@2、BLEU@3、BLEU@4） [40], ROUGE（ROUGE-L） [14],METEOR [5] 和 CIDEr [50]来评估视觉问答模型在我们数据集上的表现。

5.2.1. BLEU

该指标主要取决于精确率指标的取值。Papineni 等人 [40] 根据两项观察设计了 BLEU 指标：（1）候选文本中 n 元标记的出现次数不应超过其在参考文本中的出现次数；（2）长度超过参考文本的候选文本应被赋予较低的权重（惩罚权重）。具体而言，给定参考文本时，候选文本的得分定义如下：

$$score_{token} = \frac{Count_{clip}(token)}{Count(token)} \tag{13}$$

然后基于此公式，将数据集中所有候选文本的得分设计为：

$$p_n = \frac{\sum_{h \in hypothesis} \sum_{token \in h} Count_{clip}(token)}{\sum_{h \in hypothesis} \sum_{token \in h} Count(token)} \tag{14}$$

其中， n 是所使用的 n 元语法的值。

p_n 的公式已经处理了候选文本长度大于参考文本长度的情况，剩下需要考虑的是候选文本长度小于参考文本长度的情况。为处理这种情况，设 c 为数据集中所有候选文本的长度， r 为数据集中所有参考文本的长度，则为长度大于参考文本的候选文本设计的惩罚权重如下：

$$BP = e^{1-\frac{r}{c}} \tag{15}$$

显然，当 $c > r$时，BP $= 1$ 。

最后，BLEU 分数计算如下：

$$logBLEU = min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n log p_n \tag{16}$$

本文使用了 $n \in \{1, 2, 3, 4\}$，分别为 BLEU@1、BLEU@、BLEU@3和 BLEU@4。

5.2.2. ROUGE

ROUGE 与召回率具有相同的特性。Ganesan 等人[14]设计了 ROUGE 指标，用于表示候选文本与参考文本之间共有的 n 元标记数量占参考文本中 n 元标记数量的比例。此外，我们还需要解决如何定义共有词元的问题。本文使用了最常见的 ROUGE 指标类型，即 ROULE-L，它采用最长公共子序列方法（LCS）来确定共有的 n 元标记。

具体而言，Ganesan 等人 [14] 将候选文本与参考文本之间基于 LCS 的召回率 R 和精确率 P定义如下：

$$R_{LCS} = \frac{LCS(hypo, ref)}{m} \tag{17}$$

$$P_{LCS} = \frac{LCS(hypo, ref)}{n} \tag{18}$$

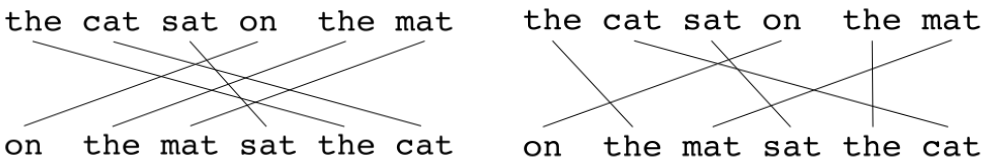
然后，ROUGE-L 定义为：

$$ROUGE = \frac{(1+\beta)^2 R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \tag{19}$$

其中， m 是参考文本的长度， n 是候选文本的长度。

5.2.3. METEOR

BLEU 和 ROUGE 基于 n 元标记来定义词元，而 ME- TEOR [5] 则采用另一种方法来衡量候选文本与参考文本之间的相似度。Banerjee 等人 [5] 认为，许多情况下，即使 n 元标记交换了位置，整个句子的含义也不会改变，但 BLEU 和 ROUGE 指标在这些情况下仍会给出较低的分数。为了解决这一问题，Banerjee 等人 [5]首先定义了候选文本与参考文本之间的对齐 。对齐则定义为一组映射，其中每个映射都表示候选文本与参考文本中的词元之间的一种关联。请注意，在这种情况下，词元被定义为 1 元标记。



(a) (b)

图 11：候选文本与其参考文本之间存在许多对齐。

实际上，特定候选文本与参考文本之间存在许多对齐方式，但选定的对齐方式是交叉点最少的一种。例如，在图 11 中，第一种对齐方式被选为候选文本与参考文本之间的对齐方式。然后，基于二者的对齐方式，候选文本与参考文本之间的精确率 P 定义如下：

$$P = \frac{m}{w_h} \tag{20}$$

候选文本与参考文本之间的召回率 R 定义如下:

$$R = \frac{m}{w_r} \tag{21}$$

那么 P 和 R 之间的相关性为:

$$F_{mean} = \frac{10PR}{R + 9P} \tag{22}$$

与 BLEU 类似，METEOR 会根据候选文本和参考文本的共有词元，设定候选文本长于或短于参考文本时的惩罚权重。具体而言，惩罚权重 p 定义为:

$$p = 0.5 \left(\frac{c}{u_m} \right)^3 \tag{23}$$

其中， c 是候选文本与其参考文本的共同的一元词长度， u_m 是候选文本和参考文本中出现的一元词总数。

最后，通过精确率 P 与召回率 R 之间的惩罚权重和相关性，候选文本与其参考文本之间的 METEOR 分数定义为:

$$M = F_{mean}(1 - p) \tag{24}$$

5.2.4. CIDEr

尽管克服了 BLEU 和 ROUGE 的缺点，METEOR 仍无法考虑候选文本和参考文本的语义相似度。具体而言，Vedantam 等人 [50]指出，METEOR 基于候选文本和参考文本之间的映射来定义对齐，因而会隐式地为所有一元词赋予相同的重量。尽管如此，答案中包含与查询信息完全无关的词元的情况并不少见；候选文本中也常常包含一个完全匹配的词元，使其更接近参考文本，但 METEOR 对待这个词元的方式却与其他词元相同。Vedantam 等人 [50] 指出，计算候选文本与参考文本之间的距离时，应为候选文本中的这类关键词元赋予更高的重量，因此他们提出了能够考虑候选文本和参考文本语义相似度的 CIDEr。

更多细节请参阅：CIDEr 基于两项观察构建：（1）参考文本中的 n 元语法不应出现在候选文本中；（2）在许多图像中同时出现的 n 元语法所携带的信息与任何特定图像都无关。为对这两项观察进行建模，Vedantam 等人

[50] 利用了词频和逆文档频率（TF-IDF）。设 $h_k(s)$ 为 n 元语法 w_k 在句子 s 中的总出现次数，则有：

$$g_k(ij) = \frac{h_k(r_{ij})}{\sum_{w_l \in \Omega} h_l(r_{ij})} log\left(\frac{|\mathrm{Q}|}{\sum_{Q_p \in Q} min(1, \sum_q h_k(s_{qp}))}\right) \tag{25}$$

其中，$g_k(ij)$ 是问题 j 的参考文本 i 中 n 元语法 w_k 的 TF-IDF 权重，Ω 是所有 n 元语法的集合，Q 是数据集中所有问题的集合。

确定所有 n 元语法的 TF-IDF 权重后，候选文本及其参考文本的 CIDEr 分数根据二者的余弦相似度确定：

$$CIDEr(h_i, r_{ij}) = \frac{1}{m} \sum_j \frac{g^n(h_i) \cdot g^n(r_{ij})}{g^n(h_i) \cdot g^n(r_{ij})} \tag{26}$$

5.3. 特征提取

5.3.1. 图像特征提取

本文使用 Faster R-CNN [42] 和 ResNeXt152++ [23]作为其骨干网络，为 LoRRA、FST、MCAN、QuMLAG 和 M4C 从图像中提取区域特征 [2]。我们还遵循 [23]，使用 ResNeXt152++为 SAAA 和 FST 提取网格特征。

