SPRING: Situated Conversation Agent Pretrained with Multimodal Questions from Incremental Layout Graph

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] Beijing University of Posts and Telecommunications, Beijing, China
[2] Independent Researcher
{longyuxing, fulong_ye, hanzhuoxin, yuancx, xjwang}@bupt.edu.cn, lyb821@gmail.com

Abstract

Existing multimodal conversation agents have shown impressive abilities to locate absolute positions or retrieve attributes in simple scenarios, but they fail to perform well when complex relative positions and information alignments are involved, which poses a bottleneck in response quality. In this paper, we propose a **S**ituated Conversation Agent **PR**etrained with Multimodal Questions from **IN**cremental Layout **G**raph (**SPRING**) with abilities of reasoning multi-hops spatial relations and connecting them with visual attributes in crowded situated scenarios. Specifically, we design two types of Multimodal Question Answering (MQA) tasks to pretrain the agent. All QA pairs utilized during pretraining are generated from novel Incremental Layout Graphs (ILG). QA pair difficulty labels automatically annotated by ILG are used to promote MQA-based Curriculum Learning. Experimental results verify the SPRING's effectiveness, showing that it significantly outperforms state-of-the-art approaches on both SIMMC 1.0 and SIMMC 2.0 datasets. We release our code and data at Github LYX0501/SPRING repository.

1 Introduction

Building multi-modal conversation agents that can communicate with people in visual situations is an attractive goal for the AI community. Lots of specific tasks and datasets for visual dialog, like VisDial (Das et al. 2017), GuessWhat (De Vries et al. 2017), GuessWhich (Chattopadhyay et al. 2017), are proposed in recent years. Among them, the Situated Interactive Multi-modal Conversation (SIMMC 1.0) (Moon et al. 2020) aims to study task-oriented dialogues that encompass a situated multi-modal user context in the form of a virtual reality (VR) environment. The updated dataset SIMMC 2.0 (Kottur et al. 2021b) provides a more challenging test bed for multi-modal conversation agents. There are many assets with a complex layout in each image. Figure 1 gives an example of a scene and a fragment of dialogue in SIMMC 2.0. There are dozens of clothes in the image. Each cloth is a digit asset with a unique asset ID and a set of attributes (*e.g.* type, color) in the metadata. But there is no information on the scene layout except for a few labels on four relations (up, down, left, and right) between the assets.

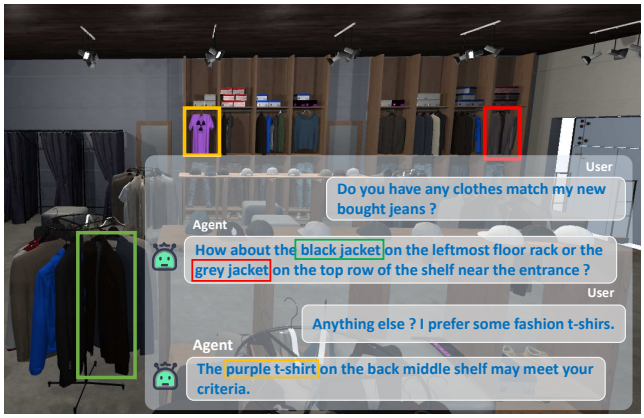

Figure 1: An example of a virtual scene and a fragment of dialogue in SIMMC 2.0. Since there are many clothes with similar visual attributes, it is difficult to talk about a asset only by its visual attributes.

A number of works have been established on SIMMC 2.0. Based on different multi-modal Visual-Language pretraining Models (VLM), previous researchers pay more attention to learning the visual attributes of assets. QS Goal Diggers (Kottur et al. 2021a) and Kakao Enterprise (Lee and Han 2021) directly insert visual attributes into models input, while Sogang University (Kottur et al. 2021b) and A-STAR (Nguyen et al. 2021) build a set of visual attributes prediction tasks in pre-training stage. KAIST (Lee et al. 2022) designs an auxiliary task to predict visual attributes. However, less attention has been paid to building spatial relations between assets. All existing models only use the coordinates of asset bounding boxes as positional information, which cannot capture spatial relations in the scenes.

As a result, the models can accurately generate *"the black jacket"* but fail to describe more natural referring expressions like *"the black jacket on the leftmost floor rack"*. It is obvious that the later expression is more useful in a scene including lots of clothes with similar attributes. The combination of attributes and spatial relations helps people locate assets quickly. To generate this type of expression, a model needs to learn not only the visual attributes of each asset but also spatial relations between different items.

