场景图增强的视觉语言常识推理生成

袁凡, 李丕绩*
计算机科学与技术学院/人工智能学院,
南京航空航天大学
江苏省，南京市，210016
{fanyuan, pjli}@nuaa.edu.cn

摘要

视觉语言常识推理是一类旨在理解视觉场景的任务，常用于评估人工智能系统的多模态常识推理能力。然而，可靠的常识推理需要细致的场景理解，而现有的基于预训练模型微调的方法却无法有效地利用具体场景中存在的物体关系信息，因此其推理的合理性存在较大的局限性。为解决上述问题，本研究提出了一种场景图增强的视觉语言常识推理生成框架SGEVL。该框架首先使用图像补丁序列提供视觉信息，并通过一种包含注意力模块的门控机制，赋予大型语言模型理解视觉信息的能力。基于该框架的视觉语言能力，进一步提出了一种无位置信息的场景图生成方法。生成的场景图能够显著提升模型对场景信息的理解，从而引导生成高质量的回答和推理。通过在VCR，VQA-X和e-SNLI-VE数据集上分别实验，实验结果表明本文提出的视觉语言常识推理框架性能优于基线模型。此外，通过消融实验和结果可视化，进一步证明了该框架中每个模块的有效性。

关键词： 多模态融合；视觉语言常识推理；场景图生成

Scene Graph Enhanced Visual Language Commonsense Reasoning Generation

Fan Yuan, Piji Li*
College of Computer Science and Technology/Artificial Intelligence,
Nanjing University of Aeronautics and Astronautics
Nanjing, Jiangsu 210016, China
{fanyuan, pjli}@nuaa.edu.cn

Abstract

Visual Language Commonsense Reasoning is a type of task aimed at understanding visual scenes, often used to evaluate the multimodal commonsense reasoning capabilities of AI systems. However, reliable commonsense reasoning needs a detailed understanding of the scene, and existing methods based on fine-tuning pre-trained models are often incapable of effectively utilizing the object relationship information present in specific scenes, thereby limiting the rationality of their reasoning. To address this issue, this study proposes a scene graph enhanced visual language commonsense reasoning generation framework SGEVL. This framework first uses a sequence of image patches to provide visual information and employs a gating mechanism that incorporates an attention mechanism to empower large language models with the ability to understand visual information. Based on the visual language capabilities of this framework, we further propose a method for generating scene graphs without positional

*通讯作者

information. The generated scene graphs can significantly enhance the model's understanding of scene information, thereby guiding the generation of high-quality answers and reasoning. Experiments on the VCR, VQA-X, and e-SNLI-VE datasets show that the performance of the proposed framework surpasses that of baseline models. Furthermore, through ablation studies and visualization results, the effectiveness of each module in this framework is further demonstrated.

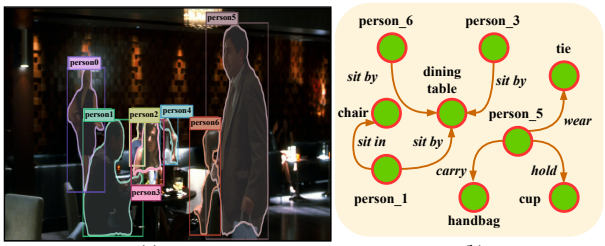
Keywords: Multimodal fusion , Visual language commonsense reasoning , Scene graph generation

1 引言

深度学习模型在视觉-语言任务中已经取得了显著的进步，并展示出其卓越的性能。然而，由于这些模型的“黑箱”特性，其可解释性已经成为了一个日益关注的问题。视觉语言常识推理是一类需要模型从识别和认知两个层面进行推理的视觉推理任务。其中，视觉语言自然语言解释（Vision-language Natural Language Explanation，VL-NLE）(Dua et al., 2021; Kayser et al., 2021; Plüster et al., 2022; Sammani et al., 2022; Whitehouse et al., 2023)近期受到了广泛的关注，其研究中心在于模型是否能够理解视觉场景中的现象和常识，并生成细粒度的推理来解释自身做出的回答或是一些给定的判断。

通过分析，我们发现在VL-NLE多个数据集中的问题、答案和推理解释中包含大量的物体对象及其关系信息。例如，在视觉常识推理数据集（Visual Commonsense Reasoning，VCR）(Zellers et al., 2019)中，有99.7%的问题、97.6%的答案和98.8%的推理解释同时包含了对象和它们的位置关系。平均每一个问题、答案和推理解释的句子都包含超过1个对象和2个以上的关系。特别地，每个推理解释平均包含4个对象和3.8个关系。例如，“He is sitting in a cushioned piece of furniture. He has his briefcase on his lap and he is engrossed in a book.”包含了如下的关系：“He is sitting in furniture”，“briefcase on lap”和“he is engrossed in a book”。因此，对对象及其关系的深入理解在视觉语言常识推理中起着关键的作用。以图1为例，当被问到“what are person1 and person3, and person6 doing?”时，一个没有对象关系提示的模型生成了以下的答案：“person1, person3, and person6 are having a conversation. Because: person1 is smiling and person2 is looking up at person3.”。然而，在实际的场景中，像“dining table”和“cup”这样的关键物体能够明确地帮助我们识别出这是一个用餐的场景，而不仅仅是一个对话的场景。而上述生成的推理过于强调人的因素，忽视了周围环境中的其他对象。因此，考虑到对象关系所生成的推理“they are eating dinner together. Because: they are sitting at a table and have drinks on the table.”则更为精确。