5.3.2. 场景文本特征提取

Hu 等人 [19] 使用 Rossetta 系统 [7] 获取图像中场景文本的边界框，然后使用 ROI 对齐，从 Faster R-CNN [42]的骨干网络特征图中为每个场景文本提取区域特征，并将这些特征用作描述图像中场景文本外观的特征。此外，他们使用 PHOC [1]提取场景文本的字符外观或细节外观，并使用 FastText [6] 获取其语言含义。最后，他们将场景文本的上下文外观特征、细节外观特征和语言含义 [19] 结合起来，生成场景文本的上下文特征。

然而，Rossetta 系统并未公开，且不支持越南语。因此，我们必须使用其他在越南场景文本数据集（如 VinText [39]）上训练的场景文本模型。具体而言，我们使用 SwinTextSpotter [20] 提取边界框以及场景文本的特征。

5.4. 实验配置

每种方法都有各自的一组超参数和训练配置，因此，本章节将详细介绍各基线方法和提议生成器方法的超参数及相关配置。总体而言，我们使用大小为 64 的批量数据训练了所有基线方法和所提方法。对于基于分类器的方法，我们将学习率固定为 0.01，并使用交叉熵损失作为目标函数。需要注意的是，在基于分类器的方法的原始配置中，[25, 55, 43],作者使用的是多标签交叉熵损失，而非交叉熵损失。然而，我们的实验表明，使用多标签交叉熵损失并不起作用，大多数基于分类器的方法在所有评估指标上的结果都收敛至 0。对于基于生成器的方法，我们借鉴了 Transformer [49]训练过程中的学习率调度器。具体而言，[49] 中使用的学习率调度器形式如下：

$$learning_rate = d_{model}^{-0.5} \times min(step_num^{-0.5},\ \ step_num \times warmup_steps^{-1.5})$$

在我们的设置中，我们将 $warmup_steps = 10,000$。我们使用提前停止技术：如果 CIDEr 分数在 5 个训练轮次后没有提升，就停止训练过程。所有实验均使用 Adam [26]作为优化方法。所有实验均在 A100 GPU 上完成。各方法的具体配置如下。

5.4.1. 基线模型的实验设置

SAAA：SAAA 的大部分训练配置均与[25] 保持一致，唯一的例外是使用 ResNeXt$152{+}{+}$ [23]提取图像的网格特征，具体如第 5.3.1 节所述。

MCAN：对于信息融合模块，我们为 GA 模块使用了 4 层，SA 模块也使用了 4 层。GA 和 SA 中的每个注意力模块均有 8 个头，隐藏维度为 512。此外，由于 OpenViVQA 数据集中的问题为越南语，我们使用 FastText [6]为问题词元生成词嵌入，而非像[55]中那样使用 GloVe [41] 。

LoRRA：我们实验中的 LoRRA 保持了 [43] 中提出的原始版本。不过，我们对从图像中提取的场景文本以及问题词元使用 FastText [6] ，而不是对问题词元使用 GloVe [41]。图像中的场景文本使用 SwinTextSpotter 识别

[20]。信息融合模块中的上下文注意力模块、自注意力模块和空间注意力模块均有 8 个头，隐藏维度大小为 512。

M4C：在实验中，我们保留了 M4C 的所有原始配置，包括使用 BERT-based-uncased [12] 作为问题嵌入模块，以及使用 BERT 的编码器模块作为多模态模块。问题嵌入模块中的 BERT 模型从其预训练权重加载，并在 OpenViVQA 数据集上训练时进行更新。

5.4.2. 所提出模型的实验设置

FST：我们使用 ResNeXt152++ [23] 从图像中提取网格特征。问题嵌入模块中的长短期记忆网络 [17] 包含 1 层，隐藏维度为 512。融合模块的堆叠注意力模块使用 2 个注意力视角来执行堆叠注意力。对于答案生成器模块，我们使用具有 3 层的 Transformer [49] 解码器。每层都包含一个执行自注意力机制的注意力模块和一个执行交叉注意力机制的注意力模块。每个注意力模块有 8 个头，隐藏维度为 512。

QuMLAG：我们实验中的 GMCAN 使用 FasterRCNN [42] 和 ResNeXt152++ [23] 作为骨干网络，从图像中提取区域特征。GMCAN 的文本嵌入模块也使用 FastText [6] 代替 GloVe [41] ，以适应越南语。信息融合模块包含 4 层 GA 模块和 4 层 SA [55]模块。GA 和 SA 中的每个注意力模块均有 8 个头，隐藏维度为 512。答案生成器模块包含 3 层，每层各包含一个用于自注意力的注意力模块和一个用于交叉注意力的注意力模块。每个注意力模块均有 8 个头，隐藏维度为 512。

MLPAG：MLPAG 使用 FasterRCNN [42] 和 ResNeXt152++ [23] 提取图像的区域特征。图像中的场景文本使用 SwinTextSpotter [20]进行识别。场景文本和问题均使用 FastText [42]进行嵌入。MLPAG 的答案生成器模块所使用的解码器包含 3 层，每层有两个注意力层，其中一层执行自注意力，另一层执行交叉注意力。每个注意力层包含 8 个头，隐藏维度为 512。

5.5. 实验结果

我们评估了基线方法和我们提出的三种方法，并在表 6 中报告了它们的结果。

方法	BLEU@1	BLEU@2	BLEU@3	BLEU@4	METEOR	ROUGE	CIDEr
SAAA	0.0091	0.0043	0.0001	0.0000	0.0533	0.0605	0.1230
MCAN	0.2366	0.1808	0.1446	0.1179	0.1531	0.2718	1.0613
LoRRA	0.1923	0.1430	0.1128	0.0917	0.1271	0.2137	0.8005
M4C	0.3737	0.2755	0.2078	0.1597	0.2107	0.3789	1.5073
FST（我们的方法）	0.1710	0.1154	0.0813	0.0582	0.1157	0.2010	0.6141
QuMLAG（我们的方法）	0.3548	0.2863	0.2157	0.2037	0.2129	0.3786	1.7082
MLPAG（我们的方法）	0.3811	0.2927	0.2328	0.1889	0.2230	0.3772	1.6104

表6：基线方法和所提方法在 OpenViVQA 数据集上的实验结果。

根据表6，我们提出的方法取得了优于所有基线方法的结果（图 12）。这些结果证明，基于分类器的方法在 OpenViVQA 数据集上表现不佳，我们认为这类方法在开放式视觉问答数据集上通常也无法取得良好效果。具体而言，FST 仅优于同样采用堆叠注意力机制的 SAAA。问题引导注意力优于堆叠注意力；即使 QuMLAG 不读取和使用场景文本，也能在 BLEU@4、ROUGE 和 CIDEr 指标上取得比 MLPAG 更高的分数。尤其是，M4C 的参数量最多，但其结果却低于我们提出的方法。我们将在第6.4节中对 M4C 进行详细讨论。

6. 结果分析

6.1. 问题和答案长度对实验结果的影响

直观而言，我们可以使用 LCS 或 LSS算法（第 3 节），通过语义依存关系或语义树的总数来分析实验方法的结果。然而，这种方法并不具有普遍性，因为它侧重于语言的语言复杂度。我们认为，根据句子长度进行分析更自然，因为人们不会关注自己说出的词语有多复杂，而会关注自己说出的句子有多长。通常，喜欢用词语或短语回答给定问题的人，说话速度比用句子礼貌作答的人更快。此外，如图13b所示，句子越长，其语言复杂度就越高。这很直观，因为更多词元会带来更多依存关系