To address above problem, we propose **S**ituated Conver-

*Corresponding author

sation Agent **PR**etrained with Multimodal Questions from **IN**cremental Layout **G**raph (**SPRING**), which is pretrained with multimodal questions generated from incremental layout graph. In our method, we design **Incremental Layout Graph (ILG)** for each scene to capture rich spatial relations between different scene items. Unlike scene graph (Chang et al. 2020), an ILG is built using pure textual information and can be extended incrementally with newly added dialogue. And then, two types of **Multimodal Question Answering (MQA)** pre-training tasks and corresponding QA pairs are collected by traversing nodes (digital assets and background items) on the ILG. According to the spanned path length, QA pairs can be automatically annotated with difficulty levels. Curriculum Learning (Bengio et al. 2009) is therefore employed for pre-training on a Transformer (Vaswani et al. 2017) encoder-decoder backbone. Experiments on both SIMMC 1.0 and SIMMC 2.0 show that our method improves the response quality by a large margin compared to previous best models.

The main contributions of our work are as follows:

- We first propose a novel approach to build ILGs for virtual scenes from dialogue text incrementally. The ILGs include scene items with relations. It is worth noting that this process does not rely on any human annotation.
- Based on ILGs, we introduce two types of new MQA pretraining tasks that can facilitate model understanding of visual metadata and spatial relations between different assets. Pre-training samples are automatically generated by traversing the ILG, which also generates an accompanying difficulty label for curriculum learning.
- We conduct thorough experiments to verify that our approach effectively enhances response quality. Our approach outperforms existing state-of-the-art methods by a significant margin consistently on all metrics on both SIMMC 1.0 and SIMMC 2.0.

2 Related Works

Situated Interactive Multimodal Conversations. Conversation systems have developed rapidly in recent years, *e.g.*, task-oriented conversations pretraining (He et al. 2022c,a,b), knowledge-based conversations (Hui et al. 2022; Wang et al. 2022a) and so on. Among them, multimodal conversations are the new trend. META releases multimodal conversation datasets SIMMC (Kottur et al. 2021b) based on VR shopping stores. There are hundreds of scene snapshots from different angles. Compared with the previous multimodal dialogue datasets MMD (Saha, Khapra, and Sankaranarayanan 2018) and VisDial (Das et al. 2017), the situated agent is required to generate more complex visual attributes and more detailed spatial relations to infer digital assets in the scene. Kottur et al. (2021a) has preliminary explorations on utilizing visual attributes and spatial relations. Concretely, DialVinVL (Kottur et al. 2021a) incorporates slot values about visual attributes with dialogue history as textual input and concatenates original box coordinates to region features extracted by the object detector as visual input. JMGPT (Kottur et al. 2021b) and JointGM (Nguyen et al. 2021) apply language model to predict visual attributes and system response jointly. MMBart (Lee et al. 2022) adds embedded box coordinates to textual embedding as Transformer input and designs auxiliary tasks to predict visual attributes according to the output of encoder hidden states. We can find that their utilized spatial information is all from the bounding box. Unlike these methods, we first notice the lack of VLM's capability for visuality and spatiality, and then propose MQA pretraining tasks based on incremental layout graphs which have been successfully applied to (Qiu et al. 2021; Liao et al. 2021; Hui et al. 2021; Qiu et al. 2022).

Visual Language Pretraining. To improve models' perception of text and image and help them establish connections between multimodal information, kinds of visual language pretraining models are designed. ViLBERT (Lu et al. 2019) and UNITER (Chen et al. 2020) propose to consider the raw output of the detector, a distribution of object classes, as soft labels and optimize the KL-divergence between two distributions. LEXMERT (Tan and Bansal 2019) and UNIMO (Li et al. 2021) propose Masked Region Feature Regression (MRFR) regresses the masked region feature to its corresponding original region feature, where represents images as a sequence of region features by Faster R-CNN (Ren et al. 2015). Furthermore, SOHO (Huang et al. 2021b) is designed to avoid information leakage from neighbor features when images are converted into grid features or patch features.

Recently, CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021) leverage large-scale image-text pairs to learn transferable visual representations and exhibit surprising zero-shot transfer to image classification tasks. VL-T5 (Cho et al. 2021) and OFA (Wang et al. 2022b) introduce downstream tasks, like visual grounding and grounded caption, into pretraining tasks to narrow the gap between pretraining and fine-tuning. Unlike these efforts, we design new pretraining tasks through a unified QA paradigm to improve existing methods' visual attributes and spatial relations modeling without adding new modules.

3 Methods

Let $\mathcal{D} = \{(U_t, R_t, I_t)\}_{t=1}^{T}$ be an ongoing dialogue between a user and an agent with T rounds, where U_t is the user utterance at time step t, R_t is the language response to U_t by the agent, and I_t is the accompanying scene image. The task here is to predict the optimal language response $R_t^{'}$, given the dialog history $H_t = [U_i, R_i]_{i=1}^{t-1}$, current user utterance U_t and scene image I_t, as modeled in Eq (1)

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

where θ is the model learnable parameters.

