arXiv:2301.01949v1 [cs.CL] 5 Jan 2023

SPRING：基于增量布局图中的多模态问题预训练的情境对话代理

Yuxing Long [1], **Binyuan Hui** [2], **Fulong Ye** [1], **Yanyang Li** [2], **Zhuoxin Han** [1], **Caixia Yuan** [1], **Yongbin Li** [2], **Xiaojie Wang** [1*]
[1] 北京邮电大学，中国北京 [2] 独立研究者
{longyuxing, fulongye, hanzhuoxin, yuancx, xjwang}@bupt.edu.cn
lyb821@gmail.com

摘要

现有的多模态对话智能体已展现出在简单场景中定位绝对位置或检索属性的出色能力，但在涉及复杂相对位置和信息对齐时表现不佳，这成为响应质量的瓶颈。本文提出一种**S**情境对话智能体**PR**基于多模态问题预训练的**IN**增量布局**G**图（**SPRING**），使其能够推理多跳空间关系，并在拥挤的情境场景中将这些关系与视觉属性联系起来。具体而言，我们设计了两类多模态问答（MQA）任务来预训练该智能体。预训练期间使用的所有问答对均由新颖的增量布局图（ILG）生成。利用 ILG 自动标注的问答对难度标签，促进基于 MQA 的课程学习。实验结果验证了 SPRING的有效性，表明其在 SIMMC 1.0 和 SIMMC 2.0 数据集上均显著优于最先进的方法。我们在以下地址发布代码和数据：
https://github.com/LYX0501/SPRING。

1 引言

构建能够在视觉情境中与人交流的多模态对话智能体，是人工智能领域的一项重要目标。近年来，人们提出了许多用于视觉对话的特定任务和数据集，例如 VisDial（Das 等人，2017年）、GuessWhat（De Vries 等人，2017年）和 GuessWhich（Chattopadhyay 等人，2017年）。其中，情境交互式多模态对话（SIMMC 1.0）（Moon 等人，2020年）旨在研究面向任务的对话，这类对话涵盖了以虚拟现实（VR）环境为形式的情境化多模态用户上下文。更新后的数据集 SIMMC 2.0（Kottur 等，2021b）为多模态对话智能体提供了更具挑战性的测试平台。每张图像中都有许多布局复杂的物品。图1展示了 SIMMC 2.0 中的一个场景和一段对话片段。图中有几十件衣物。每件衣物都是一个数字资产，具有唯一的资产 ID，以及元数据中的一组属性（例如类型、颜色）。然而，除了标注物品之间上方、下方、左方和右方这四种关系的少量标签外，并没有关于场景布局的信息。

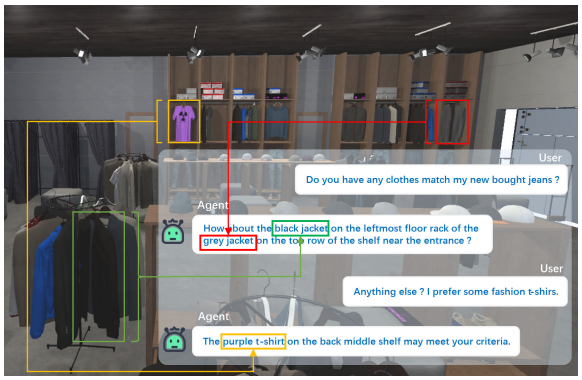

图1：SIMMC 2.0 中虚拟场景及一段对话示例。由于许多衣物的视觉属性相似，仅凭视觉属性很难描述某件物品。在这种情况下，布局中的空间关系必不可少。

已有不少研究基于 SIMMC 2.0 展开。以不同的多模态视觉语言预训练模型（VLM）为基础，先前的研究者更加关注学习物品的视觉属性。QS Goal Diggers（Kottur 等人，2021a）和 Kakao Enterprise（Lee 和 Han，2021）将视觉属性直接输入模型，而西江大学（Kottur 等，2021b）和 A-STAR（Nguyen 等，2021）则在预训练阶段构建了一组视觉属性预测任务。KAIST（Lee等，2022）设计了一项辅助任务来预测视觉属性。然而，物品之间空间关系的构建却较少受到关注。现有模型都只将资产边界框的坐标用作位置信息，无法捕捉场景中的空间关系。

因此，模型可以准确生成“黑色夹克”，但无法描述更自然的指代表达，例如“最左侧落地式货架上的黑色夹克”。显然，在包含许多属性相似衣物的场景中，后者更有用。属性与空间关系相结合，可以帮助人们快速找到物品。要生成这类表达，模型不仅需要学习每件物品的视觉属性，还需要学习不同物品之间的空间关系。

为解决上述问题，我们提出了 SituatedConver- sation Agent **PR**etrained with Multimodal Questions from

*通讯作者
版权所有 © 2023，人工智能促进协会 (www.aaai.org)。保留所有权利。

INcremental Layout**G**raph (**SPRING**)，即 SPRING。该模型使用从增量布局图生成的多模态问题进行预训练。在我们的方法中，我们为每个场景设计了 **增量布局图（ILG）**，以捕捉不同场景元素之间丰富的空间关系。与场景图（Chang 等，2020）不同，ILG 仅使用纯文本信息构建，并且可以随着新增对话逐步扩展。随后，通过遍历 ILG 上的节点（数字资产和背景物品），收集两类 **多模态问答（MQA）**预训练任务及其对应的问答对。根据路径跨度长度，可以自动为问答对标注难度级别。因此，我们采用课程学习（Bengio 等，2009），在 Transformer（Vaswani 等，2017）编码器-解码器骨干网络上进行预训练。在 SIMMC 1.0 和 SIMMC 2.0 上的实验表明，与此前表现最佳的模型相比，我们的方法显著提升了响应质量。

我们工作的主要贡献如下：

- 我们首先提出一种新颖的方法，可根据对话文本逐步构建虚拟场景的 ILG。ILG 包含带有空间关系的场景物品。值得注意的是，这一过程不依赖任何人工标注。
- 基于 ILG，我们提出两类新的 MQA预训练任务，以帮助模型理解视觉元数据以及不同资产之间的空间关系。通过遍历 ILG 自动生成预训练样本，同时也会为课程学习生成相应的难度标签。
- 我们进行了全面的实验，以验证所提方法能够有效提升响应质量。在 SIMMC 1.0 和 SIMMC 2.0 的所有指标上，我们的方法始终大幅优于现有最先进方法。