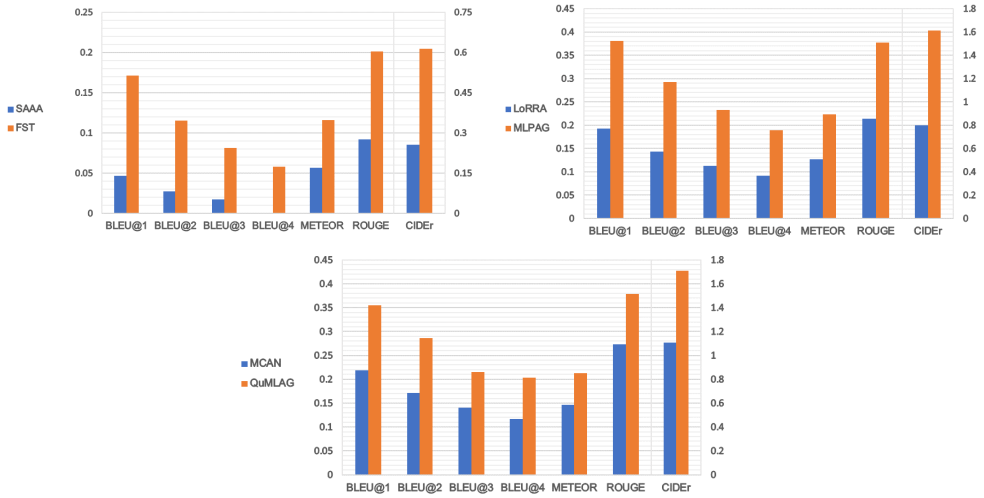


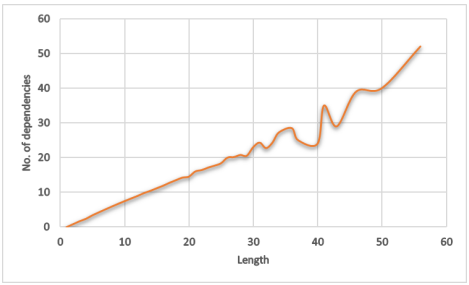
图 12：我们所提方法的结果优于各自的基线方法。

在词元之间产生。因此，根据句子长度分析结果更全面，因为这既涵盖了人们使用语言时的习惯，也涵盖了句子的语言复杂度。

为便于分析，我们根据问题和答案的长度（即所含词元的总数），将它们划分为不同的组。本章节中的词元与第 3 节中的词元采用相同的方式获取，即使用 VnCoreNLP[51] 进行分词。更具体地说，我们将组定义为

组	**长度（n）**
短（S）	$n \leq 5$
中（M）	$5 < n \leq 10$
长（L）	$10 < n \leq 15$
超长 (XL)	$n > 15$

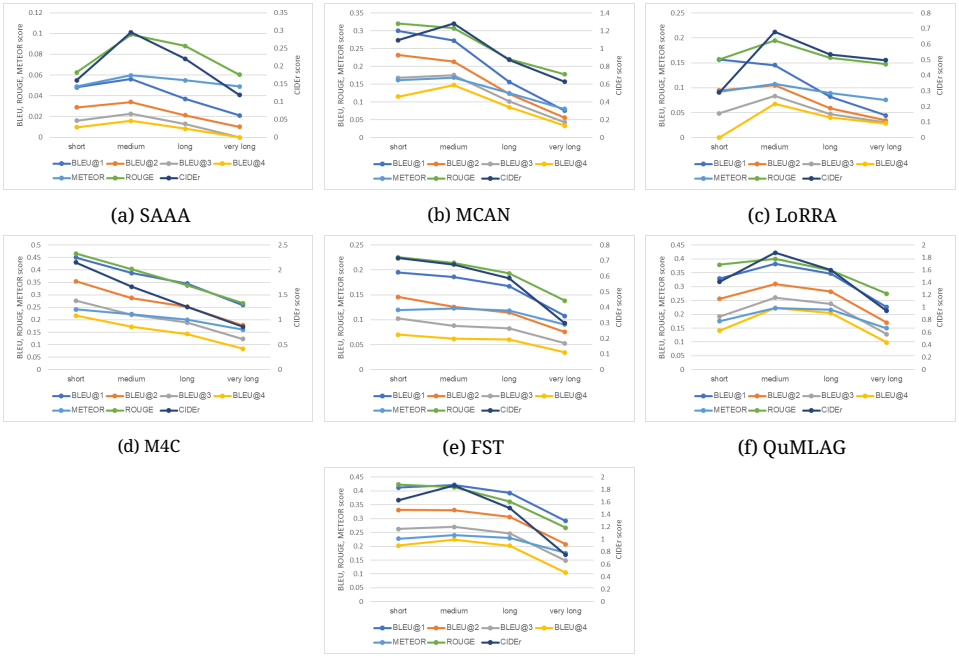
(a)



(b)

图 13：基于词元总数的问题和答案组（a），以及基于句子长度的依存关系总数（b）。

如图13a所示的问题组和答案组。根据这一划分，我们评估了基线方法和所提方法在每组问题和答案上的表现。首先，我们考察基线方法和所提模型在各问题组上的结果。细节见图14。



(a) SAAA (b) MCAN (c) LoRRA

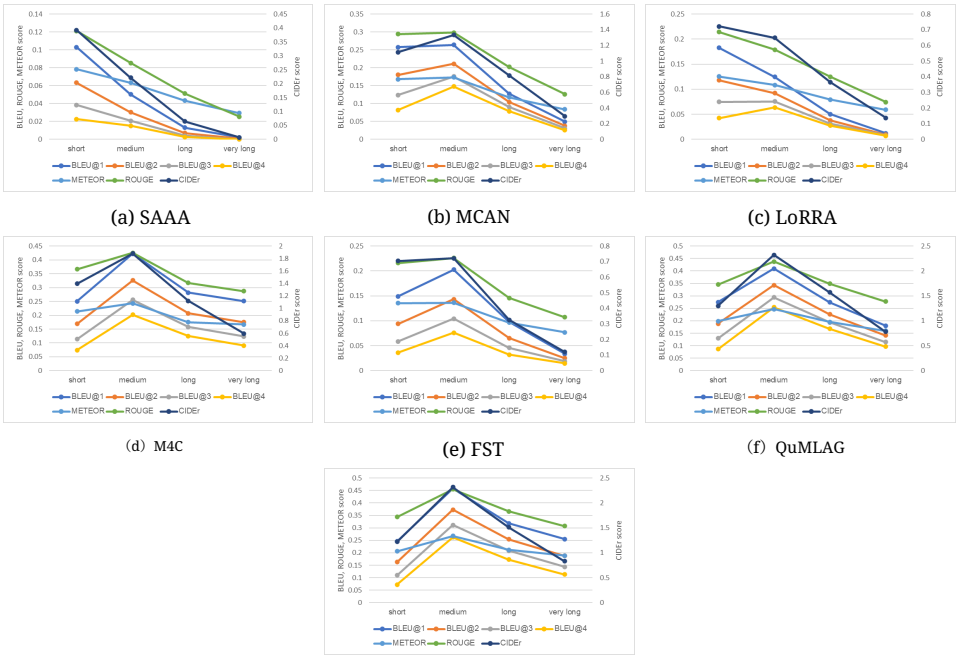
(d) M4C (e) FST (f) QuMLAG

(g) MLPAG

图14：不同基于问题长度的分组上的实验方法结果。

根据图 14，大多数基于分类器的方法在多数指标上，对M组问题取得了最高的结果。此外，基于分类器的方法的 BLEU分数相对较低，甚至低于 0.2。这些分数表明，基于分类器的方法生成的答案与真实答案匹配度较低，因此其结果不可靠，难以据此进行解读和总结。另一方面，基于生成器的方法在S组问题上表现最佳，随后随着问题长度增加，表现逐渐下降；但QuMLAG例外，它在M组问题上表现最佳，而MLPAG在S组和M组问题上的结果大致相同，但面对更长的问题时，其得分会下降。这些结果表明，基于生成器的方法更容易处理S组和M组问题，因为它们是数据集中最常见的问题。然而，这也表明这些方法

