M-MRE：将互相增强效应扩展至多模态信息提取

Chengguang Gan
Yokohama National University
Yokohama, Japan
ganchengguan@yahoo.co.jp

Sunbowen Lee
College of Science, Wuhan University of Science and Technology
Wuhan, China
zhenglei2016@qq.com

Zhixi Cai
Monash University
Melbourne, Australia
zhixi.cai@monash.edu

Yanbin Wei
Southern University of Science and Technology
Shenzhen, China
Hong Kong University of Science and Technology
Hong Kong, China
yanbin.ust@gmail.com

Lei Zheng
Computer Science, Shanghai Jiao Tong University
Shanghai, China
zhenglei2016@qq.com

Yunhao Liang
University of Chinese Academy of Sciences
Beijing, China
liangyunhao22@mails.ucas.ac.cn

Shiwen Ni
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
Shenzhen, China
sw.ni@siat.ac.cn

Tatsunori Mori
Yokohama National University
Yokohama, Japan
tmori@ynu.ac.jp

Abstract

互惠增强效应（MRE）是信息提取和模型可解释性交叉领域的新兴子领域。MRE 旨在利用不同粒度任务之间的相互理解，通过联合建模来提高粗粒度和细粒度任务的性能。尽管 MRE 已经在文本领域中进行了探索和验证，但其在视觉和多模态领域的适用性仍未被探索。在这项工作中，我们首次将 MRE 扩展到多模态信息提取领域。具体而言，我们引入了一个新任务：多模态互惠增强效应（M-MRE），并构建了一个相应的数据集来支持该任务。为了应对 M-MRE 提出的挑战，我们进一步提出了一种完全兼容于多种大型视觉-语言模型（LVLM）的方法：Prompt Format Adapter（PFA）。实验结果表明，MRE 也可以在 M-MRE 任务中观察到，即一种多模态文本-图像理解场景。这提供了有力的证据，证明 MRE 在三个相关任务中促进了相互收益，证实了其在超越文本领域的普适性。

CCS Concepts

• **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Keywords

Do, Not, Us, This, Code, Put, the, Correct, Terms, for, Your, Paper

1 介绍

信息抽取（IE）[? ? ?] 传统上是自然语言处理中的一项基础任务，旨在从非结构化文本中提取特定字段并将其分类到预定义的类别中，最终将原始文本转化为结构化数据。代表性的 IE 任务包括命名实体识别（NER）[?]、关系抽取（RE）[?] 和事件抽取（EE）[?]，所有这些任务都专注于识别文本中有意义的词或跨度，并将它们分配到诸如人、组织或地点等类别中。随着近年来多模态模型的快速发展 [? ? ? ? ?]，多模态信息抽取 [? ? ?] 越来越受到研究界的关注。与仅基于文本的 IE 不同，它仅根据文本上下文对提取的实体进行分类，而多模态 IE 进一步将视觉信息纳入分类过程。换句话说，模型必须将提取的词或实体与输入图像中的相应图像段进行匹配。

如 Figure 1 所示，这是一个多模态命名实体识别（MNER）任务的例子。输入由左上角的原始图像及其附带的文本描述组成，通常是新闻标题。模型需要联合执行三个子任务。第一个任务是在输入文本上进行 NER，提取相关实体并将其分类为特定标签（例如，`organization`）。第二个任务是检测图像中的对象，并将其分割成单独的图像块。例如，在 Figure 1 中，原始图像中的三个人物被检测到并分割成三个标记为 `a`、`b` 和 `c` 的块。第三个任务是将提取的文本实体与对应的图像块进行匹配。在图中，每个人都正确匹配到与之关联的组织实体。这定义了多模态 NER 任务，它是本文的核心关注点。需要注意的是，原始的 MNER 任务仅涉及细粒度的预测，缺乏任何形式的粗粒度任务。为了在 MNER 任务中引入并展示相互增强效应（MRE）[?]，有必要引入一个多模态粗粒度任务，并将其与细粒度任务结合，以形成一个完整的 MMRE 数据集。