2 相关工作

情境交互式多模态对话。近年来，对话系统发展迅速，例如，面向任务的对话预训练（He 等人，2022c,a,b）、知识型对话（Hui 等人，2022；Wang 等人，2022a）等。其中，多模态对话是新的趋势。META 发布了基于虚拟现实购物商店的多模态对话数据集 SIMMC（Kottur 等，2021b），其中包含数百张来自不同角度的场景快照。与以往的多模态对话数据集 MMD（Saha、Khapra 和 Sankaranarayanan，2018年）和 VisDial（Das 等人，2017年）相比，情境智能体需要生成更复杂的视觉属性和更详细的空间关系，以推断场景中的数字资产。Kottur 等人（2021a）对利用视觉属性和空间关系进行了初步探索。具体而言，DialVinVL（Kottur 等人，2021a）将有关视觉属性的槽位值与对话历史结合，作为文本输入；并将原始边界框坐标与目标检测器提取的区域特征拼接，作为视觉输入。JMGPT（Kottur 等，2021b）和 JointGM（Nguyen 等，2021）使用语言模型联合预测视觉属性和系统响应。MMBart（Lee等，2022）则添加了嵌入的边界框坐标作为 Transformer 输入的文本嵌入，并设计辅助任务，根据编码器隐藏状态的输出预测视觉属性。我们发现，它们所使用的空间信息都来自边界框。与这些方法不同，我们首先注意到视觉语言模型在视觉和空间能力方面的不足，随后提出了基于增量布局图的多模态问答预训练任务；该类布局图已成功应用于相关研究（Qiu 等人，2021；Liao 等人，2021；Hui 等人，2021；Qiu 等人，2022）。

视觉语言预训练。为了提升模型对文本和图像的感知能力，并帮助其建立多模态信息之间的联系，研究者设计了多种视觉语言预训练模型。ViLBERT (Lu et al. 2019) 和 UNITER (Chen et al. 2020) 提出将检测器的原始输出——对象类别的分布——视为软标签，并优化两个分布之间的 KL 散度。LEXMERT (Tan and Bansal 2019) 和 UNIMO (Li et al. 2021) 提出掩码区域特征回归（MRFR），将被掩码的区域特征回归到其对应的原始区域特征；其中，图像由 Faster R-CNN (Ren et al. 2015) 表示为区域特征序列。此外，SOHO (Huang et al. 2021 b) 旨在避免图像转换为网格特征或图块特征时来自邻近特征的信息泄漏。

近年来，CLIP（Radford 等，2021年）和 ALIGN（Jia 等，2021年）利用大规模图文对学习可迁移的视觉表征，并在图像分类任务上展现出令人惊讶的零样本迁移能力。VL-T5（Cho 等，2021年）和 OFA（Wang 等，2022年）将视觉定位和视觉定位描述等下游任务纳入预训练任务，以缩小预训练与微调之间的差距。不同于这些工作，我们通过统一的问答范式设计新的预训练任务，以改进现有方法对视觉属性和空间关系的建模，而无需添加新模块。

3 方法

设 $\mathcal{D} = \{(U_t,\ R_t,\ I_t)\}_{t=1}^{T}$ 为用户与智能体之间正在进行的对话，共有 T 轮。其中，U_t 表示时间步 t的用户话语，R_t 表示智能体针对 U_t 给出的语言回复，I_t 表示随附的场景图像。本文的任务是根据对话历史 $H_t = [U_i, R_i]_{i=1}^{t-1}$、当前用户话语$U_t$ 和场景图像 I_t，预测最优语言回复 $R_t^{'}$，如公式（1）所示。

$$R_t^{'} = argmax\, P_\theta\left(R_t \mid H_t, U_t, I_t\right) \tag{1}$$

其中 θ 是模型的可学习参数。

为解决上述问题，我们提出了一种多模态对话模型 **SPRING**，该模型通过增量布局图生成的多模态问题进行预训练。接下来的章节将依次介绍模型架构、ILG生成和MQA预训练任务。

3.1 架构

SPRING 模型的骨干网络是基于编码器-解码器的单流视觉语言模型框架，由多个 Transformer（Vaswani 等人，2017年）层堆叠而成。场景图像 $I_t \in$

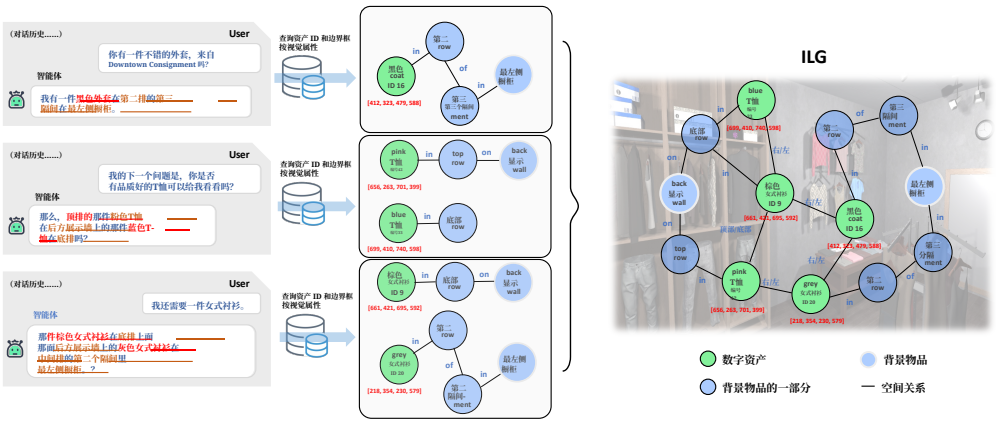

图2：从对话中构建增量布局图。数字资产和背景物品构成 ILG 节点，而空间关系构成 ILG边。随着同一场景中不断加入新的对话，ILG 也会持续增量更新。

$\mathbb{R}^{h\times w\times c}$ 被切分为 P个图像块。然后，每个图像块都会被投影为模型隐藏维度的视觉嵌入。对话历史和当前用户话语经过字节对编码（BPE）转换为子词序列，然后嵌入为文本嵌入。所有视觉嵌入和文本嵌入会拼接在一起，作为模型输入。

为了让**SPRING** 更好地理解具身场景的信息，我们提出了一系列基于布局图的多模态问答预训练任务 $\mathcal{G}_i$。由于 SIMMC 数据集中没有标注的布局图，我们提出了一种基于自然语言对话历史的无监督 ILG 构建方法。

3.2 增量布局图（ILG）

我们观察到，对话历史中包含视觉属性和空间描述。与数据集标注相比，对话中的信息更加详细。例如，SIMMC 2.0 的标注仅提供数字资产的边界框，以及它们之间四种类型的相对位置（上方、下方、右方、左方）；而对话则包含背景物品的绝对位置，以及它们与物品之间的相对位置。一个关键发现是，对话历史与响应之间存在共指关系，即待生成的响应中出现的同一物品。

因此，我们提出了一种 ILG 生成算法，从对话中提取高质量信息，并生成增量布局图（ILG）$\mathcal{G}_i = \langle \mathcal{V}_i, \mathcal{E}_i \rangle$ 来表示这些信息，其中 $\mathcal{V}_i$ 表示节点集，包含对话历史中的数字资产和背景物品；$\mathcal{E}_i$ 表示边集，用于描述场景元素之间的空间关系 $\mathcal{V}_i$。