难以阅读和理解越南语中的长问题，而这些问题的语言复杂度与相应句子的长度成正比（图13b）。



(a) SAAA (b) MCAN (c) LoRRA

(d) M4C (e) FST (f) QuMLAG

(g) MLPAG

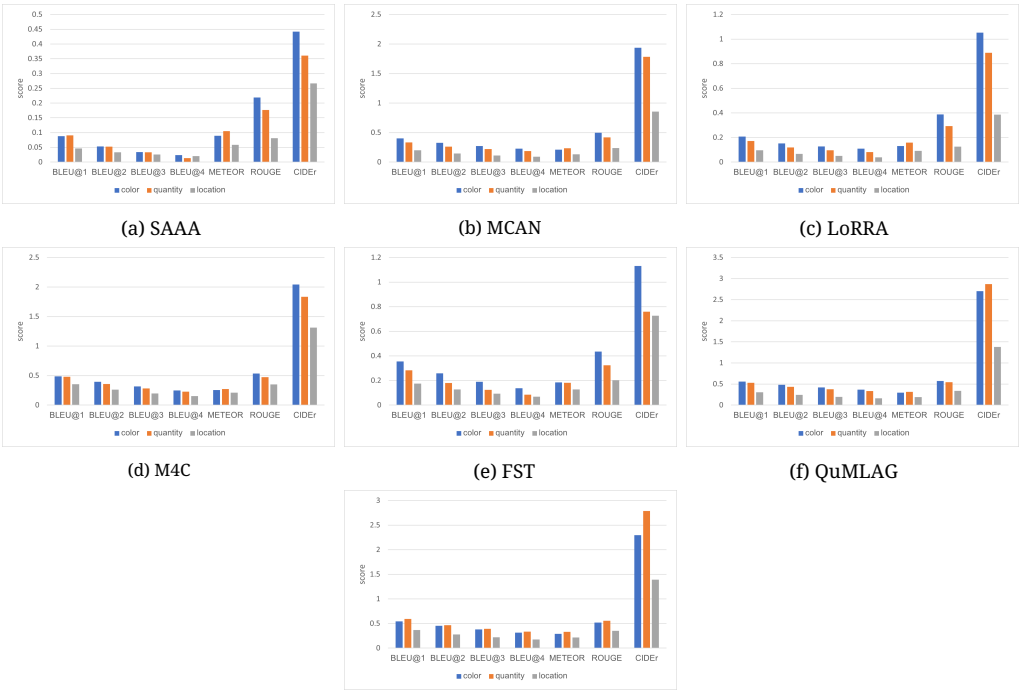
图15：实验方法在基于答案长度的各组上的结果。

转到答案组，与基于问题长度的分析类似，基于分类器的方法的结果较低，尤其是在 BLEU@2、BLEU@3 和 BLEU@4 得分方面，这导致对其结果进行分析和解释并不可靠。因此，我们将重点放在基于生成器的方法的结果上。根据图15，所有基于生成器的方法在所有答案组上的结果都呈现出相同的模式，并且它们在 M组答案上均取得了最佳得分。从图5可以看出，大多数答案属于 M组。这种分布直接影响了这些方法的结果，因为它们主要接受训练以生成 M组答案，从而产生了图15所示的结果。另一方面，基于生成器的方法难以生成更长的答案，例如 L组和 XL组中的答案。这表明，这些方法在用越南语生成长答案时存在困难。

令人惊讶的是，尽管 S 组答案在答案分布中占有相当大的比例，但基于生成器的方法在这些答案上的表现不如在 M 组答案上的表现，尽管通常而言，S 组答案比其他组的答案更容易构造。为了解释这一现象，我们研究了基于生成器的方法生成的答案与真实答案的匹配情况。我们发现，在答案较短的问题中，基于生成器的方法生成的答案约有 9.29% 的词元重复自问题；而在答案为中等长度的问题中，这一比例为 13%。对于答案较长和超长的问题，这一比例分别为 13.13% 和 10.87%。这表明，对于 M 组答案，基于生成器的模型在开始构造答案时，倾向于重复问题中的一些首部词元，就像越南人通常回答问题时那样（例如“**chiếc xe đạp đang được dựng ở đâu?**”（自行车靠在哪里？） ⇒ “**chiếc xe đạp đang được dựng ở bên cạnh cái giếng**”（自行车靠在井旁））。因此，基于生成器的方法在 M 组答案上容易取得最高成绩，因为它们与真实答案有许多匹配的词元。然而，对于较长的答案，由于与人工提供的答案相比，生成的答案缺乏信息，其表现会下降。此外，短答案主要由词语或短语构成。然而，基于生成器的方法倾向于生成 M 组答案。这种基于生成器的方法的答案构造策略不仅会造成影响整体表现的冗余词元，还可能在答案中对物体、位置、数量、颜色或场景文本提供错误信息，因此在 S 组答案上的得分较低。

6.2. 问题类型对实验结果的影响

为了分析模型在问题的不同目标内容上的表现，我们在此考察模型对若干差异较大的方面（如颜色、数量和位置）相关问题的预测。我们根据问题可能涉及的这些方面来定义问题类型。此项分析旨在观察模型如何关注视觉细节，并将描述性词语融入期望的答案中。我们首先使用正则表达式，为问题类型分类构建基于规则的算法。对于每种问题类型，我们都像前面各章节中那样，计算每种实验方法的评估分数。如图16所示，所有模型都呈现出相同的规律：与颜色和数量相关的问题相比，关于位置的问题更具挑战性



(a) SAAA (b) MCAN (c) LoRRA

(d) M4C (e) FST (f) QuMLAG

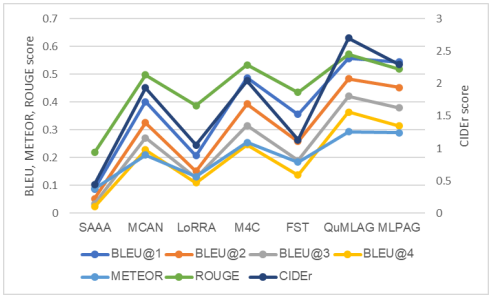
(g) MLPAG

图16：各实验方法在不同问题类型上的结果。

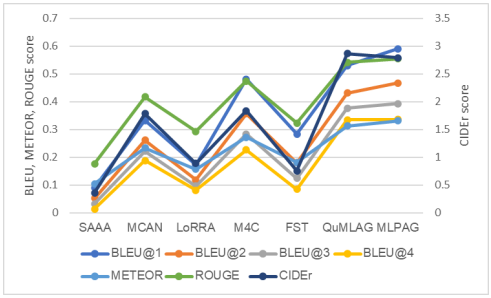
这可能是因为位置信息更为复杂，而颜色和数量则更容易理解。

在图17中，我们比较了各实验方法在不同问题类型上的得分。具体而言，基于Transformer架构的更复杂模型，在描述回答这些类型问题所需的细节方面表现更好。总体来看，各模型在不同问题类型上的表现顺序相似，其中生成式方案的得分高于原始模型。不过，这些发现与第5.5节的主要结果成比例，且可能混杂在实验方法给出完整答案时的总体表现中；这些总体表现并未说明模型在答案中准确生成用于诠释颜色、数量或位置的具体词语的能力。

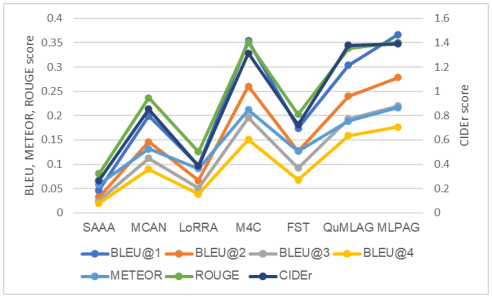
为了分析实验方法对各类问题给出答案的正确性，我们建议采用准确率进行衡量，其定义如下：



(a) 颜色



(b) 数量



(c) 位置

图17：不同问题类型下实验方法的结果。

$$Accuracy_{type} = \frac{Number_of_correct_terms_{type}}{Total_questions_{type}} \tag{27}$$