To solve above problem, we propose a mutimodal dialogue model **SPRING**, which is pretrained with mutlimodal questions generated from incremental layout graph. In the following sections, we will introduce model architecture, ILG generation and MQA pretraining tasks in order.

3.1 Architecture

The backbone of **SPRING** model is encoder-decoder based single-stream VLM framework, which are stacks of Trans-

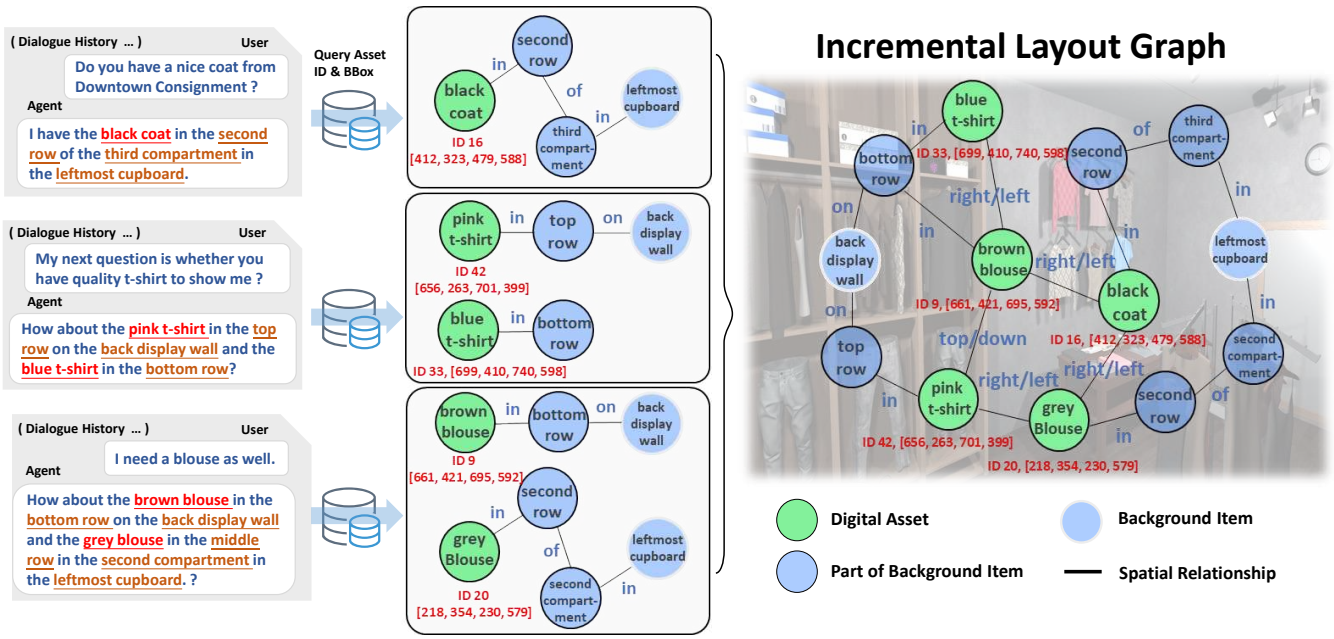

Figure 2: Construction of Incremental Layout Graph from dialogue. Digital assets and background items constitute ILG nodes while spatial relations form ILG edges. ILG is continuously incremented with newly added dialogue under the same scene.

former (Vaswani et al. 2017) layers. The scene image $I_t \in \mathbb{R}^{h\times w\times c}$ is splitted to P patches. And each patch is projected to visual embedding of the model hidden size. The dialogue history and current user utterance are converted to sub-word sequence by Byte-Pair Encoding (BPE) and then embedded to textual embedding. All visual embedding and textual embedding are concatenated as model input.

To facilitate **SPRING** to better understand the information of embodied scenes, we propose a series of MQA pre-training tasks based on layout graphs $\mathcal{G}_i$. As there is no annotated layout graph in the SIMMC dataset, we propose an unsupervised ILG construction method based on natural language dialog history.

3.2 Incremental Layout Graph (ILG)

We observe that visual attributes and spatial descriptions exist in the dialogue history. Compared with dataset annotations, the information from dialogue is more detailed. For example, SIMMC 2.0 annotation only gives bounding boxes of digital assets and four types of relative position (up, down, right, left) between them, while dialogues include the absolute position of background items and their relative position with assets. A crucial discovery lies in the co-coreference between dialogue history and response, the same assets present in the responses to be generated.

Therefore, we propose an ILG generation algorithm to extract high quality information from dialogues and generate Incremental Layout Graph (ILG) $\mathcal{G}_i = \langle \mathcal{V}_i, \mathcal{E}_i \rangle$ to dispose them, where $\mathcal{V}_i$ denotes the node set containing the digital assets and background items from dialog history and $\mathcal{E}_i$ represents the edge set depicting spatial relations between scene items $\mathcal{V}_i$.