文本信息提取与对齐 我们考虑采用一种基于规则的文本信息提取方法，即，正则表达式，从对话历史中提取视觉属性和空间描述，无需人工标注。用于**视觉属性**和**空间描述**的正则表达式 RegExp_{va}和 RegExp_{sd} 如下。

$$\text{RegExp}_{va} = \text{(art.) (color) (asset type)} \quad (2)$$

$$\text{RegExp}_{sd} = \text{(positional prep.) (art.) (.*?) (punc.)} \quad (3)$$

其中，art. 表示冠词，prep. 表示介词，punc. 表示标点符号。详情请参阅附录。借助这两个正则表达式，如图2左侧所示，我们可以从对话历史中提取视觉属性“黑色外套”和空间描述“在最左侧橱柜的第三个隔间的第二排”。

尽管由于语言特征，通过上述正则表达式提取的视觉属性和空间描述天然对齐，但它们与资产 ID 并不对齐，因此资产框坐标无法使用。为了解决这个问题，我们根据资产 ID 从数据库中查询资产的颜色和类型，以组合成诸如 “黑色外套” 这样的视觉属性，然后尝试将其与提取出的视觉属性（如“黑色外套”）配对。如果这两项视觉属性匹配，就可以确定相应的资产 ID“16”，进而获得配对的资产 ID、视觉属性和空间描述。我们进一步设计了以下两个正则表达式，从提取出的空间描述中提取**背景物品**和**相对空间关系** 。

$$\text{RegExp}_{bi} = \text{(background item)} \quad (4)$$

$$\text{RegExp}_{sr} = \text{(positional prep.)} \quad (5)$$

其中，RegExp_{bi} 和 RegExp_{sr} 表示用于背景物品和空间关系的正则表达式，prep. 表示介词。借助这两个正则表达式，如图2中间部分所示，我们可以从先前获得的空间描述中提取背景物品“第二排”、“第三个隔间” 和“最左侧橱柜”，以及相对空间关系“在……里面”、“……的”。

增量布局图生成 从对话历史样本中提取丰富的信息后，可以生成布局子图，如图2中部所示。在布局子图中，数字资产节点存储其视觉属性，例如“黑色外套”以及资产 ID“16”；而背景物品节点则存储诸如“第二行”、“第三个隔间”和“最左侧橱柜”这样的背景物品。

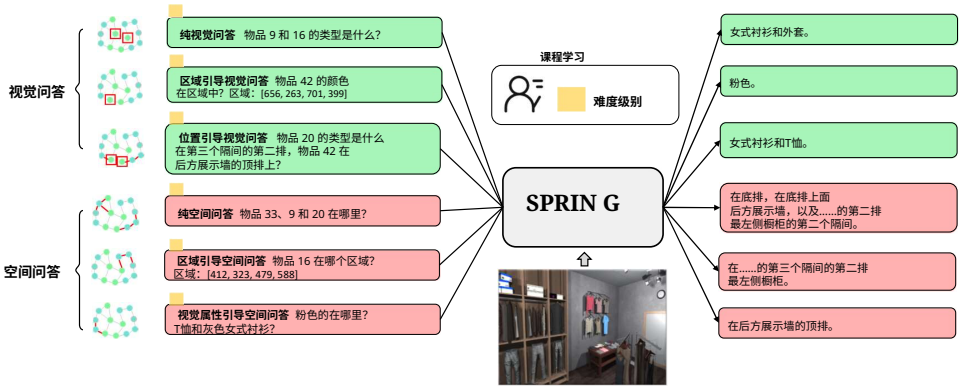

图3：SPRING模型和两种多模态问答预训练任务（视觉问答与空间问答）的示例。基于问答对难度级别的课程学习激发了多模态问答任务的潜力。

场景图像	问答类型	问题模板	答案
	纯视觉问答	什么是 **[视觉属性类型]** 物品 **[资产 ID]**?	**[视觉属性值]**
	区域引导视觉问答	什么是 **[视觉属性类型]** 物品 **[资产 ID]** 在区域中? 区域：**[x1、y1、x2、y2]**	**[视觉属性值]**
	位置引导视觉问答	物品 **[视觉属性类型]**.....的物品 **[资产 ID]** **[位置]**?	**[视觉属性值]**
	纯空间问答	物品 **[资产 ID]**在哪里?	**[位置]**
	区域引导空间问答	物品 **[资产 ID]** 在区域的什么位置? 区域：**[x1, y1, x2, y2]**	**[位置]**
	视觉属性引导空间问答	位置在哪里? **[物品颜色]** **[物品类型]** **[资产 ID]**?	**[位置]**

表 1：问答对模板。方括号“[∗]”表示通过遍历 ILG 进行填充的槽位 s.

空间关系和查询得到的边界框用于定义布局子图边。如图2右侧所示，场景ILG会随着有关同一场景的新子图不断加入而持续扩展，最终可涵盖该场景中的所有数字资产、背景物品及它们之间的空间关系。挖掘ILG中的信息简单而有效。通过遍历ILG 节点，可以轻松获取视觉属性；沿着ILG边遍历，则可推断出多种类型的空间关系。

3.3 基于 ILG 的多模态问答预训练任务

为提升视觉属性和空间关系的响应生成质量，我们基于多模态问答（MQA）设计了视觉问答预训练任务和空间问答预训练任务，二者分别包含三种新型子任务。如图3和表 1 所示，所有问答对均通过遍历 ILG 并填充相应模板自动生成。问答对生成算法如算法1所示。形式上，我们使用问题模板填充函数 $Q_{type}(\cdot)$ 生成问题，A_{type} 表示相应的答案，Type_{va} 表示视觉属性类型，ID_{asset} 表示资产 ID，I_{scene} 是场景图像，BBox_{asset} 表示物品区域坐标，t_{sr} 表示空间关系，t_{va} 是视觉属性，t_{bi} 表示背景物品。

3.3.1 视觉问答

纯视觉问答（PVQA）作为最基础的视觉问答任务，纯视觉问答旨在帮助模型在提供场景图像时，建立资产 ID 与相应视觉属性之间的联系。我们设计了 PVQA 模板，其中问题会提示视觉属性的类型以及资产 ID。纯视觉问题可以通过遍历 ILG 的资产节点并将其填入模板中的[资产ID］槽位来生成，而答案则依据资产节点中存储的视觉属性生成。PVQA 任务的目标如下。

$$L_\theta = -\sum_{i=1}^{N} \log P_\theta\left(A_{pv} \mid Q_{pv}(\text{Type}_{va}, \text{ID}_{asset}), I_{scene}\right) \quad (6)$$

区域引导视觉问答（RVQA）为了提升模型根据区域视觉上下文定位物品并描述其视觉属性的能力，我们基于 PVQA 设计了 RVQA 模板，其中问题由资产区域坐标和资产 ID 引导。通过遍历 ILG 的资产节点，并将 ［资产 ID］，边界框坐标`[x1, y1, x2, y2]` 填入模板中的槽位，即可生成区域引导视觉问题。对应的答案则根据资产节点中存储的视觉属性生成。RVQA 任务的目标如下。

$$L_\theta = -\sum_{i=1}^{N} \log P_\theta\left(A_{rgv} \mid Q_{rgv}(\text{Type}_{va}, \text{ID}_{asset}, \text{BBox}_{asset}), I_{scene}\right) \quad (7)$$