在正式介绍 MMRE 任务之前，我们简要解释文本领域中的 MRE 概念。在基于文本的信息抽取中，任务通常分为两个层次：字级和文本级。例如，考虑这句话：“这家餐厅的食物很好吃。”文本级任务为其分配一个如 `positive` 的情感极性标签。同时，字级任务可能输出一个标签-实体对，如 `positive: delicious`。从人的角度来看，理解一个句子通常涉及推断积极词汇的存在表明整体上的积极情感，反之亦然：知道句子表达积极情感有助于识别哪些词传达了这种积极性。因此，结合这两个层次可以让模型从多个角度理解输入。对一个任务的理解增强了另一个任务的表现。这就是信息抽取中 MRE 的本

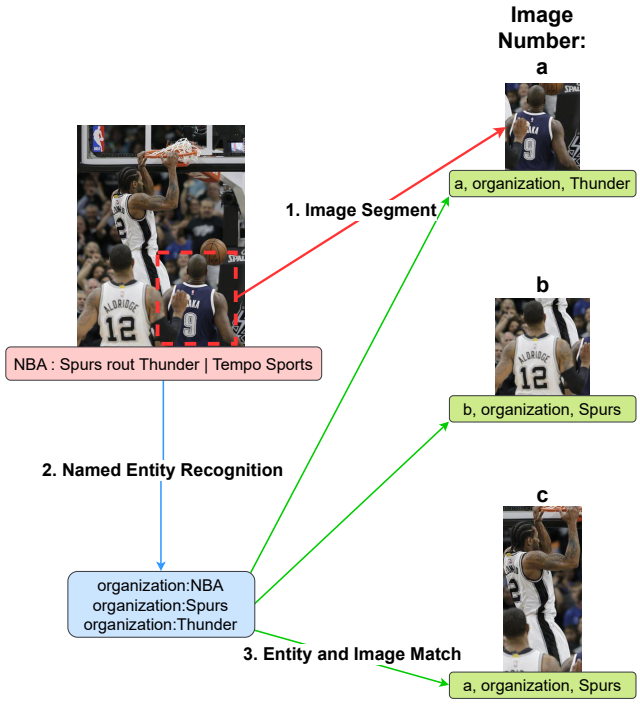

Figure 1: 多模态命名实体识别任务。包含三个子任务。文本的命名实体识别。图像的识别分割。图像和实体的匹配分割。

质：对不同粒度任务的联合建模比单独建模能产生更好的性能。我们旨在探索这种 MRE 在多模态设置中的实际效用。

我们现在正式引入将 MRE 纳入多模态 NER 任务的概念，并提出多模态互相增强效应任务（M-MRE）。如前所述，标准的 MNER 任务在细粒度水平上运行。为了构建 MRE 设置，我们引入了一个相关的粗粒度任务与 MNER 共同建模，使我们能够观察和验证 MRE 在多模态场景中的存在。

如 Figure 2 所示，我们通过图像描述任务增强了原始的 MNER 任务，该任务在多模态设置中充当粗粒度组件。在图的上半部分，模型首先对输入文本执行命名实体识别，将图像分割成对象块，并将文本实体范围匹配到相应的图像区域。在图的下半部分，模型生成整个图像的整体描述。这个描述通常包含丰富的视觉细节，如队名、球衣号码和其他上下文线索。值得注意的是，细粒度和粗粒度任务本质上是相关的：一个任务的知识可以增强另一个任务，从而产生互相增强的效果。

这种相互强化可以提升任务的表现。例如，为摘要生成所需的更深入的图像内容理解可以帮助模型在视觉上下文中更好地进行实体识别决策。反过来，识别关键的文本实体可以帮助将描述任务的重点放在图像最显著的方面。