Textual Information Extraction and Alignment we consider adopting a rule-based textual information extraction method, *i.e.* , regular expression, to extract visual attributes and spatial descriptions from dialogue history without human annotation. The regular expressions RegExp_{va} and RegExp_{sd} for **visual attribute** and **spatial description** are as follows.

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

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

where art. is article, prep. represents preposition and punc. means punctuation. Please refer to Appendix for details. With these two regular expressions, as left part of Figure 2 shows, we can extract visual attribute *"black coat"* and spatial description *"in the second row of the third compartment in the leftmost cupboard"* from dialogue history.

Although the visual attributes and spatial descriptions extracted by the above regular expressions are naturally aligned because of language features, they are not aligned with asset IDs, making asset box coordinates unusable. To solve this problem, we query the color and type of assets from the database by their IDs to compose visual attributes like *"black coat"* and then try pairing them with the extracted visual attributes like *"black coat"*. If these two visual attributes match, the corresponding asset ID "16" can be determined, from which we can get the paired asset IDs, visual attributes, and spatial descriptions. We further design the following two regular expressions to extract **background item** and **relative spatial relation** from extracted spatial descriptions.

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

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

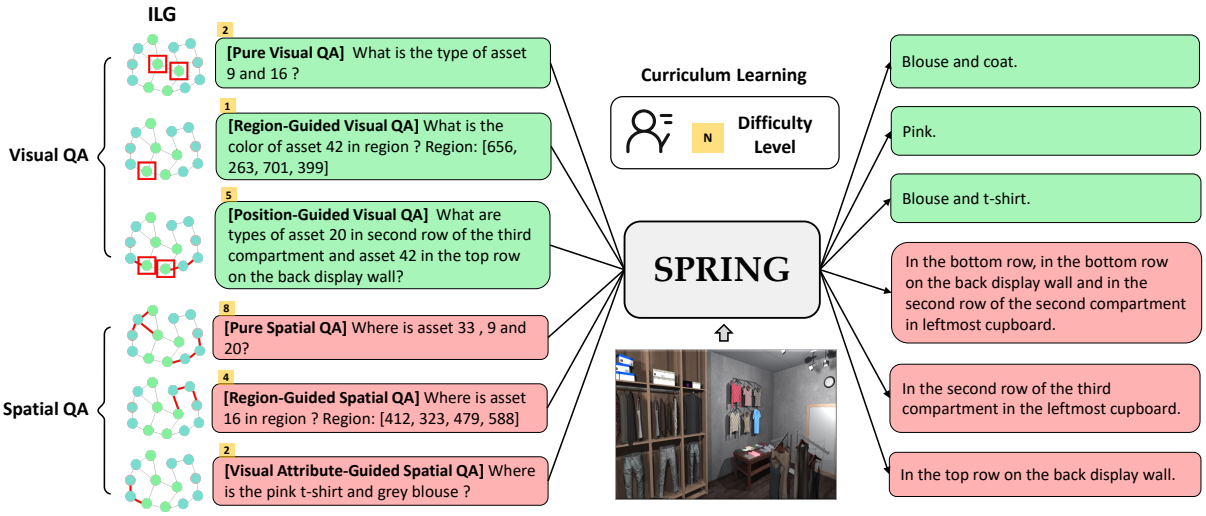

Figure 3: Demonstration of SPRING model and two types of MQA pretraining tasks, Visual QA and Spatial QA.

QA TYPE	QUESTION TEMPLATE	ANSWER
PVQA	What is the **[visual attribute type]** of item **[asset ID]**?	**[visual attribute value]**
RVQA	What is the **[visual attribute type]** of item **[asset ID]** in region? Region: **[x1, y1, x2, y2]**	**[visual attribute value]**
POVQA	What is the **[visual attribute type]** of item **[asset ID]** **[position]**?	**[visual attribute value]**
PSQA	Where is the item **[asset ID]**?	**[position]**
RSQA	Where is the item **[asset ID]** in region? Region: **[x1, y1, x2, y2]**	**[position]**
VSQA	Where is the **[item color]** **[item type]** **[asset ID]**?	**[position]**

Table 1: QA pair template. Square brackets '$[*]$' represent slots to be filled by traversing ILGs.

where RegExp_{bi} and RegExp_{sr} denote regular expressions for background item and spatial relation, prep. represents preposition. With these two regular expressions, as middle part of Figure 2 shows, we can extract background items *"second row"*, *"third compartment"* and *"leftmost cupboard"* and relative spatial relations *"in"*, *"of"* from spatial description obtained previously.