基于以上的观察，我们认为当前的模型在进行视觉语言常识推理时，应当着重于考虑场景中对象的关系。场景图 (Krishna et al., 2017; Qiu et al., 2023; Tang et al., 2020a; Tang et al., 2020b)被认为是一种对对象关系进行建模的有效方法。场景图是由节点和边组成的图结构，其中每个节点代表一个对象，每个有向边表示两端对象之间的关系。例如，如图1 (b)所示，在三元组“⟨person, holding, cup⟩”中，“person”和“cup”是对象，“holding”则是关系。通过提取对象之间的各种位置和语义关系，场景图在某些任务中已经展示出了其潜力 (Yu et al., 2021; Wang et al., 2022b)。考虑到VCR提供的图像无法提供如此详细的位置和关系信息，场景图非常适合在推理过程中作为一座桥梁来填补这个空白。高质量的场景图是提升模型视觉语言常识推理能力的关键。在构建场景图过程中，最常见的方法是首先使用例如Faster-RCNN (Ren et al., 2015)等的目标检测器进行目标检测，然后预测检测到的目标之间的关系。然而，过去的大部分场景图生成方法依赖于对象的位置和边界框信息，采用了复杂的机制和流程 (Dhamo et al., 2020; Dhamo et al., 2021; Johnson et al., 2015)，这对于大多数下游任务来说不是必要的。考虑到上述的不足，我们认为被赋予视觉理解能力的大型语言模型（Large Language Models，LLMs）可以在无显示位置信息的设定下更好地区分两个对象之间的潜在空间关系。



(a) (b)

问题	what are person1, person3 , and person6 doing?
无场景图辅助的回答和推理	person1, person3 and person6 are having a conversation. Because: person1 is smiling and person2 is looking up at person3.
有场景图辅助的回答和推理	they are eating dinner together. Because: they are sitting at a table and have drinks on the table.
真实标签	they are eating and drinking together. Because: they are all sitting at the same table and they have food and drinks in front of them.

(c)

Figure 1: 一个视觉语言常识推理的例子。（a）是VCR数据集中的一个场景。（b）是一个从（a）中产生的场景图。在表格（c）中，第一行是问题，第二行是不含场景图辅助的回答和推理，第三行是包含场景图辅助的回答和推理，最后一行是来自数据集的真实标签。

为了解决上述问题，本文提出了一个场景图增强的视觉语言常识推理生成框架SGEVL。为了获取更全面的视觉特征表示，考虑采用CLIP (Radford et al., 2021)模型。CLIP模型的主要结构是一个文本编码器和一个图像编码器，它们分别接收原始文本和图像，然后将输出的图像和文本向量表示映射到一个联合的多模态空间。CLIP模型在训练时采用了对比学习方法，训练其对图像和文本匹配关系的理解能力。对于输入的图像文本对，让正样本对的相似度更高，负样本对的相似度更低。所以CLIP模型拥有较强的视觉图像识别能力以及特征提取能力，能够提供质量较高的视觉特征表示。因此，使用CLIP中的图像补丁序列作为输入，并通过包含注意力模块的门控机制，赋予大型语言模型理解视觉信息的能力。为了实现更可靠的推理感知，我们首先在Visual Genome（VG）数据集 (Krishna et al., 2017)上训练场景图生成，并为视觉语言自然语言解释数据集生成对应的场景图。接着构建prompt模板，结合场景图引导生成高质量的回答和推理。我们在VCR (Zellers et al., 2019)，VQA-X (Park et al., 2016)和e-SNLI-VE (Kayser et al., 2021)三个公共数据集上进行了实验，结果显示我们提出的方法在自动和人工评估指标上都超过了基线方法。消融实验和可视化结果进一步证明了该框架中每个模块的有效性。本文的主要贡献如下：

- 本文结合图像补丁序列和大型语言模型，引入了一种包含注意力模块的门控机制进行模态融合，赋予大型语言模型理解视觉信息的能力。
- 本文提出了一种无位置信息的场景图生成方法和一种基于阈值的场景图筛选方法，并利用筛选后的场景图提供辅助信息，引导生成高质量的回答和推理。
- 本文在三个英文公共数据集上进行了大量的实验，验证了本文提出的框架取得了优于研究进展方法的效果。

2 相关工作

场景图包含了详细的对象、属性和关系的语义信息。具体来说，它是由节点和边组成的图结构，其中每个节点代表一个对象，每个有向边表示两端节点之间的关系。场景图的主要表示形式是三元组“⟨主语, 关系, 宾语⟩”，其中主语和宾语用节点表示，而关系则用有向边表示。场景图生成（Scene Graph Generation，SGG） (Lu et al., 2016; Li et al., 2017a; Zhang et al., 2017)是一项从图像场景中提取此类语义表示的任务。现有的场景图生成方法可以分为两类。第一类包括两个阶段：首先使用Faster-RCNN (Ren et al., 2015)等目标检测方法进行目标检测，然后在此基础上进行关系预测 (Lu et al., 2016; Dai et al., 2017; Liao et al., 2019)。考虑到图像中的元素可能表示其他元素的上下文，第二种方法采用对象和对象关系的联合预测 (Li et al., 2017a; Li et al., 2017b; Xu et al., 2017)。一些工作专注于场景图中的长尾问题 (Zhang et al.,

2017; Krishna et al., 2019; Zareian et al., 2020)，并提出了无偏场景图生成 (Yan et al., 2020; Wang et al., 2020)。此外，一些研究已经提出了无位置场景图生成的概念 (Özsoy et al., 2023)，其目标是在不使用边界框等位置信息的情况下进行场景图生成。尽管现有的研究已经能够生成相对高质量的场景图，但要实现其实用性仍面临着巨大的挑战。

视觉语言自然语言解释（Vision-language Natural Language Explanation，VL-NLE）(Dua et al., 2021; Kayser et al., 2021; Plüster et al., 2022; Sammani et al., 2022; Whitehouse et al., 2023)是自然语言解释任务（Natural Language Explanation，NLE）在多模态领域的扩展。它旨在通过生成对人类友好且细粒度的自然语言句子来解释黑盒模型的决策过程。e-ViL (Kayser et al., 2021)是一个针对VL-NLE任务的评估基准，其中包含VQA-X (Park et al., 2018)、VCR (Zellers et al., 2019)和e-SNLI-VE (Kayser et al., 2021)三个数据集。视觉常识推理（Visual Commonsense Reasoning，VCR）(Zellers et al., 2019)是其中较为复杂的一项多模态任务。VCR数据集提供了大量的图像和相关问题，并基于这些信息推断出答案和解释。每条数据都包含四个备选答案和四个备选推理选项。与经典的视觉问题回答任务不同，VCR提供的选项通常涉及到多组对象关系或事件的全面描述，要求模型对场景信息和常识信息有深刻的理解，因此同时做出正确的选择是具有挑战性的。为了评测模型的生成推理能力，e-ViL将VCR任务从分类任务转化为生成任务。过去已经有许多工作 (Dua et al., 2021; Kayser et al., 2021; Whitehouse et al., 2023)致力于提高模型的解释能力。例如，e-UG (Kayser et al., 2021)使用UNITER (Chen et al., 2020)作为预测模块进行解释预测，NLX-GPT (Sammani et al., 2022)通过训练一个蒸馏的GPT-2 (Radford et al., 2019)来结合答案和解释的生成。这些方法在捕捉场景中对象之间精确关系以进行推理的能力上仍存在困难，而解决这个问题正是本文的主要目标。