我们使用这一指标来衡量给出与标准答案中相同术语的比例。对于关于颜色的问题，我们会追踪预测中的颜色词，并将其与相应标准答案中的颜色词进行比较。对于关于数量的问题，我们也采用相同的做法；但对于位置问题则不然，因为文本呈现中的这一因素变化很大。因此，衡量颜色和数量预测的正确性更为可行，也足以观察模型在关注物体视觉属性方面的表现。图18中的图表显示，M4C在颜色和数量预测方面优于其他模型。M4C对模态间和模态内关系的建模 [19] 使其能够有效地将视觉信息与问答中词语之间的搭配联系起来，从而利用视觉信息和文本信息，预测颜色和数量等细节。另一方面，

基线方法的生成式改进显著提升了对这类术语的预测准确率。这表明，在经过修改以生成更长的答案后，实验方法往往能在答案中更准确地使用这些术语，这证明注意力生成器模块在最终阶段提升预测质量和细节方面发挥着重要作用。

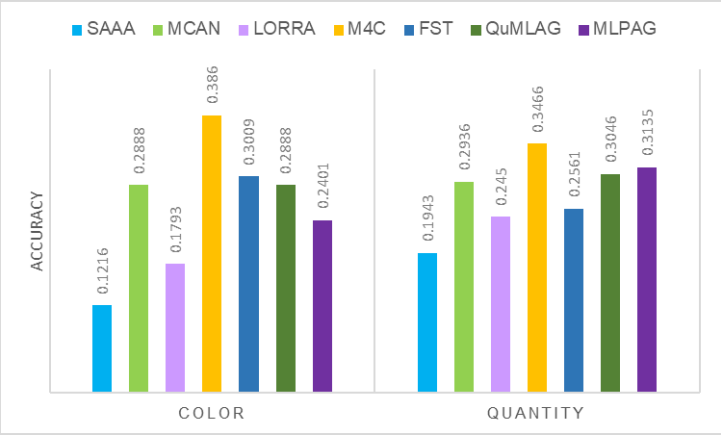


图18：实验方法针对每类问题在预测中给出正确术语的准确率。

6.3. 场景文本图像对实验结果的影响

从逻辑上讲，M4C 和 MLPAG 在其解码器模块中采用了动态指针网络 [19]，换言之，它们同样具备在答案中使用场景文本的能力，因此与 SAAA、MCAN、FST 和 QuMLAG 等无法读取文本的方法相比，理应取得更好的结果。然而，从表6可以看出，M4C 和 MLPAG 的实际结果甚至低于无法读取和使用场景文本的 QuMLAG，尤其是 MLPAG 的结果甚至优于 M4C。为了分析这些结果，我们首先观察了不同视觉问答数据集之间答案特征的差异。根据图5，尽管 TextVQA 数据集被宣称包含开放式问题和答案，但其答案分布与其他非文本问答视觉问答数据集（如 VQAv2）具有相同的特征，这意味着 TextVQA 中的大多数答案长度为1、2或3个词语。或者简单来说，在 TextVQA 数据集中，场景文本是从图像中复制出来作为答案的。因此，M4C 方法在接受该数据集训练时，



图19：我们所提的三种方法和M4C的若干示例。绿色词元是从图像中的场景文本复制而来的。

TextVQA 数据集往往会学习如何复制图像中的场景文本并将其用作答案，而不是利用场景文本在答案中提供详细信息。在 OpenViVQA 数据集上训练时，对于开放式答案，该方法必须具备判断要“复制”哪些可用场景文本，以及将其“粘贴”到答案中哪个位置最合适的能力。因此，复杂性有所增加，而动态指针网络实现的复制机制未能达到预期效果。

对于 FST 和 QuMLAG 等非阅读型架构，从图19可以明显看出，这两个模型无法读取图像中的场景文本。因此，它们生成的答案与给定的问题和图像无关。此外，MLPAG 和 M4C 也表现出相同的行为。根据图

19，这两个模型在答案中使用场景文本方面表现不佳。场景文本的使用频率较低，而且通常出现在答案的末尾。这些结果证明，针对 M4C 提出的动态指针网络 [19] 适用于 TextVQA 数据集中的简单答案，即整个答案通常是从图像中复制的场景文本，采用常规自然语言表达。然而，该模块无法很好地解决如何利用场景文本为给定问题提供额外信息的问题。我们需要为 OpenViVQA 数据集面临的这一挑战寻找更好的解决方案。

6.4. 为什么 *QuMLAG* 和 *MLPAG* 优于 *M4C*?

		BLEU@1	BLEU@2	BLEU@3	BLEU@4	METEOR	ROUGE	CIDEr
文本问答	FST	0.1020	0.0674	0.0445	0.0302	0.0809	0.1331	0.3566
	QuMLAG	0.3064	0.2408	0.1963	0.1649	0.1845	0.3222	1.2504
	MLPAG	**0.3589**	**0.2710**	**0.2115**	**0.1701**	**0.2075**	**0.3604**	**1.3758**
	M4C	0.3315	0.2375	0.1730	0.1303	0.1962	0.3547	1.3191
非文本问答	FST	0.2514	0.1743	0.1259	0.0909	0.1559	0.2747	0.8638
	QuMLAG	0.4090	0.3371	0.2870	**0.2478**	0.2408	**0.4354**	**2.1359**
	MLPAG	**0.4464**	**0.3565**	**0.2942**	0.2449	**0.2534**	0.4184	1.9377
	M4C	0.3960	0.3001	0.2341	0.1838	0.2252	0.4025	1.6545

表7：三种所提方法与 M4C 在文本问答和非文本问答上的比较。

从表6中，我们观察到了一些异常结果：尽管 M4C 在 TextVQA 和 OCR-VQA 等英语数据集上的表现更好，但其结果却低于 QuMLAG 和 MLPAG。如表7所示，在文本问答对上，MLPAG 在所有评估指标上都取得了最佳结果，而 M4C 的表现较差。然而，这两种模型的答案生成器模块采用相同的架构，并且在若干评估指标（BLEU@2、BLEU@3、BLEU@4）上，M4C 的表现低于 QuMLAG。在非文本问答任务上，MLPAG 几乎在所有评估指标上都优于 M4C。此外，QuMLAG 在 BLEU@4、ROUGE 和 CIDEr 上取得了最佳结果，并且在 CIDEr 上大幅领先于 MLPAG 和 M4C。仔细观察表7中的文本问答和非文本问答结果，以及表6中的总体结果，可以发现，参数规模较小（按总参数量衡量）的 MLPAG 和 QuMLAG（图中颜色相同）表现优于规模更大的 M4C。这表明，在 OpenViVQA 数据集上，M4C 这类大型模型并不奏效；因此，我们需要提出更好的方法来有效应对 OpenViVQA 数据集，而不是依赖大型模型的能力和计算资源。

7. 结论与未来工作

本文介绍了首个高质量、大规模的越南语视觉问答基准数据集，并为越南语定义了一项新颖且具有挑战性的开放式视觉问答任务。通过实验，我们展示了该任务在语言学方面的复杂性，以及在回答 OpenViVQA 数据集中的问题时，灵活有效地读取和利用图像中的场景文本的必要性。该数据集极具挑战性，以至于在 VQAv2、OCR-VQA 和 TextVQA 数据集上达到最先进水平的方法，在此数据集上的表现均有所下降。相比之下，我们提出的 FST、QuMLAG 和 MLPAG 方法在将视觉问答任务作为答案生成任务处理时，初步取得了更好的结果，从而证明以往的视觉问答方法并不适用于我们这种新型视觉问答任务。此外，尽管 MLPAG 属于单跳注意力方法，但凭借其读取和利用场景文本的能力（表7），其结果可与作为多跳注意力方法的 QuMLAG 相媲美（表6）。我们将解决这些初步方法的不足，并提出更好的方案，以有效应对 OpenViVQA 数据集带来的挑战。