位置引导视觉问答（PoVQA）在对话中，无论是理解用户话语还是提供推荐，智能体都必须依据物品的空间信息来定位物品，而不是依赖区域坐标。为了让问题更贴近真实对话，我们将 RVQA 中的区域坐标替换为空间关系，设计了 PoVQA 模板。对于位置引导视觉问答模板，［物品ID］槽位可通过遍历 ILG 的资产节点来填充，而 ［位置］槽位则由资产节点与背景物品节点之间的空间关系路径填充。对应的答案根据资产节点中存储的视觉属性生成。PoVQA 任务的目标如下。

$$L_\theta = -\sum_{i=1}^{N} \log P_\theta\left(A_{pgv} \mid Q_{pgv}(\text{Type}_{va}, \text{ID}_{asset}, t_{sr}), I_{scene}\right) \quad (8)$$

算法1：问答对生成

输入：
ILG $\mathcal{G}_i = \langle \mathcal{V}_i, \mathcal{E}_i \rangle$，问答模板列表 T
输出：
问答对列表 QA，难度标签列表 DL
初始化问答对列表 QA 和难度标签列表 DL
对于 $node$ 在 $\mathcal{E}_i$ **执行**
  如果 $TypeOf(node) =$ "$background\ item$" **则**
    跳过 $node$
  /* Get 来自数字资产节点的信息 */
  $(t_{va},\ ID_{asset},\ BBox_{asset}) \leftarrow GetInfo(\mathcal{G}_i,\ node)$
  /* Walk 从节点获取空间关系 */
  $(t_{bi},\ t_{sr}) \leftarrow Walk(\mathcal{G}_i,\ node)$
  $t_{slot} \leftarrow (t_{va}, t_{bi}, t_{sr}, BBox_{asset}, ID_{asset})$
  for $template$在.....里面 T **do**
    /* 填写模板 */
    $(qa,\ dl) \leftarrow FillIn(template,\ t_{slot})$
    添加 $QA \leftarrow qa, DL \leftarrow dl$

3.3.2 空间问答

纯空间问答（PSQA）作为最基础的空间问答任务，PSQA 旨在帮助模型在给定场景图像时，建立资产 ID 与相应空间关系之间的联系。我们设计了 PSQA 模板，其中问题仅提示"在哪里"以及资产 ID。纯空间问题可通过遍历 ILG 的资产节点并填充模板中的 [物品ID] 槽位来生成，而答案则根据背景物品节点与资产节点之间的空间关系路径生成。PSQA 任务的目标如下。

$$L_\theta = -\sum_{i=1}^{N} \log P_\theta \left(A_{ps} \mid Q_{ps}(\mathrm{ID}_{asset}), I_{scene} \right) \quad (9)$$

区域引导空间问答（RSQA）为了提升模型根据区域视觉上下文定位物品并描述其空间关系的能力，我们基于 PSQA 设计了 RSQA 模板，其中问题由物品区域坐标和资产 ID 引导。通过遍历 ILG 的资产节点，并将模板中的槽位 [资产 ID]，边界框坐标 [x1, y1, x2, y2] 填入，即可生成区域引导的视觉问题。相应的答案则根据背景物品节点与资产节点之间的空间关系路径生成。RSQA 任务的目标如下。

$$L_\theta = -\sum_{i=1}^{N} \log P_\theta \left(A_{rgs} \mid Q_{rgs}(\mathrm{ID}_{asset}, \mathrm{BBox}_{asset}), I_{scene} \right) \quad (10)$$

视觉属性引导空间问答（VSQA）在对话中，无论是理解用户话语还是提供推荐，智能体都必须根据物品的视觉属性而非区域坐标来定位物品。为了让问题更贴近真实对话，我们通过将 RSQA 中的区域坐标替换为视觉属性（例如颜色、类型），设计了 VSQA 模板。对于空间引导的视觉问题模板，[物品编号] 槽位可通过遍历 ILG 的资产节点来填充，而 [物品颜色] 和[物品类型] 槽位则由资产节点中存储的视觉属性填充。对应的答案根据背景物品节点与资产节点之间的空间关系路径生成。VSQA 任务的目标如下所示。

$$L_\theta = -\sum_{i=1}^{N} \log P_\theta \left(A_{vags} \mid Q_{vags}(\mathrm{ID}_{asset}, t_{va}), I_{scene} \right) \quad (11)$$

3.4 基于多模态问答的课程学习

难度级别自动标注通过遍历 ILG 生成问答对时，可以记录路径经过的节点数量。路径经过的节点越多，相应问答对包含的场景信息就越多，这意味着多模态对话模型需要进行更多步推理。因此，我们根据路径经过的节点数量，自动标注每个问答对的难度级别。例如，生成问题 *"棕色夹克 83 & 1055 在哪里？"*及答案 *"它在入口附近的落地式货架上。"*时，ILG 中经过了一个资产节点*"棕色夹克 83 & 1055"* 和两个背景物品节点。该问答对的难度级别标注为 3。形式化表达如下。

$$d = \frac{|V_{spanned}|}{D} \quad (12)$$

其中，d 表示问答对的归一化难度级别，$|V_{spanned}|$ 表示对应路径跨越的 ILG 节点数量，D 是数据集中问答对路径所跨越的 ILG 节点数的最大值。

预训练策略 借助自动标注的难度标签，我们提出了基于多模态问答的课程学习，以激发所设计的 MQA预训练任务的潜力。我们将模型能力 c 定义如下。

$$c(t) = \gamma \sqrt{\alpha \tfrac{t}{T} + \beta \left(1 - \tfrac{t}{T}\right) \min^2(d)} \quad (13)$$

其中，t 是当前训练步骤的索引，T 表示训练步骤的最大数量，$\min^2(d)$表示难度级别 d的最小值，α 和 β 是超参数，γ由 α确定，具体为 $\sqrt{\frac{1}{\alpha}}$这里我们将 α 设为 1.2，将 β 设为 0.8。在给定的训练步骤 t，难度小于或等于 $c(t)$（即 $d \leq c(t)$）的问答对将被采样用于训练。因此，我们的预训练策略在早期阶段侧重于难度较低的问答对，旨在帮助模型形成对场景元素的初步感知和推理能力。在中后期阶段，将加入难度更高的问答对，以提升模型为多个物品生成视觉属性和空间关系的能力。

完成多模态问答预训练后，**SPRING** 模型会在 SIMMC 响应生成任务上进行微调。自回归语言建模目标如下。

$$L_\theta = -\sum_{i=1}^{N} \log P_\theta \left(R_i \mid H_i, U_i, I_i \right) \quad (14)$$