Incremental Layout Graph Generation With rich information extracted from a sample of dialogue history, layout sub-graph can be generated as middle part of Figure 2 shows. In the layout sub-graph, digital asset node store its visual attributes like *"black coat"* and asset ID "16" while background item nodes store background items like *"second row"*, *"third compartment"* and *"leftmost cupboard"*. Spatial relations and queried bounding boxes are utilized to define layout sub-graph edges. As the right part of Figure 2 shows, the scene ILG continuously increments with newly added sub-graph about the same scene, which finally can include all digital assets, background items, and spatial relations between them under this scene. Mining information on the ILG is simple but effective. The visual attributes can be easily obtained by traversing the ILG nodes, while multiple types of spatial relations can be inferred by walking along the ILG edges.

3.3 ILG-Based MQA Pre-training Tasks

To enhance response generation quality of visual attributes and spatial relations, we design visual QA pre-training task and spatial QA pre-training task based on Multimodal Question Answering (MQA), which respectively contain three types of novel sub-tasks. As shown in Figure 3 and Table 1, all QA pairs are automatically generated by traversing ILG and filling the corresponding template. The QA pair generation algorithm is displayed in Algorithm 1. Formularly, we use the question template filling function $Q_{type}(\cdot)$ to generate question, A_{type} represents corresponding answer, Type_{va} means visual attribute type, ID_{asset} denotes asset ID, I_{scene} is scene image, BBox_{asset} means asset region coordinates, t_{sr} represents spatial relation, t_{va} is visual attribute, t_{bi} denotes background item.

Visual QA

Pure Visual QA (PVQA) As the most basic visual QA task, the goal of Pure Visual QA is to help the model establish connections between asset ID and corresponding visual attributes when a scene image is provided. We design PVQA template in which the question prompts the type of visual attribute and asset ID. The pure visual question can be generated by traversing the asset nodes of ILG and filling `[asset ID]` slot in the template while answers are generated based on the visual attributes stored in asset nodes. The objective of PVQA task is the following.

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

Region-Guided Visual QA (RVQA) To improve the model's ability of locating asset and describing its visual attribute by region visual context, we design RVQA template based on PVQA, in which the question is guided by

asset region coordinates and asset ID. The region-guided visual question can be generated by traversing asset nodes of ILG and filling `[asset ID]`, bounding box coordinates `[x1, y1, x2, y2]` slots in the template. The corresponding answer is produced based on the visual attributes stored in asset nodes. The objective of RVQA task is the following.

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

Algorithm 1: **QA Pair Generation**

Input:
ILG $\mathcal{G}_i = \langle \mathcal{V}_i, \mathcal{E}_i \rangle$, QA template list T
Output:
QA pair list QA, difficulty label list DL

Initialize QA pair list QA and difficulty label list DL
for $node$ in $\mathcal{E}_i$ **do**
  if $TypeOf(node) =$ "$background\ item$" **then**
    Skip $node$
  /* Get information from digtal asset node */
  $(t_{va}, ID_{asset}, BBox_{asset}) \leftarrow GetInfo(\mathcal{G}_i, node)$
  /* Walk from node to get spatial relations */
  $(t_{bi}, t_{sr}) \leftarrow Walk(\mathcal{G}_i, node)$
  $t_{slot} \leftarrow (t_{va}, t_{bi}, t_{sr}, BBox_{asset}, ID_{asset})$
  for $template$ in T **do**
    /* Fill in the template */
    $(qa, dl) \leftarrow FillIn(template, t_{slot})$
    Add $QA \leftarrow qa, DL \leftarrow dl$

Position-Guided Visual QA (POVQA) In the conversations, instead of region coordinates, an agent has to locate asset by its spatial information no matter when understanding user utterances or making recommendations. To bring the question closer to a real conversation, we design POVQA template by replacing region coordinates in RVQA with spatial relations. For position-guided visual question template, the `[asset ID]` slot can be filled by traversing asset nodes of ILG while the `[position]` slot is filled by spatial relation path between asset nodes and background item nodes. The corresponding answer is produced based on the visual attribute stored in asset nodes. The objective of POVQA task is the following.

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

Spatial QA

Pure Spatial QA (PSQA) As the most basic spatial QA task, the goal of PSQA is to help the model establish connections between asset ID and corresponding spatial relations when a scene image is provided. We design PSQA template in which the question only prompts "where" and asset ID. The pure spatial question can be generated by traversing the asset nodes of ILG and filling `[asset ID]` slot in the template, while answers are generated based on the spatial relation paths between the background item node and the asset node. The objective of PSQA task is the following.

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

Region-Guided Spatial QA (RSQA) To improve the model's ability of locating an asset and describing its spatial relations by region visual context, we design RSQA template based on PSQA, in which the question is guided by asset region coordinates and asset ID. The region-guided visual question can be generated by traversing asset nodes of ILG and filling the slots of `[asset ID]`, bounding box coordinates `[x1, y1, x2, y2]` in the template. The corresponding answer is produced based on the spatial relation paths between the background item node and the asset node. The objective of RSQA task is the following.