3 方法

为了增进模型生成合理回答和解释的能力，本文提出了一种场景图强化的视觉语言常识推理生成方法SGEVL。该方法基于输入的图像和问题，首先根据图像生成对应的场景图，然后利用场景图来改善模型视觉语言常识推理的性能。整体框架如图2所示。其中，定义I为给定的图像，Q为相应的问题，$O = (o_1, o_2, \ldots, o_n)$为对应图像中的对象，其中$n$表示对象的数量。在无位置的场景图构建阶段，将图像$I$的补丁序列$P = (p_1, p_2, \ldots, p_m)$作为视觉输入$X_{\mathrm{v}}$，将包含对象的文本提示prompt作为文本输入$X_{\mathrm{t}}$。目标是预测一个场景图$G = (V, E)$，其中$V$和$E$分别代表实体节点和关系。随后，在生成视觉语言常识推理的答案和解释时，使用图像I的补丁序列$P = (p_1, p_2, \ldots, p_m)$作为视觉输入$X_{\mathrm{v}}$，对问题$Q$和场景图$G$构建文本提示prompt作为输入$X_{\mathrm{t}}$。目标是生成答案和解释$(I, G, Q) \rightarrow (A, R)$，其中$A$代表答案，$R$表示解释。我们期望生成的$A, R$在$G$的帮助下是合理并符合常识的。

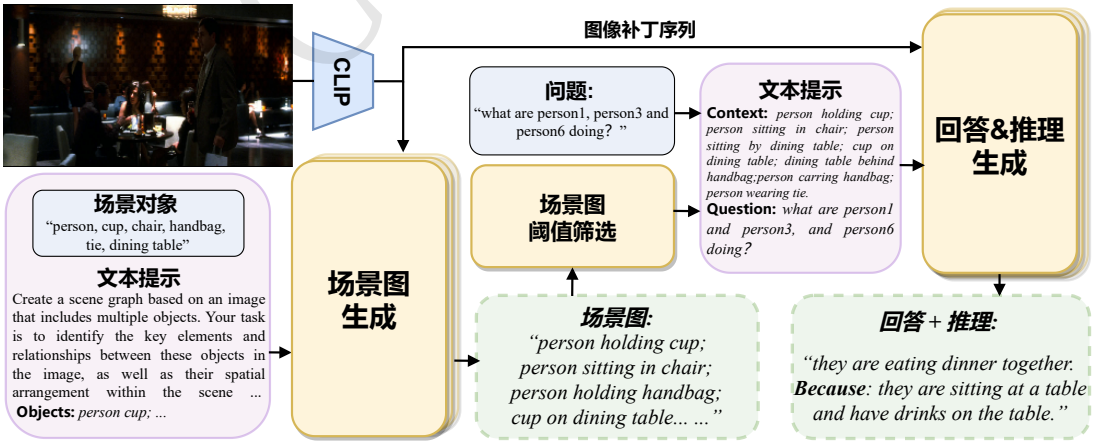


Figure 2: SGEVL框架的结构图。该框架首先根据图像的补丁序列和物体提示生成场景图。然后，结合问题和图像，生成符合常识的回答和解释。

3.1 视觉语言模态融合

由于大型语言模型是文本单模态模型，无法利用图像场景信息，因此本文采用了一种基于单头注意力的门控机制对为大型语言模型提供多模态能力。为了使不同模态的特征进行有效的融合，并更大程度的保留视觉信息，我们使用CLIP模型 (Radford et al., 2021)的图像补丁序列输出$P = (p_1, p_2, \ldots, p_m)$作为视觉输入$X_\mathrm{v}$，并使用一个线性层将其投影到与文本输入相同的隐层维度，定义为视觉隐层向量$\mathbf{H}_\mathrm{v}$。接着，使用一个单头跨模态注意力模块对视觉输入和文本输入进行模态融合获得跨模态隐层向量$\mathbf{H}_\mathrm{c}$:

$$\mathbf{H}_\mathrm{t} = \mathbf{TextEncoder}(X_\mathrm{t}), \tag{1}$$

$$\mathbf{H}_\mathrm{v} = \mathbf{W_h} \cdot X_\mathrm{v}, \quad \text{s.t. } X_\mathrm{v} = \mathbf{CLIP}(I), \tag{2}$$

$$\mathbf{H}_\mathrm{c} = \mathbf{Attention}\left(\mathbf{H}_\mathrm{t}\mathbf{W^Q}, \mathbf{H}_\mathrm{v}\mathbf{W^K}, \mathbf{H}_\mathrm{v}\mathbf{W^V}\right), \tag{3}$$

其中，文本隐层向量$\mathbf{H}_\mathrm{t}$作为查询向量，视觉隐层向量$\mathbf{H}_\mathrm{v}$作为键值向量和值向量，$\mathbf{W^Q}$，$\mathbf{W^K}$，$\mathbf{W^V}$和$\mathbf{W_h}$分别表示可学习的参数。

为了更好地利用大型语言模型的文本语义理解能力，我们引入了一种门控机制 (Zhang et al., 2023)帮助平衡两个不同的模态。首先，对跨模态隐层向量$\mathbf{H}_\mathrm{c}$和文本隐层向量$\mathbf{H}_\mathrm{t}$在特征维度上进行拼接。接着，使用线性层将其维度调整至与隐层向量相同，并通过sigmoid函数计算出门控参数λ。最后，使用λ对跨模态隐层向量$\mathbf{H}_\mathrm{c}$和文本隐层向量$\mathbf{H}_\mathrm{t}$加权相加获得融合隐层向量$\mathbf{H}_\mathrm{f}$。过程可以表示为:

$$\lambda = \mathbf{Sigmoid}\left(\mathbf{W_t} \cdot \mathbf{H}_\mathrm{t} \oplus \mathbf{W_v} \cdot \mathbf{H}_\mathrm{c}\right), \tag{4}$$

$$\mathbf{H}_\mathrm{f} = (1 - \lambda) \cdot \mathbf{H}_\mathrm{t} + \lambda \cdot \mathbf{H}_\mathrm{c}, \tag{5}$$