其中 N 表示训练样本总数。

4 实验

4.1 设置

数据集。为了评估所提出模型的性能，我们首先在广泛使用的情境多模态对话数据集 SIMMC 1.0 和 SIMMC 2.0 上开展实验。

模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	视觉	空间
SIMMC 1.0									
MN-MAG（Kim 等，2021年）	27.28	16.75	12.32	9.50	16.62	32.35	0.8694	9.49	9.10
Tom（Jeong 等，2021年）	28.95	18.81	14.23	11.10	18.83	38.18	1.5014	11.13	10.17
JBi-encoder（Huang 等，2021a）	26.76	16.76	12.49	9.60	17.65	36.46	1.2345	9.73	9.43
SPRING（我们的方法）	**32.46**	**22.15**	**17.23**	**13.77**	**20.75**	**40.51**	**1.6329**	**13.53**	**12.60**
SIMMC 2.0									
MTN（Kottur等，2021b）	62.38	44.52	32.90	21.70	21.38	38.50	1.1207	19.91	14.95
JMGPT（Kottur等，2021b）	51.05	35.03	24.66	19.20	14.73	29.18	0.7738	13.67	11.54
JMGPT-BS（Kottur 等人，2021a）	64.86	48.86	37.91	28.38	22.43	43.88	1.9669	22.10	14.56
JointGM（Nguyen等，2021）	64.40	48.54	37.69	34.62	21.91	42.44	1.8265	21.77	15.82
MMBart（Lee等，2022）	69.89	52.99	41.32	33.10	24.79	46.60	2.1887	26.19	21.11
DialVinVL（Kottur 等人，2021a）	75.38	57.42	44.92	34.90	27.09	51.24	2.3426	29.92	22.55
GPTDeIT（Lee 和 Han，2021）	68.43	52.23	40.95	28.50	24.81	47.80	2.2271	25.04	18.06
GLIMMeR（Hemanthage 等人，2021）	74.05	56.85	44.88	35.31	27.48	50.92	2.4952	32.70	22.58
SPRING（我们的方法）	**83.29**	**64.75**	**52.41**	**42.49**	**31.90**	**57.12**	**3.1351**	**38.87**	**30.77**

表2：在 SIMMC 1.0、SIMMC 2.0 数据集、视觉子集和空间子集上的比较。我们的模型在 7 个广泛使用的指标上始终大幅优于强基线模型。特别是，通过 BLEU-4 对**视觉子集**和**空间子集**进行评估，有效验证了模型的巨大提升源于其对视觉属性和空间关系的响应更优。

SIMMC 2.0 数据集分别包含 7.2k 条时尚对话和 4k 条家具对话。时尚场景中约有 290 件数字资产，家具场景中有 110 件物品；这些物品在种子场景中重新排列，生成了 160 个不同场景。SIMMC 1.0 数据集包含 6.6k 条时尚对话和 6.4k 条家具对话。我们在 SIMMC 1.0 和 SIMMC 2.0 的开发测试集上评估模型性能，该测试集的规模与 test-std [1] 数据集相同。

此外，我们邀请人工专家从 SIMMC 1.0 和 SIMMC 2.0 开发测试集中筛选包含视觉属性或空间关系的响应，以构建**视觉子集** 和**空间子集**。我们进一步在这两个子集上评估模型，以证明我们模型的有效性。**评估指标**。 SIMMC 2.0 响应生成任务采用的官方指标是 BLEU-4，它只关注预测响应与目标响应之间的 n 元组重叠。为了进行更全面的比较，我们还采用了广泛使用的机器生成指标：BLUE-n（Papineni 等人，2002）、METEOR（Banerjee 和 Lavie，2005）、ROUGE（Lin 和 Hovy，2003）和 CIDEr（Vedantam、Zitnick 和 Parikh，2015）。与基于准确率的 BLEU 指标相比，METOR 和 ROUGH 关注召回率，并计算目标响应中的多少 n 元组出现在预测响应中，而 CIDEr 使用词频-逆文档频率为不常见的短语赋予更高权重。

实现细节。 我们的模型基于 Transformer（Vaswani 等人，2017）架构，共有 12 层，每个 Transformer 块包含 768 个隐藏单元和 12 个注意力头。每个图像块都会映射为与隐藏单元维度相同的特征。我们使用预训练视觉语言模型 OFA（Wang 等人，2022b）初始化 SPRING 的参数。即，OFA（Wang 等人，2022b）。在预训练期间，我们的模型在 8 块 Tesla V100 GPU 上以批量大小 8 训练 4 个 epoch。我们采用 Adam（Kingma 和 Ba，2015）作为优化器，学习率设为 4e-4。此外，为防止过拟合，dropout 率设为 0.2。在微调阶段，我们使用 SIMMC 训练集训练 60 个 epoch，学习率为 4e-5，批量大小为 16。

[1]未公开作为 DSTC竞赛的测试集。

对比方法。 我们将 SPRING 与 SIMMC 1.0 和 SIMMC 2.0 中强大的基线方法进行比较。在 SIMMC 1.0 上，MN-MAG（Kim 等人，2021）采用记忆网络作为编码器，并设计了多模态融合门来融合信息。Tom（Jeong 等人，2021）集成了多个 GPT-2 模型的预测结果。JBi-encoder（Huang 等人，2021a）通过联合训练来预测信念状态和响应。在 SIMMC 2.0 上，MTN（Le 等人，2019）分别对多模态输入进行编码，同时由查询感知注意力编码器引导视觉编码器。JMGPT（Kottur 等人，2021b）训练了一个多任务 GPT2-large，将对话历史和扁平化的多模态上下文作为输入。此外，JMGPT-BS（Kottur 等人，2021a）通过采用不同大小的束搜索扩展了 JMGPT。MMBart（Lee 等人，2022）将边界框坐标嵌入添加到文本输入中，并提出辅助任务来预测资产属性。DialVinVL（Kottur 等人，2021a）基于 VinVL-Base（Zhang 等人，2021），将原始边界框坐标与区域特征拼接作为视觉输入，并将对话历史与对话策略结合起来作为文本输入。GPTDeIT（Lee 和 Han，2021）使用 GPT2-large（Radford 等人，2019）作为文本模型，对对话历史和扁平化的槽值进行编码；并使用 DeIT-I（Touvron 等人，2021）作为图像模型，对当前轮次话语提及的物品进行编码。JointGM（Nguyen 等人，2021）利用 BART-large（Lewis 等人，2020），根据输入的对话历史联合预测消歧标签、信念状态和响应。与 GPTDeIT 类似，GLIMMeR（Hemanthage 等人，2021）也使用 GPT2-large，并利用资产场景 ID 帮助模型理解每个资产的语义。值得注意的是，GLIMMeR 是 SIMMC 2.0 上最先进的方法，并在 DSTC10 中获胜。

4.2 总体性能

表2展示了模型在 SIMMC 1.0 和 SIMMC 2.0 数据集响应生成任务上的结果。可以看出，SPRING 大幅超越了先前模型，并在所有代表性的机器生成指标上取得了最先进的结果。在 SIMMC 2.0 上，