此外，尽管 OpenViVQA 数据集是规模最大的越南语视觉问答数据集，但与英文视觉问答数据集相比，其规模仍相对较小，因此 M4C 等大型视觉语言模型未能达到其在英文视觉问答数据集上的最佳表现。我们计划在下一项研究中扩充 OpenViVQA 数据集，增加图像和问答对的数量，同时为每个问题构建更多答案，以便全面评估开放式答案。此外，受 Changpiny 等人 [9], 的启发，我们将把 OpenViVQA 数据集扩展为多语言视觉问答数据集，从而为包括越南语在内的多语言视觉问答研究提供高质量资源。OpenViVQA 数据集经过人工标注，可作为评估预训练视觉语言模型的高质量资源之一，尤其适用于越南语模型。

8. 若干知名视觉问答数据集示例。

我们在图20中展示了若干答案最短、中等长度和最长的问题，以便对视觉问答基准数据集进行定性观察。从图14和图20可以看出，OCR-VQA 数据集中的问题仅限于一组特定的问题。具体而言，这组问题包括 {"这本书的版本是什么？", "这本书属于什么类型？",

"这本书的标题是什么？", "这本书的版本是什么？", "这张日历上印的是哪一年？", "这是什么类型的书？", "这是哪一年的日历？"?, "这本书的作者是谁？", "这本书是谁写的？", "这本书与 <标题>有关吗？", "这是一本 <类别>书吗？"}，其中 <标题> 和 <类别> 可以从在线图书的元数据中获取。这表明，OCR-VQA 数据集的作者从互联网收集图书元数据，然后根据这些元数据，从预先定义的问题集中选取合适的问题。对于 TextVQA 数据集，答案只是图像中的场景文本；而在 OpenViVQA 数据集中，答案会将图像中的场景文本作为一个因素，从而提供更详细的信息。对于 VQAv2 数据集，答案通常较短（图15），而其最长的答案实际上是图像中的场景文本。ViVQA 与 VQAv2 使用相同的颜色。

9。关于颜色、数量和方向的问答示例。

以下是我们从 OpenViVQA 数据集中选取的若干典型示例（图21），用于展示我们的问题如何利用可见物体的颜色、数量和方向等因素，以及相应答案中包含的表示这些因素的词语。

10. 更多关于文本问答的 MLPAG 和 M4C 示例

我们提供了更多 MLPAG（图22）和 M4C（图23）的示例，以便全面观察它们将图像中的场景文本复制到答案中的表现。

参考文献

[1] J. Almazán、A. Gordo、A. Fornés 和 E. Valveny。利用嵌入属性进行词定位与识别。*IEEE*模式分析与机器智能汇刊，36(12)：2552–2566，2014年。

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould 和 L. Zhang。用于图像描述和视觉问答的自底向上与自顶向下注意力。*2018*年*IEEE/CVF*计算机视觉与模式识别会议，第6077–6086页，2017年。

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick 和 D. Parikh。VQA：视觉问答。载于国际计算机视觉会议（*ICCV*），2015年。[4] D. Bahdanau、K. Cho 和 Y. Bengio。通过联合学习对齐与翻译实现的神经机器翻译。*arXiv*预印本 *arXiv:1409.0473*，2014年。[5] S. Banerjee 和 A. Lavie。METEOR：一种用于机器翻译评估的自动指标，与人工判断的相关性更高。载于*ACL*机器翻译和/或摘要内在与外在评估指标研讨会论文集，第65–72页，密歇根州安娜堡，2005年6月。计算语言学协会。[6] P. Bojanowski、E. Grave、A. Joulin 和 T. Mikolov。利用子词信息丰富词向量。计算语言学协会会刊，5:135–146，2017年。[7] F. Borisyuk、A. Gordo 和 V. Sivakumar。Rosetta：用于图像文本检测与识别的大规模系统。载于第*24*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第71–79页，2018年。[8] S. Changpinyo、L. Xue、I. Szpektor、A. V. Thapliyal、J. Amelot、X. Chen 和 R. Soricut。迈向多语言视觉问答。*ArXiv*，abs/2209.05401，2022年。[9] S. Changpinyo、L. Xue、I. Szpektor、A. V. Thapliyal、J. Amelot、X. Chen 和 R. Soricut。迈向多语言视觉问答。*arXiv*预印本 *arXiv:2209.05401*，2022年。[10] Y.-C. Chen、L. Li、L. Yu、A. El Kholy、F. Ahmed、Z. Gan、Y. Cheng 和 J. Liu。UNITER：通用图文表征学习。载于计算机视觉——*ECCV 2020*：第*16*届欧洲会议，格拉斯哥，英国，*2020*年*8*月*23*日至*28*日，论文集，第*XXX*部分，第104–120页。施普林格，2020年。[11] J. Cho、J. Lu、D. Schwenk、H. Hajishirzi 和 A. Kembhavi。X-LXMERT：使用多模态变换器绘制、描述图像并回答问题。*arXiv*预印本 *arXiv:2009.11278*，2020年。

[12] J. Devlin、M.-W. Chang、K. Lee 和 K. Toutanova。BERT：面向语言理解的深度双向 Transformer 预训练。*arXiv*，abs/1810.04805，2019 年。[13] J. L. Fleiss。衡量多位评定者之间名义量表的一致性。心理学公报，76(5):378，1971 年。[14] K. Ganesan。ROUGE 2.0：用于评估摘要任务的更新和改进指标。*arXiv* 预印本 *arXiv:1803.01937*，2018 年。[15] Y. Goyal、T. Khot、D. Summers-Stay、D. Batra 和 D. Parikh。让 VQA 中的“V”发挥作用：提升图像理解在视觉问答中的作用。载于*IEEE*计算机视觉与模式识别会议论文集，第 6904–6913 页，2017 年。[16] K. He、X. Zhang、S. Ren 和 J. Sun。用于图像识别的深度残差学习。*2016 IEEE*计算机视觉与模式识别会议（*CVPR*），第 770–778 页，2015 年。[17] S. Hochreiter 和 J. Schmidhuber。长短期记忆。神经计算，9(8):1735–1780，1997 年。[18] M. Honnibal 和 I. Montani。spaCy 2：结合 Bloom 嵌入、卷积神经网络和增量解析的自然语言理解。即将发表，2017 年。[19] R. Hu、A. Singh、T. Darrell 和 M. Rohrbach。通过指针增强的多模态 Transformer 进行迭代式答案预测，用于 TextVQA。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 9992–10002 页，2020 年。[20] M. Huang、Y. Liu、Z. Peng、C. Liu、D. Lin、S. Zhu、N. Yuan、K. Ding 和 L. Jin。SwinTextSpotter：通过改善文本检测与文本识别之间的协同实现的场景文本定位。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 4593–4603 页，2022 年。[21] Z. Huang、Z. Zeng、B. Liu、D. Fu 和 J. Fu。Pixel-BERT：通过深度多模态 Transformer 对齐图像像素与文本。*arXiv* 预印本 *arXiv:2004.00849*，2020 年。