值得注意的是，M-MRE 任务的 MNER 组件并不完全等同于原始的公式。为了在多模态模型中隔离和分析 MRE，我们通过去除物体检测和分割步骤来简化 MNER 任务。相反，模型直接提供候选图像片段，只需将其与提取的文本实体进行匹配。这个设计选择使 M-MRE 能够强调语言与视觉语言理解之间的互动，而不是视觉处理本身。通过这样做，我们控制了混杂因素，专注于多模态学习中 MRE 的出现。

为了减少构建 M-MRE 任务的标注成本，我们利用了现有的多模态命名实体识别 (GMNER) 数据集 [?] 。GMNER 已经包含了原始图像、从这些图像得出的分割图像补丁、文本内容以及相应的标签实体对。基于此数据集，我们使用大型语言模型 (LLM) 为粗粒度任务生成图像描述。这些描述然后经过人工审查和修改以确保质量，最终形成了完整的 M-MRE 数据集。

为了有效处理 M-MRE 任务，我们进一步提出了一种提示格式适配器（PFA），这是一个基于提示的框架，旨在完全兼容任何大型视觉语言模型（LVLM）。PFA 提供了一个统一的输入输出接口，使得在处理 M-MRE 中的三个子任务时能够同时进行处理。

实验结果表明，MRE 确实可以在多模态信息提取任务中观察到，从而验证了我们所提出方法的实用性。这项工作为未来多模态信息提取研究开辟了一个新的子领域。我们的主要贡献总结如下：

2 相关工作

2.1 多模态信息抽取

?] 引入了“多模态命名实体识别与定位（GMNER）”任务，该任务需要从社交媒体上的文本-图像对中提取实体-类型-区域三元组，并提出一种分层索引生成框架（H-Index），在新标注的数据集上优于现有的基线。?] 提出了“细粒度多模态命名实体识别与定位（FMNERG）”任务，其目标是通过细粒度的类别提取实体-类型-对象三元组，并介绍了一种基于 T5 的生成框架（TIGER），在新构建的 Twitter 数据集上实现了最新的结果。?] 提出了生成型多模态数据增强（GMDA）框架，用于低资源命名实体识别（MNER），结合标签感知文本生成与图像合成技术，以在全监督与有限监督的情况下提高实体识别性能。?] 提出了一个基于图卷积的信息提取模型，用于从视觉丰富文档（VRDs）中提取信息，有效整合文本和视觉布局特征，在真实世界数据集上优于 BiLSTM-CRF 基线。此外，还有许多其他多模态信息抽取 ?] 子任务，包括关系抽取 [?] 、事件抽取 [?] 、情感分析 [?] 等等。

2.2 相互增强效应

互相增强效应（MRE）最初是为文本信息提取领域提出的。?] 提出了句子到标签生成（SLG）框架，用于多任务学习日语句子分类（SC）和命名实体识别（NER），将这两个任务统一在生成范式下。他们通过在现有的日语维基百科 NER 语料库上注释句子级类别，构建了一个新的 SCNM 数据集，并证明联合学习 SC 和 NER 任务可以产生 MRE，使每个任务的表现从其他任务中受益。他们的结果表明，SLG 相较于单任务基线分别提高了 SC 和 NER 的准确率 1.13 和 1.06 分。他们还引入了一种约束机制（CM），以确保输出格式的正确性，并与 Shinra NER 语料库进行增量学习，以提高模型的 NER 能力。后续研究已将 MRE 扩展到其他语言和信息提取中的各种子任务，通常与在专业训练范式下的大型语言模型（LLMs）结合使用。基于 MRE 的方法持续优于传统的将每个 IE 任务单独处理的训练方法。

尽管 MRE 仍然是一个相对未被充分研究的领域，但其结合细粒度和粗粒度任务的形式尤其适合多模态信息提取。这是因为文本和图像自然而然地互补并相互强化，使得相互增强不仅成为可能，而且是有益的。这一观察是我们工作的主要动机。