$\mathbf{W_t}$和$\mathbf{W_v}$分别表示可学习的参数。融合隐层向量$\mathbf{H}_\mathrm{f}$作为模态联合输入参与模型的训练与推理。

3.2 无位置信息的场景图生成

无位置信息的场景图生成使用Visual Genome（VG）数据集 (Krishna et al., 2017)进行训练。在VG数据集中，每一张图像包含不同数量的场景图，而这些场景图中的对象也在大小上有所不同。在一般的认知中，特别小的对象，如“hair”和“finger”，通常会包含更普遍的关系信息，在其关联的场景图中能提供的信息较少。而较大的对象往往能够引领场景中的更多关系信息，在推断中起着更关键的作用。因此，为了提升模型训练效率，我们对VG数据集进行了一定程度的数据清理。根据不同场景图三元组中主语对象和宾语对象的平均大小对每个图像对应的所有场景图进行排序，并选择前50个进行训练。这些场景图已经可以囊括大部分有效的对象，并有效排除部分信息量较小的三元组。

由于大型语言模型无法通过边界框等显示位置数据定位对象，因此，为了训练模型获得无位置信息的场景图生成能力，我们不使用场景中各个对象的边界框等显示位置数据作为输入。但是，为了防止图像场景中存在相同命名的对象影响模型的预测效果，根据数据集中所选对象从左往右出现的位置给定一个序号。例如，场景中存在多个“person”对象，将左边的对象定义为“person0”，右边的对象定义为“person1”，并以此类推。同时，我们也加入了个别无具体关系的对象，构成“⟨主语, None, 宾语⟩”的三元组，以训练模型应对无关对象的情况。

具体来说，使用基于Transformer (Vaswani et al., 2017)的大型语言模型FLAN-T5 (Chung et al., 2022)作为主干模型，它是一个编码器-解码器架构。受到Vision Transformer (Dosovitskiy et al., 2021; Liu et al., 2021)在视觉领域成功的启发，我们使用经过训练的CLIP模型提取图像的补丁序列作为视觉输入。将标签中的场景图拼接作为待预测的场景图三元组序列$G = \{sub_1\ obj_1\ rel_1;\ \ldots\ ; sub_\sigma\ obj_\sigma\ rel_\sigma\}$，其中$\sigma$表示三元组的数量。并基于此将筛选过的场景图中的对象按照$X_\mathrm{o} = \{sub_1\ obj_1; sub_2\ obj_2;\ \ldots\ ; sub_\sigma\ obj_\sigma\}$的形式构成文本输入。为了能够利用大型语言模型理解指令的能力，我们构建了文本提示“Create a scene graph based on an image that includes multiple objects. The task is to identify the key elements and relationships between these objects in the image, as well as their spatial arrangement within the scene. Objects:$\{X_\mathrm{o}\}$ Scene:”。然后将文本提示作为文本输入X_t，图像补丁序列作为视觉输入X_v，经过第3.1节的模态融合模块后生成对应的场景图G。

由于预测场景图的方式是从序列到序列的，因此，采用交叉熵损失作为无位置信息的场景图生成的训练损失：

$$\mathcal{L}_{SG} = -\sum_{l}^{L} \log P\left(g_l \mid g_{<l}, X_{\mathrm{t}}; X_{\mathrm{v}}\right), \tag{6}$$

其中，g_l是第l个生成的词，X_{t}包含对象信息的X_{o}文本提示，X_{v}表示视觉输入。

3.3 视觉语言常识推理生成

在视觉语言常识推理阶段，视觉语言自然语言解释（VL-NLE）的目标是先生成对应的回答，再生成能够帮助解释回答的推理。为了提升模型推理的一致性，我们将这两个离散的任务合并为一个，即同时生成回答和推理。具体来说，同样使用基于Transformer (Vaswani et al., 2017)的大型语言模型FLAN-T5 (Chung et al., 2022)作为主干模型。使用文本提示："Context:{G} Question:{Q}"作为文本输入X_{t}。其中，$G = \{sub_1\ obj_1\ rel_1;\ \ldots\ ; sub_\sigma\ obj_\sigma\ rel_\sigma\}$表示在第3.2节中生成的场景图三元组序列。问题$Q$表示为$Q = \{q_1, q_2, \ldots, q_h\}$，$h$是输入问题的长度。对于视觉输入$X_{\mathrm{v}}$，我们使用与构建场景图时相同的CLIP图像补丁序列，并使用第3.1节中的跨模态注意力和门控机制进行模态融合。

需要注意的是，在场景图生成阶段之后，每个图像都会得到一个相应的由一系列的三元组组成的场景图。然而，并非所有的三元组都是准确的，不准确的三元组不仅对生成过程没有正面的帮助，而且可能还引入额外的噪音，对推理造成误导。因此，为了确保生成更准确和合理的推理，筛选出高质量生成的三元组至关重要。本文采用了一种基于阈值的选择方法：对于原始图像I和生成的场景图G，我们使用经过训练的CLIP模型计算每个三元组和图像区域的标准化相似度得分。输出的得分按降序排序，得分越高表示相似度越高，表示它们更符合图像场景。我们为选择过程设定一个阈值，然后从有序集合中持续选择具有最高置信度得分的三元组。即当累计的置信度得分之和小于阈值时，将不断继续选择新的三元组，直到达到阈值为止。