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

Visual Attribute-Guided Spatial QA (VSQA) In the conversations, instead of region coordinates, an agent has to locate an asset by its visual attribute no matter when understanding user utterances or making recommendations. To bring the question closer to a real conversation, we design VSQA template by replacing region coordinates in RSQA with visual attributes (*e.g.* color, type). For spatial-guided visual question template, the `[asset ID]` slot can be filled by traversing asset nodes of ILG while the `[item color]` and `[item types]` slots are filled by the visual attribute stored in asset nodes. The corresponding answer is produced based on the spatial relation paths between the background item node and the asset node. The objective of VSQA task is the following.

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

3.4 MQA-Based Curriculum Learning

Automatic Difficulty Level Annotation When generating QA pairs by walking on the ILG, the number of nodes spanned by the pathway can be recorded. The more nodes the path passes through, the more scene information contained in the corresponding QA pair, which means that the multimodal dialogue model needs more hops to make inferences. Therefore, we automatically label the difficulty level of each QA pair according to the number of nodes the path spans. For example, when generating the question *"Where is the brown jacket 83 & 1055?"* and the answer *"it is on the floor rack near the entrance."*, one asset node *"brown jack 83 & 1055"* and two background item nodes are spanned on the ILG. The difficulty level of this QA pair is annotated as 3. The following is the formal expression.

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

where d denotes the normalized difficulty level of QA pair, $|V_{spanned}|$ represents the number of ILG nodes spanned by corresponding path, D is the maximum value of ILG nodes spanned by the QA pair path in the dataset.

Pretraining Strategy With automatically annotated difficulty labels, we propose MQA based curriculum learning to activate the potential of our designed MQA pretraining tasks. We define the model competence c as follows.

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

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

Table 2: Comparison on SIMMC 1.0, SIMMC 2.0 dataset, visual and spatial subsets. Our model consistently outperforms strong baselines by a large margin on 7 widely-used metrics.

where t is the index of current training step, T represents the maximum number of training steps, $\min^2(d)$ means the minimum value of difficulty level d, α and β are hyper-parameters, γ is determined by α as $\sqrt{\frac{1}{\alpha}}$. Here we set α to 1.2 and β to 0.8. At a given training step t, QA pair with difficulty smaller than or equal to $c(t)$ (*i.e.* $d \leq c(t)$) will be sampled for training. As such, our pretraining strategy focuses on QA pairs with lower difficulty in the early stage, aiming at helping the model form preliminary perception and inference capabilities for scene items. In the middle and late stages, more difficult QA pairs are added, which improves the model's ability to generate visual attributes and spatial relations for multiple assets.

After MQA pretraining, **SPRING** model is fine-tuned on the SIMMC response generation task. The auto-regressive language modeling objective is the following.

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

where N denotes the total number of training samples.

4 Experiment

4.1 Setup

Datasets. To evaluate the performance of the proposed model, we first conduct experiments on widely-used situated multimodal dialogue datasets SIMMC 1.0 and SIMMC 2.0. The SIMMC 2.0 dataset contains 7.2k fashion dialogs and 4k furniture dialogs, respectively. There are around 290 digital assets for fashion and 110 assets for furniture, which are rearranged within seed scenes to generate 160 different scenes. The SIMMC 1.0 dataset includes 6.6k fashion dialogs and 6.4k furniture dialogs. We evaluate model performance on the dev-test split of SIMMC 1.0 and SIMMC 2.0, which has the same scale as the test-std [1] dataset.

In addition, we invite human experts to filter responses with visual attribute or spatial relation from SIMMC 1.0 and SIMMC 2.0 dev-test split to construct **Visual Subset** and **Spatial Subset**. We further evaluate models on these two subsets to prove the effectiveness of our model.

[1]Not publicly available as a test set for the DSTC competition.

Evaluation Metrics. The official metric adopted by SIMMC 2.0 response generation task is BLEU-4, which only focuses on n-grams overlap between the predicted and target response. For a more comprehensive comparison, we add widely-used machine generation metrics: BLUE-n (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), ROUGE (Lin and Hovy 2003) and CIDEr (Vedantam, Zitnick, and Parikh 2015) metrics. Compared with the accuracy based BLEU metric, METOR and ROUGH pay attention to recall and calculate how many n-grams from the target response exist in the predicted response, while CIDEr uses TF-IDF to assign larger weights to infrequent phrases.

Implementation Details. Our model is based on Transformer (Vaswani et al. 2017) structure with 12 layers, where ever Transformer block has 768 hidden units and 12 attention heads. Each patch is projected to features of the same size as the hidden units. We initialize SPRING parameters from pretrained VLM, *i.e.* , OFA (Wang et al. 2022b). During pretraining, our model is trained for 4 epochs with 8 batch sizes on 8 TESLA V100 GPU. Adam (Kingma and Ba 2015) is adopted as optimizer with a 4e-4 learning rate. Besides, the dropout rate is set to 0.2 to prevent over-fitting. During fine-tuning stage, we train 60 epochs on the SIMMC train set with a learning rate of 4e-5 and a batch size of 16.