3 多模态相互增强效应

如前所述，M-MRE 任务包括细粒度和粗粒度子任务。在本节中，我们首先描述处理 M-MRE 任务的整体处理流程，然后解

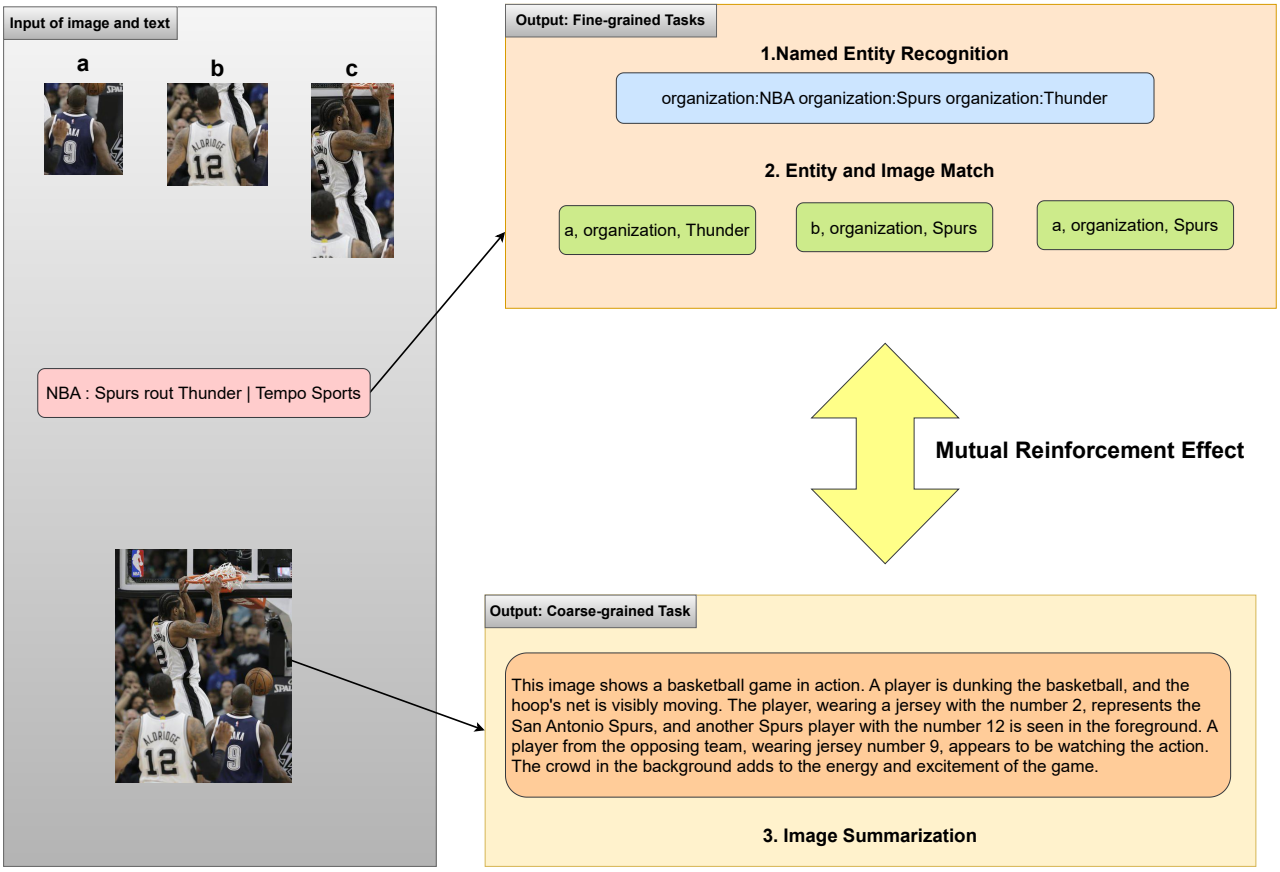

Figure 2: 多模态互相增强效应任务说明。

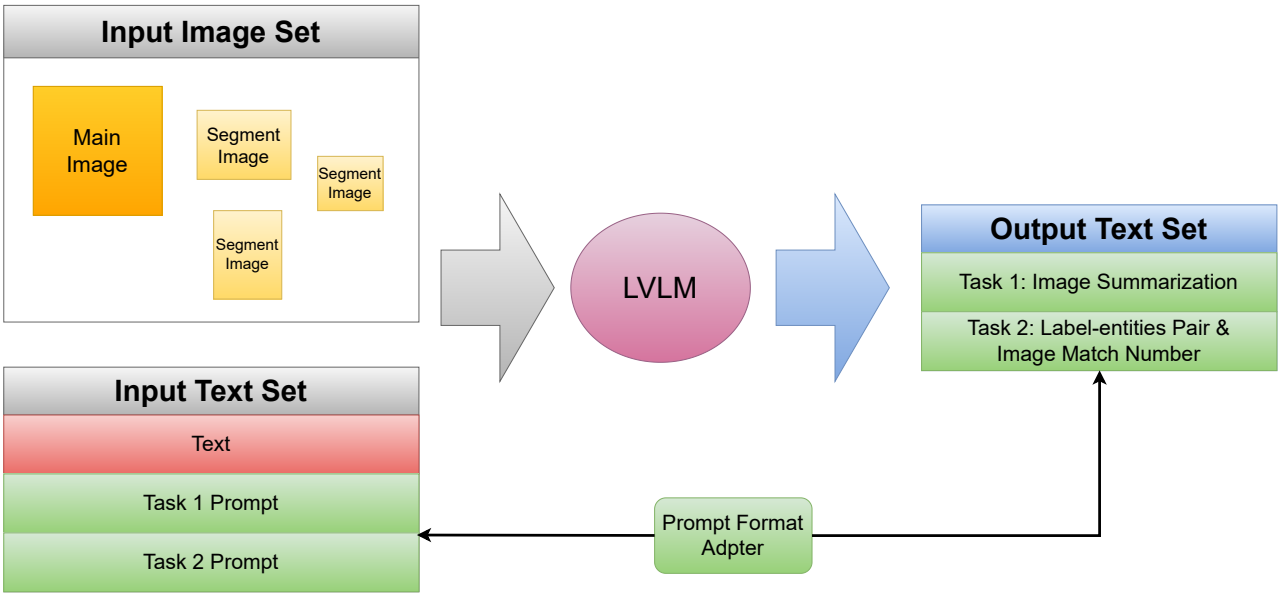

Figure 3: 大型视觉语言模型在处理 M-MRE 任务时的输入和输出。

Table 1: 在 M-MRE 数据集中任务 1 图像摘要答案的统计。

Metric	Max	Min	Median	Mean
Character Count	573	40	317	318
Word Count	97	10	54	53

释如何构建 M-MRE 数据集。接下来，我们详细描述提出的提示格式适配器（PFA）。

如 **??** 所示，处理 M-MRE 任务时，大型视觉-语言模型（LVLMs）的输入和输出结构由两个主要组件组成：图像组和文本组。图像组包括主要图像和从中分割出来的若干图像块。文本组包含附带的文本输入，作为命名实体识别（NER）的目标文本。在文本组中，用绿色高亮显示的部分代表提示格式适配器（PFA），它统一了输入和输出格式，从而简化了 LVLM 对多个子任务的处理。

输入端的 PFA 由两个提示指令组成，对应于两个子任务：任务 1（图像总结）和任务 2（命名实体识别和段落图像匹配）。需要注意的是，任务 2 并没有被分隔成两个独立的组件。这是因为段落图像匹配本质上涉及识别和标记实体范围；也就是说，它包含了从命名实体识别中提取标签实体对。因此，我们将这两个方面的提示合并为一个指令，以有效引导模型的响应生成。