同时，我们将真实标签中的回答和推理以"{A} Because:{R}"的形式组合作为$(Q \rightarrow AR)$生成的训练标签。考虑到VL-NLE的原始设定还包含了两个子任务：$(Q \rightarrow A)$和$(QA \rightarrow R)$，我们也为此构造了相应的文本提示模板。更具体地说，对于子任务$(Q \rightarrow A)$（即输入问题，生成回答），设置"Context:{G} Question:{Q} Answer:"的输入模板和"Answer:{A}"的真值标签格式。对于子任务$(QA \rightarrow R)$（即输入问题和回答，生成推理），设置了"Context:{G} Question:{Q} Answer:{A} Because:"的输入模板和"Explanation:{R}"的真值标签格式。其中，A和R分别表示生成的回答和推理解释。最后，训练中的生成损失被定义为以下的交叉熵损失：

$$\mathcal{L}_{E} = -\sum_{i}^{I} \log P\left(y_i \mid y_{<i}, X_{\mathrm{t}}; X_{\mathrm{v}}\right), \tag{7}$$

其中，y_i是生成的词，X_{t}是包含问题Q和场景图G的文本输入，X_{v}表示视觉输入。

4 实验设置

4.1 实验对比模型

为了评估我们提出的模型框架，我们与以下模型进行比较：

- **e-UG** (Kayser et al., 2021)：e-UG将GPT-2与UNITER结合，使用Faster R-CNN获取区域的视觉特征，并将位置特征编码为视觉输入。通过将图像区域和问题词的嵌入预置到文本问题和预测答案中，并利用UNITER的上下文嵌入输出来微调GPT-2。

- **OFA-X** (Plüster et al., 2022)：OFA-X将OFA (Wang et al., 2022a)模型作为主干模型，它是一种标准的编码器-解码器Transformer架构。给模型提供问题和四个不同的选择，根据模型对每一个选项输出的的可能性选择正确的答案和解释。

- **NLX-GPT** (Sammani et al., 2022)：NLX-GPT是一种编码器-解码器模型，由对图像进行编码的视觉主干模型和精简的GPT-2组成。首先在图像标注任务上预训练精简的GPT-2，然后在VL-NLE数据集上进行微调。

- **UMAE** (Whitehouse et al., 2023)：UMAE同样基于OFA模型。为多个数据集设计了相应的文本提示，并在VL-NLE任务上进行多任务训练。

4.2 评测指标

为了评估生成的场景图的质量，并衡量它们对视觉语言常识推理生成的帮助程度，我们分别对场景图生成和视觉语言常识推理生成在相应指标下进行评估 (Özsoy et al., 2023; Yu et al., 2021)。

对于无位置信息场景图生成，我们采用启发式树搜索（Heuristic Tree Search，HTS）(Özsoy et al., 2023)来评估生成质量，该方法用于计算图之间的重叠程度并获取召回率分数。在预测阶段，生成的场景图以序列形式呈现。我们首先将预测的序列和标签序列转换为图结构，然后使用这种基于树的搜索算法来计算召回率R@50和100。

对于视觉语言常识推理生成，我们采用了不同的评估方法，包括自动评估和人工评估。在自动评估中，我们使用了自然语言生成的评测指标：BLEU (Papineni et al., 2002)，METEOR (Banerjee and Lavie, 2005)，ROUGE-L (Lin, 2004)，CIDEr (Vedantam et al., 2015)，SPICE (Anderson et al., 2016)和BERTScore F1 (Zhang et al., 2020)。同时，还使用了e-ViL指标 (Kayser et al., 2021)进行评测。e-ViL指标由S_T，S_E和S_O组成，分别代表任务准确度，n-gram综合分数，以及总分（S_T和S_E的乘积）。S_E是通过计算ROUGE-L，METEOR，CIDEr，SPICE和额外的BERTScore F1的调和平均数得到的。为了与之前的工作 (Yu et al., 2021; Whitehouse et al., 2023; Sammani et al., 2022; Plüster et al., 2022))在相同的尺度下进行比较，我们同样使用过滤后的生成结果来评估我们模型的性能。即在过滤过程中，采用了0.92的BERTscore阈值来判断回答部分的正确性：BERTscore低于0.92的回答被认为是错误的，而达到0.92的回答被认为是正确的。因此，经过过滤后的回答被认为是正确的，并进一步评估其推理解释。此外，我们还使用了人工评估，方法与之前的工作 (Kayser et al., 2021)相同：随机从测试集中选择300个答案正确的例子，注释者需要从“yes”，“weak yes”，“weak no”和“no”中选择一个作为推理是否能够解释回答的评测结果。

4.3 实验参数设置

在实验中，我们使用训练过的CLIP模型[0]的补丁序列输出作为图像输入，用预训练过的FLAN-T5-large[1]的参数来初始化模型。在场景图生成阶段，我们使用大小为8的批次大小训练模型，共迭代5个轮次。在视觉语言常识推理生成阶段，我们使用大小为16的批次大小训练模型，共迭代4个轮次。在两个阶段中，都使用AdamW优化器 (Loshchilov and Hutter, 2019)来训练模型，学习率都被设定为$5e^{-5}$，热身步数分别设置为1k和2k步。训练中的模型每经过5k步进行一次验证，并保存最好的模型进行评测。实验均在8张NVIDIA RTX 3090 24Gb显卡上进行。

模型	边界框	R@50	R@100
SGTR	✓	30.38	34.85
Pix2SG	✗	24.81	26.66
ours	✗	25.93	43.61

Table 1: Visual Genome数据集上的场景图生成在R@k指标上的实验结果。

5 结果与分析

5.1 场景图生成的实验结果

与大多数场景图生成模型不同，本文提出的方法采用了无位置设定，即不提供边界框等显示的位置信息。因此，我们仅采用启发式树搜索方法对场景图进行评测，结果如表1所示。SGTR (Li et al., 2022)是一种传统方法下的端到端的场景图生成模型，它分别预测实体节点和谓词节点，通过连接两种节点，生成二分图进而构建场景图。Pix2SG (Özsoy et al., 2023)是最早也是目前唯一在无位置设定下提出场景图生成的模型。它基于Pix2Seq骨干网

[0]https://github.com/jianjieluo/OpenAI-CLIP-Feature
[1]https://huggingface.co/google/flan-t5-large