Compared Methods. We compare SPRING with strong baseline methods from SIMMC 1.0 and SIMMC 2.0. On SIMMC 1.0, MN-MAG (Kim et al. 2021) adopts a memory network as encoder and designs multimodal fusion gate to fuse information. Tom (Jeong et al. 2021) esambles prediction results from several GPT-2 models. JBi-encoder (Huang et al. 2021a) is jointly trained to predict belief state and response. On SIMMC 2.0, MTN (Le et al. 2019) separately encodes multimodal input while the visual encoder is guided by a query-aware attention encoder. JMGPT (Kottur et al. 2021b) trains a multi-task GPT2-large, which takes dialogue history and flattened multimodal contexts as input. Furthermore, JMGPT-BS (Kottur et al. 2021a) extends JMGPT by inferring with different beam search sizes. MMBart (Lee et al. 2022) adds box coordinates embedding to textual in-

MODELS	SIMMC 2.0	VISUAL	SPATIAL
VLM	38.22	34.67	25.04
VLM			
w/ PVQA	40.75	36.54	27.58
w/ RVQA	41.27	37.02	27.22
w/ POVQA	40.89	35.94	28.08
w/ (PVQA + RVQA + POVQA)	41.36	37.59	28.24
VLM			
w/ PSQA	41.18	36.05	28.30
w/ RSQA	40.77	35.42	28.18
w/ VSQA	40.40	36.34	27.97
w/ (PSQA + RSQA + VSQA)	41.56	36.25	28.49
VLM			
w/ all QA	41.92	38.52	30.18
w/ (all QA + CL)	**42.49**	**38.87**	**30.77**

Table 3: Ablation study on SIMMC 2.0 with BLEU-4.

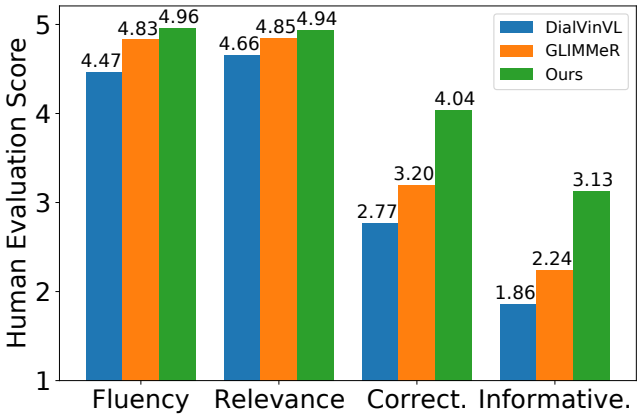

Figure 4: The human evaluation results on SIMMC 2.0.

put and proposes auxiliary tasks to predict asset attributes. DialVinVL (Kottur et al. 2021a) is based on VinVL-Base (Zhang et al. 2021), concatenates original box coordinates to region features as visual input, and incorporates dialogue history with dialogue policy as textual input. GPTDeIT (Lee and Han 2021) utilizes GPT2-large (Radford et al. 2019) as the text model to encode dialogue history and flattened slot values and DeIT-I (Touvron et al. 2021) as the image model to encode assets referenced by current turn utterance. JointGM (Nguyen et al. 2021) leverages BART-large (Lewis et al. 2020) to predict disambiguation label, belief state and response jointly according to inputted dialogue history. Similar to GPTDeIT, GLIMMeR (Hemanthage et al. 2021) also leverages GPT2-large and utilizes asset scene ID to help the model understand the semantics of each asset. Notably, GLIMMeR is the state-of-the-art method on SIMMC 2.0 and achieves the winner of the DSTC10.

4.2 Overall Performance

Table 2 displays the results of the model on the SIMMC 1.0 and SIMMC 2.0 dataset response generation task. It can be seen that SPRING has exceeded previous models by a large margin and achieved state-of-the-art results on all representative machine generation metrics. On SIMMC 2.0, the significant increased percentage on BLEU-n manifests our model successfully utilizing more accurate words and phrases to make responses. Our model also shows excellent performance on recall-based metrics METEOR and ROUGE, of which the score improvements reach 4.42 and 6.2. When the CIDEr metric pays more attention to infrequent n-grams, SPRING still outperforms GLIMMeR with 0.64 on CIDEr. Besides, according to the right part of Table 2, our model exhibits the highest BLEU-4 scores on the visual subset and spatial subset, which verifies the improvement of our model is produced by its better understanding of visual attribute and spatial relation and ability to conduct reasoning with aligned information to generate more accurate responses.