一旦所有输入图像和文本提示准备就绪，它们就会依次传递到 LVLM 中。然后，该模型生成两个任务的结构化输出。重要的是，输入中的两个绿色文本提示通过 PFA 进行了格式化，它还确保输出符合信息提取的结构要求——特别是，生成可以直接作为结构化文本存储和重用的数据。

对于任务 1，输出包括一个自然语言摘要，该摘要描述了图像的视觉内容。对于任务 2，输出是一个标签-实体对序列，以元组的形式组织，并按照其提取顺序排列。在此之后，模型生成第二个元组序列，每个元组将一个分割的图像编号与对应的实体范围配对，从而完成实体到图像的匹配任务。

在我们描述 M-MRE 数据集的构建过程之后，下一部分将详细解释 PFA 的设计。上述内容描述了将 LVLMs 应用于 M-MRE 任务的完整工作流程，使得可以通过统一的输入输出接口同时处理细粒度和粗粒度的子任务。

3.1 多模态互相增强效应数据集

验证多模态信息提取中的互相增强效果（MRE）的一个主要挑战是缺乏专门的多模态 MRE 数据集。为了解决这一缺口，我们通过使用大型语言模型（LLM）对现有的多模态 NER 数据集进行额外的粗粒度注释，构建了 M-MRE 数据集。具体来说，我们利用了已经包含细粒度多模态 NER 标签的有依据的多模态命名实体识别（GMNER）数据集。这使我们能够通过增量注释构建 M-MRE 数据集。

GMNER 数据集本身是构建在两个流行的多模态 NER 数据集 **?**] 和 **?**] 之上的。GMNER 引入了原始图像的对象级分割，并将分割的片段与相应的实体跨度对齐。在我们的工作中，我们通过生成图像级摘要进一步增强 GMNER 数据集，从而为 M-MRE 任务所需的粗粒度组件奠定基础。

如 Figure 4 所示，我们将主图像和一个预定义的指令提示模板输入到 GPT-4o 模型中，特别是 `yu-etal-2023-grounded` 版本。然后模型生成图像内容的描述性摘要。为了确保质量和一致性，每个生成的摘要都由多名注释者进行交叉检查和完善。因此，所得的 M-MRE 数据集结合了细粒度和粗粒度的标注。

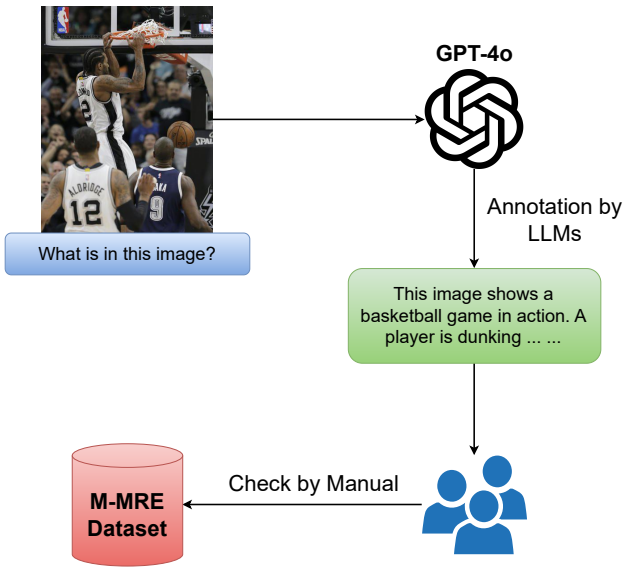

Figure 4: 大型视觉-语言模型在处理 M-MRE 任务时的输入和输出。

图像摘要组件的注释统计如 Table 1 所示。每个摘要的平均字数为 53，中位数为 54。结果表明，GPT-4o 模型产生的输出稳定且均衡——既不太短以至于遗漏重要的图像细节，也不太长以至于过度描述视觉内容。