络 (Chen et al., 2022)，这是一种能够将物体检测任务表示为序列预测的多模态模型。实验结果表明，SGEVL与传统的场景图生成模型相比，能够在R@100上得到显著的提升。而在无位置信息的设定下，本文构建的场景图生成模型在R@50和R100上都能够超过Pix2SG模型。这表明赋予多模态能力后的大型语言模型的视觉语义理解能力得到了提升，并能够有效地识别场景中的对象以及其关联信息。

5.2 视觉语言常识推理的实验结果

我们分别在VCR，VQA-X和e-SNLI-VE三个公开数据集上进行了实验，并与基准模型进行了比较，结果如表2所示。下面将根据不同数据集分别进行结果分析。

5.2.1 VCR数据集的实验结果

根据表2所示，现有的一些工作（例如e-UG，OFA-X和NLX-GPT）在VCR数据集上表现出了相近的性能。他们在BLEU和SPICE这样的n-gram指标上得分较低，表明这些模型生成的句子结构与标签相比存在一定差距。与这些模型相比，UMAE的性能取得了更加显著提升，包括将BLEU-4从6.6%提高到13.4%，展现出了更强的句子结构一致性。然而，与其他模型相比，它在S_T指标上的得分下降了大约10%，表明其生成结果的准确性存在欠缺。

本文提出的方法与先前的工作相比取得了显著的提升。具体来说，加入了场景图强化后，模型在METEOR、ROUGE-L、CIDEr和BERTScore上分别达到了20.9、34.9、57.5和90.1的得分，超过了基准模型的表现，这意味着场景图提供的对象和对象间的关系有助于模型生成与标签语义更相近的解释。同时，在e-ViL指标上，SGEVL在S_O和S_E上分别取得了28.3和45.6的分数，表明生成的推理解释能够达到更高的标准。

5.2.2 VQA-X数据集的实验结果

在VQA-X数据集上的实验结果如表2所示。SGEVL在n-gram评测指标和BERTScore指标上取得了相对较高的分数。其中，METEOR和SPICE得到了最高的分数，分别为24.0和24.3，并在S_E分数上超过基准模型，得到了50.6。但是，由于VQA-X数据集中提供的回答和推理的句子长度较短（通常以单词或短语的方式组成），而我们的模型倾向于生成长度更长，更加完整的推理过程，这会间接导致评测分数的下降。因此在结果中本文的方法在个别指标上略低于基线方法。但从综合评估的角度，SGEVL仍然表现出了良好的效果。

5.2.3 e-SNLI-VE数据集的实验结果

e-SNLI-VE数据集上的实验结果也同样表明了我们提出的框架的有效性。根据表2中的结果，本文的提出的方法在所有指标上都取得了较好的分数，特别是在METEOR和ROUGE-L等指标上。此外，与基线工作相比，我们的方法在综合得分S_O上实现了7%的提升。这进一步证实了我们提出的框架能够把握住视觉场景中的关系信息，生成符合场景信息的解释。

5.2.4 人工评测

为了更好地评估视觉语言常识推理的生成质量，本文从测试样集中随机抽取了部分经过BERTScore过滤和未经过过滤的测试样例进行了人工评测。评测人员需要根据生成的推理能否佐证回答，从“yes”，“weak yes”，“weak no”和“no”四个选项中对每一个样例进行评价。评测结果如表3所示。在经过过滤的样例中，有63.1%样例的推理部分被认为很好地证明和解释了回答部分，只有5.4%的样例在推理和解释上没有得到人工评测的认可。这表明从人类认知的角度，绝大部分的样例都是符合常识和具有解释性的。而在未经过过滤的样例中，“weak no”和“weak yes”的比例出现了增长，而被选为“no”的比例依然较小，这也表明在回答不完全正确的情况下，推理的质量依然能够保持一定的稳定性。

5.3 消融实验

5.3.1 融合不同的视觉特征

我们进行了消融实验以了解使用不同视觉特征的效果。结果如表4所示。其中，DETR (Carion et al., 2020)是一个用于目标检测的端到端的基于Transformer的模型。CLIP用于分别提取补丁序列特征和全局特征。可以观察到，相比于补丁序列特征，DETR在这些指标上只显示出轻微的落后，而全局特征在召回率上则表现出明显的不

	e-ViL Scores			*n*-gram Scores								
VCR	S_O	S_T	S_E	B1	B2	B3	B4	M	R-L	C	S	BS
e-UG	19.3	69.8	27.6	20.7	11.6	6.9	4.3	11.8	22.5	32.7	12.6	79.0
OFA-X_{VCR}	23.0	**71.2**	32.4	24.5	14.4	9.1	6.1	12.2	25.1	48.5	18.8	79.8
OFA-X_{MT}	19.2	62.0	30.9	22.3	13.0	8.0	5.2	11.3	24.3	44.6	17.8	79.3
NLX-GPT	-	-	32.6	24.7	15.0	9.6	6.6	12.2	26.4	46.9	18.8	80.3
$UMAE_{VCR}$	22.5	56.6	39.8	-	-	-	12.3	16.7	28.9	48.2	**27.4**	81.8
$UMAE_{MT}$	22.8	56.6	40.2	31.4	22.9	17.6	**13.4**	17.5	29.5	47.3	26.5	81.9
SGEVL(ours)	**26.1**	61.9	**45.3**	**35.7**	**24.2**	**17.3**	12.9	**20.7**	**34.5**	**55.6**	**27.4**	**89.4**
VQA-X												
e-UG	36.4	80.5	45.3	57.3	42.7	31.4	23.2	22.1	45.7	74.1	20.1	87.0
OFA-X_{MT}	**45.5**	**92.6**	49.2	64.0	49.4	**37.6**	**28.6**	23.1	51.0	110.2	22.6	86.8
NLX-GPT	40.6	83.0	49.0	**64.2**	**49.5**	**37.6**	28.5	23.1	**51.5**	**110.6**	22.1	**86.9**
$UMAE_{MT}$	31.5	77.6	40.6	47.5	31.4	21.4	14.6	20.2	35.1	50.3	19.1	85.4
SGEVL(ours)	43.5	87.5	**49.8**	**64.2**	48.6	36.4	27.0	**23.9**	44.8	103.8	**23.1**	86.7
e-SNLI-VE												
e-UG	36.0	**79.5**	45.3	30.1	19.9	13.7	9.6	19.6	27.8	85.9	34.5	81.7
OFA-X_{MT}	35.6	78.9	45.1	32.4	21.8	15.2	10.8	17.9	31.4	108.2	32.8	80.4
NLX-GPT	34.6	73.9	46.9	37.0	25.3	17.9	12.7	18.8	34.2	**117.4**	33.6	80.8
SGEVL(ours)	**41.7**	73.9	**56.5**	**39.8**	**27.5**	**17.9**	**12.8**	**24.9**	**39.7**	114.7	**43.9**	**89.3**