Task	模型	SIMMC 2.0	视觉	空间
	VLM	38.22	34.67	25.04
视觉问答	VLM			
	带 PVQA	40.75	36.54	27.58
	使用 RVQA	41.27	37.02	27.22
	使用 PoVQA	40.89	35.94	28.08
	使用 (PVQA + RVQA + PoVQA)	41.36	37.59	28.24
空间问答	VLM			
	使用 PSQA	41.18	36.05	28.30
	使用 RSQA	40.77	35.42	28.18
	使用 VSQA	40.40	36.34	27.97
	使用（PSQA +RSQA + VSQA）	41.56	36.25	28.49
All	VLM			
	使用全部问答	41.92	38.52	30.18
	使用（全部问答 + 课程学习）	**42.49**	**38.87**	**30.77**

表3：在 SIMMC 2.0 数据集上使用 BLEU-4 指标进行的消融研究。红色和绿色阴影分别表示视觉子集和空间子集上的优势更明显。

在 BLEU-n 上，SPRING 分别比此前最佳模型高出 7.91、7.33、7.49 和 7.18，n 的取值范围为 1 至 4。BLEU-n 上显著提升的百分比表明，我们的模型成功利用了更准确的词语和短语来生成响应。我们的模型在基于召回率的指标 METEOR 和 ROUGE 上也表现出色，分数分别提升了 4.42 和 6.2。当 CIDEr 指标更加关注低频 n 元组时，SPRING 在 CIDEr 上仍以 0.64 的优势超越 GLIMMeR。此外，根据表 2 右侧的结果，我们的模型在视觉子集和空间子集上的 BLEU-4 得分最高，这验证了模型性能的提升源于其更好地理解视觉属性和空间关系，并能够利用对齐的信息进行推理，从而生成更准确的响应。

4.3 详细分析

消融研究。 如表3所示，我们进行了消融实验，以评估 SPRING 中各项预训练任务和课程学习策略的有效性。可以看出，与基础视觉语言模型相比，每项多模态问答预训练任务都能在完整的 SIMMC 2.0 数据集上显著提升 BLEU-4。具体而言，在视觉子集上，使用全部视觉问答任务进行预训练的视觉语言模型比基线模型高出 2.92；而在空间子集上，使用全部空间问答任务进行预训练的视觉语言模型比基线模型高出 3.45。这验证了视觉问答和空间问答分别增强了模型描述视觉属性和空间关系的能力。此外，最后两行进一步证明，我们设计的课程学习预训练策略能够有效激发问答预训练任务的潜力，并提升模型性能。

人工评估。 人工评估主要关注4个方面：

流畅性、**相关性**、**正确性**和**信息量**，这些都是面向任务的对话系统的重要指标。我们从 SIMMC 2.0 开发测试数据集中随机选取 500 段对话作为候选项，然后从 DialVinVL、GLIMMeR 和我们的模型生成的结果中筛选出这些对话。我们在

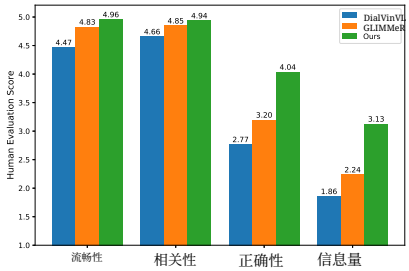

图4：SIMMC 2.0 上包含四个方面的人工评估结果。我们的模型在**正确性**和**信息量**方面均取得了显著提升。

亚马逊土耳其机器人（AMT）上发布了评估任务，并让 10 位不同的评估人员对每段选定对话的最后一条响应进行评估。每位评估人员在盲审中依据标准回复，从四个方面对 1500 条生成的回复进行 1 到 5 分的评分，以模拟真实的多模态对话场景。如图 4 所示，我们的模型在所有指标上始终优于另外两个模型，这与自动评估结果一致。

案例研究。 为了更好地展示我们模型的优势，并说明 SPRING 如何提升模型预测与背景物品相关的视觉属性和空间关系的能力，我们将模型和现有最先进模型生成的若干回复，以及相应的用户话语和场景快照进行可视化展示，如图5所示。可以明确观察到：（1）我们的模型能够利用背景物品来描述目标物体的位置。（2）我们的模型能够准确预测目标物体与背景物品之间的相对空间关系。（3）当回复中存在多个物品时，我们的模型具备将视觉属性与空间信息对齐的能力。

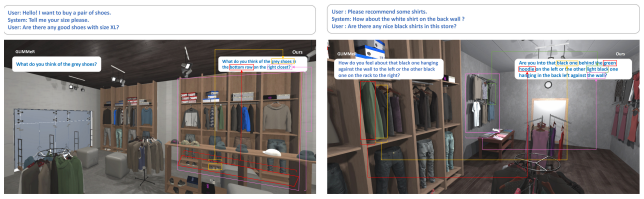

图5：SIMMC 2.0 数据集上的案例研究。

5 结论

本文提出了一种名为 SPRING 的新型情境对话代理预训练方法。具体而言，预训练中使用的所有问答对及其难度标签均由我们的增量布局图生成，无需任何额外的人工标注。在 SIMMC 1.0 和 SIMMC 2.0 上的实验结果表明，SPRING 的表现远超先前模型，并且能够更准确地描述视觉属性和空间关系。

致谢

我们衷心感谢匿名审稿人提出的建议和意见。本研究部分由国家自然科学基金委员会（NSFC62076032）资助。我们还要感谢 GuanqiZhan 提供的宝贵建议。

参考文献

Banerjee, S.; and Lavie, A. 2005. METEOR：一种用于机器翻译评估的自动指标，与人类判断的相关性更高。载于 Goldstein, J.; Lavie, A.; Lin, C.; and Voss, C. R. 编，*ACL* 机器翻译和／或摘要生成内在与外在评估指标研讨会论文集。

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. 课程学习。载于国际机器学习会议（*ICML*）论文集。

Chang, X.; Ren, P.; Xu, P.; Li, Z.; Chen, X.; and Hauptmann, A. 2020. 场景图综述：生成与应用。载于 *IEEE* 神经网络学习汇刊论文集。Chattopadhyay, P.; Yadav, D.; Prabhu, V.; Chandrasekaran, A.; Das, A.; Lee, S.; Batra, D.; and Parikh, D. 2017. 通过合作式人类—AI 游戏评估视觉对话智能体。载于第五届 *AAAI* 人类计算与众包会议论文集（*AAAI*）。Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. UNITER：通用图文表征学习。载于欧洲计算机视觉会议论文集（*ECCV*），104–120。施普林格。Cho, J.; Lei, J.; Tan, H.; and Bansal, M. 2021. 通过文本生成统一视觉与语言任务。载于国际机器学习会议（*ICML*）论文集。Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M.; Parikh, D.; and Batra, D. 2017. 视觉对话。载于*IEEE*计算机视觉与模式识别会议论文集（*CVPR*）。