3.2 提示格式适配器

我们现在介绍提示格式适配器（PFA）的设计，它在 M-MRE 框架中扮演两个关键角色。首先，它作为一个提示来指导模型的输出行为。其次，它强制执行一致的输出格式，使生成的文本数据具有结构化和机器可读的特点。

如 Figure 5 所示，PFA 由两个部分组成：输入格式和输出格式。输入部分包括三个提示模块。第一个模块是一个通用指令，它告诉模型回答接下来的两个任务，同时抑制不必要的冗长输出。在 LVLMs 中，这个指令至关重要，因为它们通常在后期训练中通过指令微调进行训练，并能有效地响应这种指导。

第二个和第三个模块分别对应任务 1 和任务 2。其中，任务 2 的指令格式更为复杂，因为它涉及两个子任务。对于 NER 子任务，我们将输出格式定义为 `:label;entities` 。受到基于文本 MRE 信息抽取的先前工作的启发，我们采用冒号（`:`）来标记标签的开始，然后用分号（`;`）将标签与其对应的实体范围分开。实体文本随后直接附加。这种格式可以很容易地扩展到多个标签-实体对的顺序表示（例如，`:label1;entities1:label2;entities2:label3;entities3...`），实现紧凑而结构化的表示。

对于任务 2 的第二个子任务，即将分割的图像块与实体匹配，我们将输出格式定义为一组键值映射，例如 `{ 'a': 'entities1' }` ，其中每个块的 ID 对应于它所代表的实体。这种标准化使得模型输出的解析更加可靠和可扩展。

在输出方面，PFA 进一步加强固定文本标记，以明确指示每个任务响应的开始。具体来说，`Task 1 Answer:` 和 `Task 2 Answer:` 被用作模型输出的标题。在任务 2 的响应中，我们包含了额外的指导层次，以区分两个子任务：`NER:` 前缀（在 Figure 5

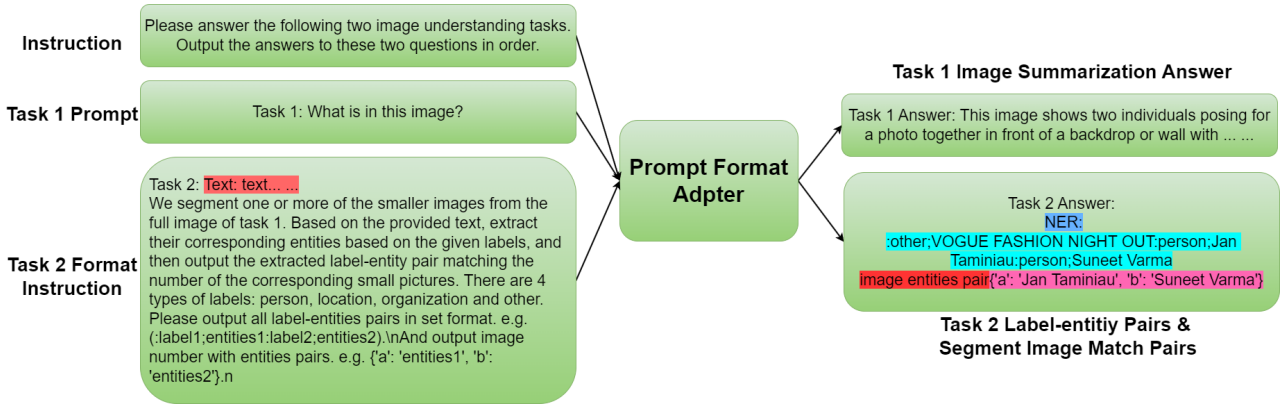

Figure 5: 大型视觉语言模型在处理 M-MRE 任务时的输入和输出。

中以绿色突出显示）表示 NER 子任务的输出，而 `image-entity pair:` 前缀（以红色突出显示）表示图像实体匹配的输出。

总之，M-MRE 框架中的所有子任务都以统一的、结构化的格式被提示和输出。由于 PFA 仅修改和约束文本输入和输出，并未改变模型架构或权重，因此它保持与模型无关，并且可以以即插即用的方式无缝应用于任何 LVLM。