Table 2: VCR，VQA-X和e-SNLI-VE数据集上的过滤评测结果。B，M，R-L，C，S和BS分别表示BLEU，METEOR，ROUGE-L，CIDEr，SPICE和BERTScore F1。模型下标的**VCR**表示该模型仅在VCR数据集上训练，而**MT**表示该模型经过多任务训练。

	No	Weak No	Weak Yes	Yes
过滤后	5.4%	5.2%	26.3%	63.1%
过滤前	9.5%	29.6%	29.1%	31.8%

Table 3: 人工评测结果。“过滤后”表示仅对回答正确的样例进行推理生成的评测。“过滤前”表示同时对所有样例的回答和推理的评测。

足。我们认为这主要是因为在获取全局特征的过程中，由于维度空间的的压缩导致了信息的丢失，而补丁序列特征可以为大型语言模型提供更原始，更完整的视觉信息。

为了验证我们的模态融合机制的有效性，我们还对是否使用模态融合机制进行了进一步的分析，并通过表5补充了我们的研究。其中，“$w/o\ I$”表示仅加入了场景图，未加入图像和模态融合机制。实验结果显示，视觉模态的融合显著地提升了结果，进一步证明了其有效性。

5.3.2 选择不同的场景图筛选阈值

我们还对不同的场景图选择方法进行了消融实验。结果显示在表6中。可以观察到，当阈值设定为0.8时，过滤后的评测结果可以达到最优的性能。然而，当阈值降低或提高时，性能显示出下降趋势，甚至低于不使用场景图的水平。不仅如此，提供较少的场景图将会导致场景信息的不足，而提供更多质量较差的场景图将导致噪声的增加，这都将降低最终推理生成的表

视觉特征	维度	R@20	R@50	R@100
DETR	(100, 256)	10.07	**26.00**	43.39
$CLIP_{global}$	(1, 512)	6.39	15.82	27.56
$CLIP_{patch}$	(49, 2048)	**10.92**	25.93	**43.61**

Table 4: 使用不同视觉特征输入在Visual Genome数据集上进行场景图生成的消融实验。

	S_O	S_T	S_E	B1	B4	R-L	BS
SGEVL ($w/o\ I$)	13.7	43.3	31.7	27.9	5.6	25.0	77.8
SGEVL	26.1	61.9	45.3	35.7	12.9	34.5	89.4

Table 5: 模态融合的消融实验结果。

方法	B1	B2	B3	B4	M	R-L	C	S	BS
SGEVL (w/o 场景图)	35.3	23.8	16.9	12.5	**20.7**	**34.5**	53.7	26.8	**90.0**
SGEVL ($w/$ 0.7)	35.4	23.7	16.8	12.3	20.0	33.6	49.8	26.5	89.8
SGEVL ($w/$ 0.8)	**35.7**	**24.2**	**17.3**	**12.9**	**20.7**	**34.5**	**55.6**	**27.4**	**90.0**
SGEVL ($w/$ 0.9)	35.3	23.8	16.9	12.4	20.0	33.8	50.6	26.5	88.5
SGEVL ($w/$ 1.0)	34.9	22.9	16.2	11.7	19.8	33.8	49.9	26.4	88.6

Table 6: 使用不同场景图筛选阈值的消融实验。

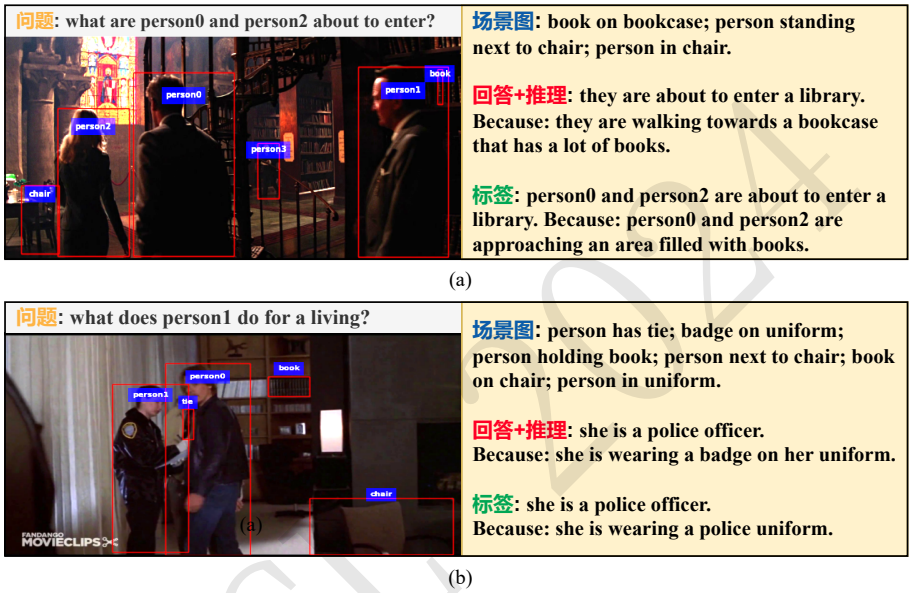


Figure 3: SGEVL在VCR数据集上的案例分析。

现。生成的场景图的质量对推理的质量存在相关性，筛选更准确的场景图有助于提升推理的合理性。因此，根据结果本文选用0.8作为最终的筛选阈值。

5.4 可视化分析

我们对部分样例进行可视化并为进行了分析，以对所提出的方法有一个更加清晰的认知。我们在VCR数据集上的可视化结果如图3所示。可以看到，生成的场景图三元组为核心对象之间的详细关系提供了信息。在图3(a)的例子中，生成的“book on bookcase”增强了这是一个图书馆的场景的可能性，推理模型正确地应用了这个三元组生成了“they are about to enter a library. Because: they are walking towards a bookcase that has a lot of books”。生成的推理能够较好地反映场景中的情况，能够提供关键的信息，并对回答生成符合常理的解释。图3(b)中展示的例子中，推理模型通过生成的“person has tie”和“person in uniform”三元组，帮助推理模型生成了一个与实际情况密切匹配的推理解释。