De Vries, H.; Strub, F.; Chandar, S.; Pietquin, O.; Larochelle, H.; and Courville, A. 2017年。Guesswhat?! 通过多模态对话发现视觉对象。载于*IEEE* 计算机视觉与模式识别会议论文集（*CVPR*）。

He,W.; Dai,Y.; Hui, B.; Yang, M.; Cao, Z.; Dong, J.; Huang, F.; Si, L.; and Li, Y. 2022年。SPACE-2：面向任务的对话理解的树结构半监督对比式预训练。载于第*29*届国际计算语言学会议论文集（*COLING*）。

He, W.; Dai, Y.; Yang, M.; Sun, J.; Huang, F.; Si, L.; and Li, Y. 2022年。SPACE-3：面向任务的对话理解与生成的统一对话模型预训练。载于第*45*届国际*ACM*信息检索研究与发展会议论文集（*SIGIR*）。

He, W.; Dai,Y.; Zheng, Y.; Wu, Y.; Cao, Z.; Liu, D.; Jiang, P.; Yang, M.; Huang, F.; Si, L.; 等. 2022c. SPACE：一种采用半监督学习和显式策略注入的任务导向对话生成式预训练模型。*AAAI* 人工智能会议论文集（*AAAI*）。Hemanthage, B.; Lemon, O.; Dondrup, C.; and Bartie, P. 2021. GLIMMeR：利用生成式预训练变换器进行共指消解的全局-局部信息感知多模态指代。发表于 *AAAI* 上的 *DSTC10* 挑战赛研讨会。Huang, X.; Tan, C. S.; Ng, Y. B.; Shi, W.; Yeo, K. H.; Jiang, R.; and Kim, J. J. 2021a. 面向情境交互式多模态对话的联合生成与双编码器。发表于*AAAI* 上的 *DSTC9* 挑战赛研讨会。

Huang, Z.; Zeng, Z.; Huang, Y.; Liu, B.; Fu, D.; and Fu, J. 2021年b. 视野之外：面向视觉-语言表征学习的端到端预训练。见*IEEE/CVF* 计算机视觉与模式识别会议（*CVPR*）论文集。

Hui, B.; Geng, R.; Ren, Q.; Li, B.; Li, Y.; Sun, J.; Huang, F.; Si, L.; Zhu, P.; and Zhu, X. 2021年。用于跨领域上下文依赖语义解析的动态混合关系探索网络。见*AAAI*。

Hui, B.; Geng, R.; Wang, L.; Qin, B.; Li, Y.; Li, B.; Sun, J.; and Li, Y. 2022年。S^2SQL：将语法注入面向文本到 SQL 解析器的问题-模式交互图编码器。见计算语言学协会论文集：*ACL 2022* 研究成果，1254–1262。都柏林，爱尔兰：计算语言学协会。

Jeong,Y.; Lee, S. J.; Ko,Y.; and Seo, J.2021年。Tom：基于 GPT-2 的端到端面向任务的多模态对话系统。收录于*AAAI* 的 *DSTC9* 挑战赛研讨会。

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021年。通过噪声文本监督扩展视觉和视觉-语言表征学习。收录于国际机器学习会议论文集（*ICML*）。

Kim, B.; Lee, I.; Jeong, Y.; Youngjoong, K.; Koo, M.-W.; and Seo, J. 2021年。通过添加对话状态和各种多模态门控改进多模态 API 预测。收录于*AAAI* 的 *DSTC9* 挑战赛研讨会。

Kingma, D. P.; and Ba, J. 2015。Adam：一种随机优化方法。收录于国际学习表征会议论文集（*ICLR*）。

Kottur, S.; Moon, S.; Geramifard, A.; 和 Damavandi, B. 2021a. 情境化交互式多模态对话（SIMMC）2.0 赛道概述，发表于 DSTC10。在*AAAI* 上的 *DSTC10* 挑战赛研讨会。

Kottur, S.; Moon, S.; Geramifard, A.; 和 Damavandi, B. 2021b. SIMMC 2.0：面向沉浸式多模态对话的任务型对话数据集。发表于*2021*年自然语言处理实证方法会议（*EMNLP*）论文集。

Le, H.; Sahoo, D.; Chen, N.; and Hoi, S. 2019年。用于端到端视频接地对话系统的多模态 Transformer 网络。载于计算语言学协会第*57*届年会论文集（*ACL*）。

Lee, H.; Kwon, O. J.; Choi, Y.; Kim, J.; Lee, Y.; Han, R.; Kim, Y.; Park, M.; Lee, K.; Shin, H.; and Kim, K.-E. 2022年。为情境式对话智能体学习多模态上下文嵌入。载于计算语言学协会（*NAACL*）论文集。Lee, J.; and Han, K. 2021年。使用预训练单模态模型进行多模态交互：SIMMC 2.0。载于*AAAI DSTC10* 挑战赛研讨会。Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020年。BART：面向自然语言生成、翻译与理解的去噪序列到序列预训练。载于计算语言学协会第*58*届年会论文集（*ACL*）。Li, W.; Gao, C.; Niu, G.; Xiao, X.; Liu, H.; Liu, J.; Wu, H.; and Wang, H. 2021年。UNIMO：通过跨模态对比学习实现统一模态理解与生成。载于计算语言学协会第*59*届年会论文集（*ACL*）。Liao, L.; Long, L. H.; Ma, Y.; Lei, W.; and Chua, T.-S. 2021年。通过增量推理进行对话状态跟踪。计算语言学协会汇刊，9：557–569。Lin, C.; and Hovy, E. H. 2003年。使用 N 元语法共现统计自动评估摘要。载于北美计算语言学协会人类语言技术会议（*HLT-NAACL*）。Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019年。ViLBERT：为视觉与语言任务预训练任务无关的视觉语言表示。载于神经信息处理系统进展论文集（*NIPS*）。Moon, S.; Kottur, S.; Crook, P. A.; De, A.; Poddar, S.; Levin, T.; Whitney, D.; Difranco, D.; Beirami, A.; Cho, E.; Subba, R.; and Geramifard, A. 2020年。情境式交互多模态对话。载于第*28*届国际计算语言学大会（*COLING*）论文集。Nguyen, T.-T.; Shi, W.; Jiang, R.; and jae Kim, J. 2021年。面向 SIMMC 2.0 的多模态联合学习生成模型。载于*AAAI DSTC10* 挑战赛研讨会。Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002年。BLEU：一种机器翻译自动评估方法。载于计算语言学协会第*40*届年会论文集（*ACL*）。Qiu, L.; Liang, Y.; Zhao, Y.; Lu, P.; Peng, B.; Yu, Z.; Wu, Y. N.; and Zhu, S. 2021年。SocAoG：用于对话中社会关系推断的增量图解析。载于计算语言学协会第*59*届年会暨第*11*届自然语言处理国际联合会议论文集（*ACL-IJCNLP*。