[22] B. K. Iwana、S. T. Raza Rizvi、S. Ahmed、A. Dengel 和 S. Uchida。不要以封面判断一本书。*arXiv* 预印本 *arXiv:1610.09204*，2016 年。[23] H. Jiang、I. Misra、M. Rohrbach、E. Learned-Miller 和 X. Chen。为视觉问答中的网格特征辩护。载于*IEEE*计算机视觉与模式识别会议（*CVPR*），2020 年。[24] S. Kantharaj、X. Do、R. T. K. Leong、J. Q. Tan、E. Hoque 和 S. R. Joty。OpenCQA：结合图表的开放式问答。*arXiv*，abs/2210.06628，2022 年。[25] V. Kazemi 和 A. Elqursh。展示、提问、关注并回答：一种强大的视觉问答基线。*arXiv* 预印本 *arXiv:1704.03162*，2017 年。[26] D. P. Kingma 和 J. Ba。Adam：一种随机优化方法。*CoRR*，abs/1412.6980，2014 年。[27] A. Kuznetsova、H. Rom、N. Alldrin、J. Uijlings、I. Krasin、J. Pont-Tuset、S. Kamali、S. Popov、M. Malloci、A. Kolesnikov、T. Duerig 和 V. Ferrari。Open Images数据集V4：大规模统一的图像分类、目标检测和视觉关系检测。国际计算机视觉杂志，2020 年。[28] G. Li、N. Duan、Y. Fang、M. Gong 和 D. Jiang。UniCoder-VL：通过跨模态预训练实现的通用视觉与语言编码器。载于*AAAI*人工智能会议论文集，第 34 卷，第 11336–11344 页，2020 年。[29] L. H. Li、M. Yatskar、D. Yin、C.-J. Hsieh 和 K.-W. Chang。VisualBERT：一种简单且高性能的视觉与语言基线。*arXiv* 预印本 *arXiv:1908.03557*，2019 年。[30] X. Li、X. Yin、C. Li、P. Zhang、X. Hu、L. Zhang、L. Wang、H. Hu、L. Dong、F. Wei 等。Oscar：面向视觉语言任务的目标语义对齐预训练。载于计算机视觉—*ECCV 2020*：第*16*届欧洲会议，英国格拉斯哥，*2020*年*8*月*23*日至*28*日，论文集，第 *XXX 16* 部分，第 121–137 页。施普林格，2020 年。[31] T.-Y. Lin、M. Maire、S. Belongie、J. Hays、P. Perona、D. Ramanan、P. Dollár 和 C. L. Zitnick。微软COCO：上下文中的常见物体。载于计算机视觉—*ECCV 2014*：第*13*届欧洲会议，苏黎世，

瑞士，*2014*年*9*月*6*日至*12*日，论文集，第*V*部分*13*，第740–755页。施普林格，2014年。[32] J. Lu、D. Batra、D. Parikh 和 S. Lee。ViLBERT：为视觉与语言任务预训练与任务无关的视觉语言表示。神经信息处理系统进展，第32卷，2019年。[33] J. Lu、C. Clark、R. Zellers、R. Mottaghi 和 A. Kembhavi。Unified-io：面向视觉、语言和多模态任务的统一模型。*arXiv* 预印本 *arXiv:2206.08916*，2022年。[34] J. Lu、J. Yang、D. Batra 和 D. Parikh。用于视觉问答的层次化问题—图像协同注意力。神经信息处理系统进展，第29卷，2016年。[35]M.-T. Luong、H. Pham 和 C. D. Manning。基于注意力的神经机器翻译的有效方法。*arXiv* 预印本 *arXiv:1508.04025*，2015年。[36] M. Mathew、D. Karatzas 和 C. Jawahar。DocVQA：一个用于文档图像视觉问答的数据集。载于*IEEE/CVF*冬季计算机视觉应用会议（*WACV*）论文集，第2200–2209页，2021年1月。[37] A. Mishra、S. Shekhar、A. K. Singh 和 A. Chakraborty。OCR-VQA：通过阅读图像中的文本进行视觉问答。载于 *ICDAR*，2019年。[38] L. T. Nguyen 和 D. Q. Nguyen。PhoNLP：一种用于越南语词性标注、命名实体识别和依存句法分析的联合多任务学习模型。载于*2021*年北美计算语言学协会分会会议论文集：演示，第1–7页，2021年。[39] N. Nguyen、T. Nguyen、V. Tran、T. Tran、T. Ngo、T. Nguyen 和 M. Hoai。词典引导的场景文本识别。载于*IEEE*计算机视觉与模式识别会议（*CVPR*）论文集，2021年。[40] K. Papineni、S. Roukos、T. Ward 和 W.-J. Zhu。BLEU：一种用于机器翻译自动评估的方法。载于计算语言学协会第*40*届年会论文集，第

311–318页，美国宾夕法尼亚州费城，2002年7月。计算语言学协会。[41] J. Pennington、R. Socher 和 C. Manning。GloVe：用于词表示的全局向量。载于*2014*年自然语言处理实证方法会议（*EMNLP*）论文集，第1532–1543页，卡塔尔多哈，2014年10月。计算语言学协会。[42] S. Ren、K. He、R. Girshick 和 J. Sun。Faster R-CNN：借助区域建议网络实现实时目标检测。神经信息处理系统进展，第28卷，2015年。[43]A. Singh、V. Natarajan、M. Shah、Y. Jiang、X. Chen、D. Batra、D. Parikh 和 M. Rohrbach。迈向能够阅读的视觉问答模型。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第8317–8326页，2019年。[44] W. Su、X. Zhu、Y. Cao、B. Li、L. Lu、F. Wei 和 J. Dai。VL-BERT：通用视觉语言表示的预训练。*arXiv* 预印本 *arXiv:1908.08530*，2019年。[45] H. Tan 和 M. Bansal。LXMERT：从 Transformer 中学习跨模态编码器表示。*arXiv* 预印本 *arXiv:1908.07490*，2019年。[46] R. Tanaka、K. Nishida 和 S. Yoshida。VisualMRC：针对文档图像的机器阅读理解。*AAAI*人工智能会议论文集，35(15)：13878–13888，2021年5月。[47] D. Teney、P. Anderson、X. He 和 A. Van Den Hengel。视觉问答的技巧与窍门：从2017年挑战赛中获得的经验。载于*IEEE*计算机视觉与模式识别会议论文集，第4223–4232页，2018年。[48] K. Q. Tran、A. T. Nguyen、A. T.-H. Le 和 K. V. Nguyen。ViVQA：越南语视觉问答。载于第*35*届亚太语言、信息与计算会议论文集，第546–554页，中国上海，2021年11月。计算语言学协会。

[49] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、L. Kaiser 和 I. Polosukhin。Attention is all you need。神经信息处理系统进展，第 30 卷，2017年。[50] R. Vedantam、C. Lawrence Zitnick 和 D. Parikh。CIDEr：基于共识的图像描述评估。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 4566–4575 页，2015年。[51] T. Vu、D. Q. Nguyen、D. Q. Nguyen、M. Dras 和 M. Johnson。Vn-CoreNLP：越南语自然语言处理工具包。载于*2018*年计算语言学协会北美分会会议论文集：演示，第 56–60 页，新奥尔良，路易斯安那州，2018年6月。计算语言学协会。[52] Z. Wang、J. Yu、A. W. Yu、Z. Dai、Y. Tsvetkov 和 Y. Cao。SimVLM：采用弱监督的简单视觉语言模型预训练。*arXiv* 预印本 *arXiv:2108.10904*，2021年。[53] P. Worley。开放式思考，封闭式提问：两种开放式和封闭式问题。《学校哲学期刊》，2015年。[54] Z. Yang、X. He、J. Gao、L. Deng 和 A. Smola。用于图像问答的堆叠注意力网络。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 21–29 页，2016年。[55] Z. Yu、J. Yu、Y. Cui、D. Tao 和 Q. Tian。用于视觉问答的深度模块化协同注意力网络。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 6281–6290 页，2019年。[56] D. Zhang、R. Cao 和 S. Wu。视觉问答中的信息融合：综述。信息融合，52：268–280，2019年。[57] P. Zhang、X. Li、X. Hu、J. Yang、L. Zhang、L. Wang、Y. Choi 和 J. Gao。Vinvl：重新审视视觉语言模型中的视觉表征。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 5579–5588 页，2021年。

[58] S. Zhang、M. Chen、J. Chen、F. Zou、Y.-F. Li 和 P. Lu。用于视觉问答的多模态特征级协同注意力方法。信息融合，73：1–10，2021年。
[59] W. Zhang、J. Yu、H. Hu、H. Hu 和 Z. Qin。通过关系推理与注意力实现用于视觉问答的多模态特征融合。信息融合，55：116–126，2020年。
[60] W. Zheng、L. Yan、C. Gou 和 F.-Y. Wang。Km4：通过知识嵌入记忆模型与相互调制实现视觉推理。信息融合，67：14–28，2021年。