6 结论

本文主要研究了如何增强模型在视觉语言常识推理能力的问题。为解决在常识推理过程中大型语言模型缺乏对场景关系的感知的问题，本文提出了一个场景图增强的的推理生成框架。这个解耦的框架将无位置信息场景图生成和视觉语言常识推理生成分离为两个阶段。在多个数

据集上的实验显示出本文的方法在与基线方法的比较中取得了更好的效果，证明了本文的方法所生成的场景图有助于产生更合理的推理，从而解释和验证回答的准确性。不仅如此，本文还针对不同模块进行了消融实验，以验证框架中各个部分的有效性。在未来的工作中，我们将探索更全面的视觉关系信息挖掘方法，并加强模型的逻辑推理能力。

7 致谢

该研究得到了国家自然科学基金（62106105）、CCF-百度松果基金(CCF-Baidu202307)、CCF-智谱大模型基金(CCF-Zhipu202315)、南京航空航天大学科研启动基金（YQR21022）和南京航空航天大学高性能计算平台的支持。

参考文献

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, volume 9909 of *Lecture Notes in Computer Science*, pages 382–398. Springer.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss, editors, *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: universal image-text representation learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer.

Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geoffrey E. Hinton. 2022. Pix2seq: A language modeling framework for object detection. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Bo Dai, Yuqi Zhang, and Dahua Lin. 2017. Detecting visual relationships with deep relational networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3298–3308. IEEE Computer Society.

Helisa Dhamo, Azade Farshad, Iro Laina, Nassir Navab, Gregory D. Hager, Federico Tombari, and Christian Rupprecht. 2020. Semantic image manipulation using scene graphs. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5212–5221. Computer Vision Foundation / IEEE.

Helisa Dhamo, Fabian Manhardt, Nassir Navab, and Federico Tombari. 2021. Graph-to-3d: End-to-end generation and manipulation of 3d scenes using scene graphs. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 16332–16341. IEEE.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Radhika Dua, Sai Srinivas Kancheti, and Vineeth N. Balasubramanian. 2021. Beyond VQA: generating multi-word answers and rationales to visual questions. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, pages 1623–1632. Computer Vision Foundation / IEEE.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3668–3678. IEEE Computer Society.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1224–1234. IEEE.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73.

Ranjay Krishna, Vincent S. Chen, Paroma Varma, Michael S. Bernstein, Christopher Ré, and Li Fei-Fei. 2019. Scene graph prediction with limited labels. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2580–2590. IEEE.

Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiaoou Tang. 2017a. Vip-cnn: Visual phrase guided convolutional neural network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 7244–7253. IEEE Computer Society.

Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. 2017b. Scene graph generation from objects, phrases and region captions. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1270–1279. IEEE Computer Society.

Rongjie Li, Songyang Zhang, and Xuming He. 2022. SGTR: end-to-end scene graph generation with transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 19464–19474. IEEE.

Wentong Liao, Bodo Rosenhahn, Ling Shuai, and Michael Ying Yang. 2019. Natural language guided visual relationship detection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 444–453. Computer Vision Foundation / IEEE.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, volume 9905 of *Lecture Notes in Computer Science*, pages 852–869. Springer.

Ege Özsoy, Felix Holm, Tobias Czempiel, Nassir Navab, and Benjamin Busam. 2023. Location-free scene graph generation. *CoRR*, abs/2303.10944.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2016. Attentive explanations: Justifying decisions and pointing to the evidence. *CoRR*, abs/1612.04757.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8779–8788. Computer Vision Foundation / IEEE Computer Society.

Björn Plüster, Jakob Ambsdorf, Lukas Braach, Jae Hee Lee, and Stefan Wermter. 2022. Harnessing the power of multi-task pretraining for ground-truth level natural language explanations. *CoRR*, abs/2212.04231.

Yue Qiu, Yoshiki Nagasaki, Kensho Hara, Hirokatsu Kataoka, Ryota Suzuki, Kenji Iwata, and Yutaka Satoh. 2023. Virtualhome action genome: A simulated spatio-temporal scene graph dataset with consistent relationship labels. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pages 3340–3349. IEEE.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.

Fawaz Sammani, Tanmoy Mukherjee, and Nikos Deligiannis. 2022. NLX-GPT: A model for natural language explanations in vision and vision-language tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 8312–8322. IEEE.

Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. 2020a. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020b. Unbiased scene graph generation from biased training. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3713–3722. Computer Vision Foundation / IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Tzu-Jui Julius Wang, Selen Pehlivan, and Jorma Laaksonen. 2020. Tackling the unannotated: Scene graph generation with bias-reduced models. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022a. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR.

Zhecan Wang, Haoxuan You, Liunian Harold Li, Alireza Zareian, Suji Park, Yiqing Liang, Kai-Wei Chang, and Shih-Fu Chang. 2022b. SGEITL: scene graph enhanced image-text learning for visual commonsense reasoning. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 5914–5922. AAAI Press.

Chenxi Whitehouse, Tillman Weyde, and Pranava Madhyastha. 2023. Towards a unified model for generating answers and explanations in visual question answering. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1648–1660. Association for Computational Linguistics.

Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3097–3106. IEEE Computer Society.

Shaotian Yan, Chen Shen, Zhongming Jin, Jianqiang Huang, Rongxin Jiang, Yaowu Chen, and Xian-Sheng Hua. 2020. PCPL: predicate-correlation perception learning for unbiased scene graph generation. In Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann, editors, *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 265–273. ACM.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 3208–3216. AAAI Press.

Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. 2020. Weakly supervised visual semantic parsing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3733–3742. Computer Vision Foundation / IEEE.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE.

Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. 2017. PPR-FCN: weakly supervised visual relation detection via parallel pairwise R-FCN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 4243–4251. IEEE Computer Society.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *CoRR*, abs/2302.00923.