4 实验设置

对于基础模型，我们选择了 `Qwen2.5-VL-7B-Instruct` [1]，这是一个以在多模态理解任务中表现出色而闻名的先进的开源 LVLM。我们使用了最终版的 M-MRE 数据集，其中包含 5,533 个标注样本，并随机将其以 7/3 的比例划分为训练集和测试集。

我们使用 `bfloat16`（BF16）精度进行全参数微调来训练模型。训练的超参数如下：批次大小为 4，文本编码器的学习率为 1e−5，视觉编码器的学习率为 2e−6，联合训练的统一学习率为 2e−6。模型训练进行 3 个周期。

我们采用 Qwen 微调框架 [2] 进行训练。多模态输入格式遵循 `LLaVA` 输入标准，确保与现有指令跟随 LVLM 管道的兼容性。所有实验均在分布式环境下使用四个 NVIDIA A100 80GB GPU 进行。

4.1 评估

由于 M-MRE 任务的多任务性质和多样化的输出结构，对其进行评价并非易事。为便于评价，我们首先根据预定义的关键词将模型的输出文本分为三个部分：（1）用于任务 1 的图像摘要文本，以及（2）任务 2 中两个子任务的结构化文本输出。

对于任务 1（图像摘要），我们采用广泛使用的自然语言生成度量标准：ROUGE-1、ROUGE-2、ROUGE-L 和 BLEU。这些度量标准评估生成的摘要与人工标注的真实情况之间的内容重叠。

对于任务 2，我们对两个子任务使用不同的度量标准。NER 子任务使用 F1 分数进行评估，F1 分数衡量提取的标签-实体对的精确度和召回率的调和平均值。图像-实体匹配子任务使用准确率进行评估，准确率衡量正确匹配的图像片段与其对应实体的比例。

F1 分数的计算方式如下：

Algorithm 1: M-MRE 任务的评估程序

Input: Predicted text T_{pred} , Reference text T_{ref}
Output: BLEU, ROUGE-1/2/L, F1_NER, ImageEntity_Accuracy

Extract summary texts from T_{pred} and T_{ref} using regex pattern “`Task 1 Answer:`”;
Extract NER text using pattern “`NER::`”;
Extract image-entity pairs using pattern “`image entities pair`”;
Parse NER label-entity pairs using delimiter : and ;;
Convert image-entity mappings to dictionaries;
Compute ROUGE and BLEU between predicted and reference summaries;
Match NER label-entity pairs and compute precision and recall;
Compute F1 score: $F1 = \frac{2PR}{P+R}$;
Match predicted and reference image-entity pairs and compute accuracy;
return All metrics: { BLEU, ROUGE-1/2/L, F1_NER, ImageEntity_Accuracy } ;

评估流程在 algorithm 1 的伪代码中进行了说明。在我们的方法中，词级预测首先解析为单独的标签-实体对，例如 `:neutral;unique` 、`:positive;boring` 、`:positive;sad` 。这些预测的对随后与真实标签-实体对进行匹配。

精度定义为正确匹配对的数量除以预测对的数量。召回率计算为正确匹配对的数量除以真实对的数量。这种基于匹配的评估为结构化预测任务中模型性能提供了一种细致入微和现实的测量。

5 结果

如 **??** 所示，我们展示了使用我们提出的 PFA 框架完成 M-MRE 任务的完整评估结果。在这里，单一 TS 表示仅执行图像到文本的摘要任务，代表一种粗粒度的多模态任务。单一 MNER 指的是仅执行 MNER 任务，并评估其两个子任务，代表一种细

[1] https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct
[2] https://github.com/2U1/Qwen2-VL-Finetune