(a) VQAv2

(b) ViVQA



(c) TextVQA

(d) OCR-VQA

图20：知名视觉问答数据集示例。连续三列分别展示长度最短、中等和最长的问题。



问题：学生们戴的围巾是什么颜色？
（学生们戴着什么颜色的围巾？）
答案：红色围巾（红色围巾）



问题：这个女孩穿的服装是什么颜色？
（这个女孩穿的服装是什么颜色？）
答案：她穿着白色奥黛和长裤
黑色（她穿着白色奥黛和一条黑色
长裤）

(a) 颜色



问题：有多少名保安在为顾客看管自行车
在这家商店？（这家店有多少名保安在看管
顾客的自行车？）
答案：有一名保安正在看管
顾客（有一名保安看守顾客的自行车）



问题：有多少人正在摆姿势？
拍照？（有多少人站着摆姿势
用于图片吗？
答案：五个人正站着摆姿势拍照
（五个人站着摆姿势拍照）

(b) 数量



问题：穿黑色上衣的女生正坐在哪里？
（穿黑色衬衫的女孩坐在哪里？）
答案：在穿蓝色上衣的女生左手边
（穿蓝色衬衫的女孩的左侧）



问题：这位女性在货架上挑选商品。
（这位女士正在购买哪个货架上的商品？）
答案：左侧的货架（左边的货架）

(c) 方向

图21：OpenViVQA 数据集中有关颜色、数量和方向的问答对示例



建筑物叫什么名字？（是什么这座建筑的名称是什么？）
人类：化学工程系（化学工程学院）
机器：建筑物是一种中央外语中心 (这栋建筑是外语中心)



这是什么品牌的商店？（这家店是什么品牌？）
人类：这是 thương 的商店。
Top City 店铺的优惠（这是 Top City 的门店）品牌)
机器：这是门的品牌
品牌 （这是该商店的品牌）



这家餐馆叫什么名字？（what is the 这家餐厅的名字是什么？
人类：Pho Nuong餐馆 (Pho Nuong 餐厅)
机器：商店 火锅 (火锅餐厅)



建筑物上方写着什么字？（头上的词语是什么？这座建筑？）
人类：有三星（这是三星）
机器：有字 三星 (它是三星)



这里的餐馆卖什么东西喝什么？（指的是饮料）
这个摊位卖什么？
人类：水果茶和奶茶（水果茶和奶茶）
机器: 奶茶，奶茶 (牛奶茶、奶茶)



该产品有什么特点？（有什么特别之处关于这些产品？）
人类：产品
产品此处
100% từ thiên nhiên
然而（这些产品是100% 来自自然）
机器：产品
这件产品有产品 洁净（这些产品都是新鲜的）



这家饭馆叫什么名字？（它叫什么？这家米摊的名字是什么？
人类：雪绒饭馆（Tuyet Nhung 米摊）
机器：商店黄米鸡饭 (Ga Boang米摊)



竹阳店做什么？（what is the service does Truc Duong商店提供？）
人类：花卉批发和零售粉末（花粉）零售和批发）
机器：理发店 nails 有做 美甲 (美甲美甲沙龙提供服务)



商店的特色菜“五达”饭是什么？（what is 南达特的特色餐食米摊？）
人类：Nam Dat餐馆特色菜是牛肉火锅（Nam Dat米摊有牛肉火锅作为特色餐点）
机器：牛肉火锅 （牛肉火锅 pot)



写在品牌上的内容
围栏是什么？（上面写着什么在路障上？）
人类：危险（dangerous）
机器：河内 (河内)



VNPT创新竞赛举行在哪一天？（哪一天VNPT创新竞赛举行？）
人类：VNPT创新竞赛于9月15日举行2016年（VNPT创新竞赛于2016年9月15日举行）
2016年)
机器：日 15 / 9 年
2016年（2016年9月15日）



辣排骨摊属于哪个城市？（辣排骨摊在哪里？）
人类：辣排骨摊属于哪个城市头顿（辣排骨摊位于头顿市）
机器：胡志明市 （胡志明明市）



越南春卷商店位于哪里？（越南春卷摊在哪里？）
人类：在理发店旁边南（紧挨着理发店）
机器：紧挨着商店售卖越南汤粉 （旁边是越南汤粉摊）

图22：MLPAG 的结果示例。绿色文本表示从图像中的场景文本复制的词元。



致敬活动
越南英雄母亲
和优抚家庭
公安部门2017年
在哪里举行？（where does the meeting to
欣赏越南语
英勇的母亲与家庭
享受优惠待遇
2017年警察部队的
（发生在何时？）
人类： 聚会活动 tri
向越南英雄母亲致敬
者及优抚
家庭和公安部门于
2017年在永隆举行
（致敬……的会议
越南英雄母亲
以及优抚家庭
的优待
警察部队在2017年举行
于永隆）
机器： 事件类型
家庭 永隆 和 2017年
（与永隆家庭会面
和2017年）



这是什么商店？（这是什么
类型的商店？）
人类： 这是一家儿童时装商店
（这是儿童
时装店）
机器： 这是商店 小
（这是 be 商店）



第四家商店
商业领域
什么商品？
（卖什么产品
四灵商店
？）
人类： 门店
销售品牌自行车
和自行车
电（商店
卖自行车和
电动自行车）
机器： 门
出售品牌的
类型 汽车 (该
商店出售各种
自行车）



哪个品牌正在做广告
在这辆车上？（what is the brand
在这辆车上打广告？）
人类： 米其林品牌
正在这辆车上做广告
（米其林品牌正在……上投放广告）
这辆车）
机器： 品牌 Primacy
正在这辆车上做广告。
（Primacy 品牌的广告投放于
这辆车）



这里的草坪禁止做什么？
（这片草坪上禁止做什么？）
人类： 禁止践踏草坪（请勿踩踏
这片草坪）
机器： 禁止汽车通行 （禁止汽车通行）



这家杂货店叫什么名字？
（这是什么名字
杂货店？）
人类： 杂货店 hằng
景（这里指 Hang Canh 杂货店）
商店）
机器： 杂货店
这个名字 mosfly (这是
大多数杂货店)



哪个数字出现在两边的
招牌上的数字是多少？
标牌的两边？
人类： 89号 (89)
机器： 090355号 （该号码
090355）



这个健身房的名字
什么？（名称是什么
这个健身中心的……？）
人类： 安庆（安
庆）
机器： 集体宿舍
这座名为安庆 （这座
集体宿舍
被命名为安庆）



美甲店的名字
是什么？（what is
the name of this
nail salon?)
人类： Truc
Dương（Truc
Duong)
机器： 美甲店
名为 Truc
(这家美甲店
名为 Truc)



提及的宣传海报
谈到文化之美
属于谁的？（whose
cultural beauty does
宣传板提到了
mention?)
人类： 属于人的
顺化古都的人们（people
顺化皇城的人们）
机器： 公司的
河内古城 （古老的
河内的公司）



这是哪个区域？（什么
这是该区域吗？
人类： 急救
（急诊区）
机器： 这是一个区域
隔离区域 (这是
隔离区）



第16街区是一个街区
gì? (là quý mấy?)
季度 16？）
人类： 文街区
化（文化街区）
机器： 街区 丘
Văp （戈瓦普区）



杂货店叫什么名字？
（这家
杂货店？）
人类： Thầy Triều商店
（Thầy Triều杂货店）
机器： 杂货店
名字 老师 (这家杂货店
名为 Thay)



电脑商店叫什么名字？（what
是这台计算机的名称
存储？）
人类： Quang Anh的名字（名为
Quang Anh）
机器： 名为 商店 （名为商店）



这场仪式在哪里举行？
（这场会议在哪里举行？）
人类： 在河内（Ha Noi）
机器： 在 省 培训 (在一台
（教育省份）



有床单、枕头和床垫的商店
叫什么名字？（what name is the
销售拖鞋、枕头和
床垫？）
人类： 商店名为
huy minh（这家商店名为
Huy Minh）
机器： 商店名为
明 （这家商店名为
明）

图23：M4C 的结果示例。绿色文本表示从图像中的场景文本复制的词元。