Qiu, L.; Zhao, Y.; Liang, Y.; Lu, P.; Shi, W.; Yu, Z.; and Zhu, S.-C. 2022年。《朝向具有心理状态转变和人类价值的社会智能体》。发表于*SIGDIAL*。

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; 等。2021年。从自然界中学习可迁移视觉模型语言监督。在国际机器学习会议（*ICML*）论文集。Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; 等。2019年。语言模型是无监督的多任务学习器。在*Arxiv*。Ren, S.; He, K.; Girshick, R.; 和 Sun, J. 2015。Faster R-CNN：借助区域建议网络实现实时目标检测。在神经信息处理系统进展会议（*NIPS*）论文集。Saha, A.; Khapra, M.; 和 Sankaranarayanan, K. 2018年。朝着构建大规模、多模态、领域感知的对话系统迈进。在*AAAI* 人工智能会议论文集。Tan, H.; 和 Bansal, M. 2019。LXMERT：从 Transformer 中学习跨模态编码器表示。在*2019*年自然语言处理实证方法会议暨第九届国际自然语言处理联合会议（*EMNLP-IJCNLP*）论文集。Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; 和 J´egou, H. 2021年。通过注意力训练数据高效的图像 Transformer 并进行蒸馏。在国际机器学习会议（*ICML*）论文集。Vaswani, A.; Shazeer, N.; Par mar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; 和 Polosukhin, I. 2017年。注意力就是你所需要的一切。神经信息处理系统进展会议（*NIPS*）论文集。Vedantam, R.; Zitnick, C. L.; 和 Parikh, D. 2015。CIDEr：基于共识的图像描述评估。在*IEEE*计算机视觉与模式识别会议（*CVPR*）。Wang, L.; Qin, B.; Hui, B.; Li, B.; Yang, M.; Wang, B.; Li, B.; Huang, F.; Si, L.; 和 Li, Y. 2022年。Proton：从预训练语言模型中探测模式链接信息以进行文本到SQL解析。第*28*届 *ACM SIGKDD*知识发现与数据挖掘会议论文集。Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; 和 Yang, H. 2022年。OFA：通过简单的序列到序列学习框架统一架构、任务和模态。在国际机器学习会议（*ICML*）论文集。Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; 和 Gao, J. 2021年。VinVL：重新审视视觉语言模型中的视觉表示。在*IEEE/CVF* 计算机视觉与模式识别会议（*CVPR*）论文集。

槽位类型		槽位值
时尚	颜色	’红色, 白色, 黄色’, ’红色, 白色’, ’紫色’, ’白色, 黑色’, ’黑色’, ’深灰色’, ’灰色, 黑色’, ’棕色’, ’深绿色’, ’灰色’, ’深蓝色’, ’蓝色, 白色’, ’灰色, 棕色’, ’白色, 蓝色’, ’黄色, 白色’, ’深绿色, 深蓝色’, ’浅灰色’, ’白色’, ’蓝色’, ’绿色’, ’栗色’, ’黄色’, ’红色’, ’紫罗兰色’, ’黑色, 红色, 白色’, ’黄色, 黑色’, ’蓝色, 黑色’, ’黑色, 白色’, ’浅蓝色’, ’红色, 黑色’, ’粉色, 白色’, ’橙色’, ’黄色, 棕色’, ’浅粉色’, ’深棕色’, ’粉色’, ’深黄色’, ’浅红色’, ’绿色, 白色’, ’灰色, 白色’, ’黑色, 红色’, ’灰色, 蓝色’, ’棕色, 白色’, ’白色, 黑色, 红色’, ’米色’, ’浅橙色’, ’橙色, 紫色’, ’脏绿色’, ’蓝色, 灰色’, ’黑色, 灰色’, ’白色, 灰色’, ’橄榄色’, ’深红色’, ’橄榄色, 黑色’, ’粉色, 黑色’, ’蓝色, 绿色’, ’绿色, 黑色’, ’浅蓝色, 浅绿色’, ’深粉色, 白色’, ’脏灰色’, ’深粉色’, ’红色, 灰色’, ’深紫罗兰色’, ’橄榄色, 白色’, ’黑色, 橙色’, ’金色’, ’栗色, 白色, 蓝色’, ’绿色, 紫罗兰色, 粉色’, ’白色, 红色, 紫罗兰色’, ’棕色, 黑色’, ’黑色, 橄榄色’
时尚	type	’女式衬衫’, ’夹克’, ’衬衫’, ’毛衣’, ’连衣裙’, ’T恤’, ’束脚裤’, ’牛仔裤’, ’帽子’, ’背心’, ’背心’, ’外套’, ’鞋子’, ’裙子’, ’西装’, ’裤子’, ’连帽衫’
家具	颜色	’红色’, ’蓝色’, ’白色’, ’灰色’, ’棕色’, ’绿色’, ’黑色’, ’黑白色’, ’木质’
家具	type	’地毯’, ’床’, ’椅子’, ’茶几’, ’沙发椅’, ’边几’, ’灯’, ’架子’, ’沙发’, ’桌子’
背景物品		’置物架’, ’墙’, ’镜子’, ’架子’, ’壁橱’, ’桌子’, ’衣柜’, ’橱柜’, ’窗户’, ’隔断’, ’门’, ’柜台’, ’小隔间’, ’小隔间’, ’衣架’, ’支架’, ’橱柜’, ’人体模型’, ’鞋盒’, ’房间隔断’, ’墙面隔断’
位置	介词	’在.....里面’, ’在.....上面’, ’在’, ’在.....后面’, ’朝向’, ’到’, ’靠着’, ’.....的’, ’沿着’, ’在.....下面’, ’朝向’, ’在.....上方’
冠词	和代词	“该”、“一个”、“那个”、“这个”、“其他”、“另一个”
标点符号	and 连词	’,’, ’.’, ’;’, ’?’, ’和’, ’或’

表4：正则表达式中的槽位类型和槽值。

附录

在表4中，我们展示了用于从对话语料库中提取视觉属性、空间描述、背景物品和空间关系的四个正则表达式的槽位类型和槽位值。所有颜色值和类型值均来自 SIMMC 2.0 元数据，而背景物品则是维基百科提供的常见家具词汇，但不包括 SIMMC 2.0 家具类型中出现的词汇。冠词、代词、并列连词和标点符号均取自《牛津英语词典》。为使说明更加清晰，下面我们采用一个简单的 Python 代码片段，展示如何使用我们设计的正则表达式提取视觉和空间信息。

```
1 导入 re 2 3
系统 响应 = ’架子上的蓝色T恤或桌子上方的红色夹克怎么样？’
5
个) (
色 |
红色
黑色
橙色
色 |
色,
绿色
色 |2
色 |
24 T
西装
在...
在...
和 |
30 \. |\?)’ 31 32 正则表达式_bi =’(置物架 |墙 |镜子 |架子 |
33
小隔
面 |
下面
re.f
蓝色
sd
(’在
桌子
51 i
方 |
sd i
bi_
sr, s
(’ 桌
方’)
```