4.3 Detailed Analysis

Ablation Study. As shown in Table 3, we perform ablation experiments to evaluate the effectiveness of each pretraining task and curriculum learning strategy in SPRING. It can be observed that each MQA pretraining task brings significantly BLEU-4 improvement on the complete SIMMC 2.0 dataset compared with the basic VLM model. Specifically, VLM models pretrained with all visual QA tasks perform 2.92 higher than baseline on the Visual Subset, while VLM models pretrained with all spatial QA tasks display 3.45 improvement compared with baseline on the Spatial Subset, which can verify that visual QA and spatial QA respectively prompt model's ability of describing visual attribute and spatial relation. Besides, the last two rows further prove that our designed curriculum learning pretraining strategy effectively activates the potential of QA pretraining tasks and boosts model performance.

Human Evaluation. The human evaluation mainly focuses on 4 aspects: **fluency**, **relevance**, **correctness**, and **informativeness**, which are important for task-oriented dialogue systems. We randomly select 500 dialogues from SIMMC 2.0 dev-test dataset as candidates, and then filter these dialogues from the results generated by DialVinVL, GLIMMeR, and our model. We release evaluation task on Amazon Mechanical Turk (AMT) and make the last response of every selected dialogue evaluated by 10 different evaluators. Each evaluator scores 1500 generated responses on 4 aspects according to golden response in blind review from 1 to 5, simulating a real-life multimodal dialogue scenario. As shown in 4, it can be observed that our model consistently outperforms the other two models on all metrics, which is in line with automatic evaluation results.

5 Conclusion

In this paper, we propose a novel situated conversation agent pretraining method named SPRING. Specifically, all QA pairs and their difficulty labels used in pretraining are generated from our Incremental Layout Graph without any extra human annotations. Experimental results on SIMMC 1.0 and SIMMC 2.0 show that SPRING greatly surpasses previous models and describes visual attributes and spatial relations more accurately.

Acknowledgements

We would like to sincerely thank anonymous reviewers for their suggestions and comments. The work was partially supported by the National Natural Science Foundation of China (NSFC62076032). We also want to express our gratitude for precious advises given by Guanqi Zhan.

References

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Goldstein, J.; Lavie, A.; Lin, C.; and Voss, C. R., eds., *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization @ ACL*.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Chang, X.; Ren, P.; Xu, P.; Li, Z.; Chen, X.; and Hauptmann, A. 2020. A survery of scene graph: Generation and application. In *Proceedings of IEEE Transactions on Neural Network Learning*.

Chattopadhyay, P.; Yadav, D.; Prabhu, V.; Chandrasekaran, A.; Das, A.; Lee, S.; Batra, D.; and Parikh, D. 2017. Evaluating Visual Conversational Agents via Cooperative Human-AI Games. In *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing (AAAI)*.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *Proceedings of the European conference on computer vision (ECCV)*, 104–120. Springer.

Cho, J.; Lei, J.; Tan, H.; and Bansal, M. 2021. Unifying vision-and-language tasks via text generation. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M.; Parikh, D.; and Batra, D. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

De Vries, H.; Strub, F.; Chandar, S.; Pietquin, O.; Larochelle, H.; and Courville, A. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, W.; Dai, Y.; Hui, B.; Yang, M.; Cao, Z.; Dong, J.; Huang, F.; Si, L.; and Li, Y. 2022a. SPACE-2: Tree-Structured Semi-Supervised Contrastive Pre-training for Task-Oriented Dialog Understanding. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*.

He, W.; Dai, Y.; Yang, M.; Sun, J.; Huang, F.; Si, L.; and Li, Y. 2022b. SPACE-3: Unified Dialog Model Pre-training for Task-Oriented Dialog Understanding and Generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

He, W.; Dai, Y.; Zheng, Y.; Wu, Y.; Cao, Z.; Liu, D.; Jiang, P.; Yang, M.; Huang, F.; Si, L.; et al. 2022c. SPACE: A Generative Pre-trained Model for Task-Oriented Dialog with Semi-Supervised Learning and Explicit Policy Injection. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Hemanthage, B.; Lemon, O.; Dondrup, C.; and Bartie, P. 2021. GLIMMeR: Global-Local Information-aware Multimodal grounding with GPT for Co-reference Resolution. In *DSTC10 challenge wokrshop at AAAI*.

Huang, X.; Tan, C. S.; Ng, Y. B.; Shi, W.; Yeo, K. H.; Jiang, R.; and Kim, J. J. 2021a. Joint generation and bi-encoder for situated interactive multimodal conversations. In *DSTC9 challenge wokrshop at AAAI*.

Huang, Z.; Zeng, Z.; Huang, Y.; Liu, B.; Fu, D.; and Fu, J. 2021b. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hui, B.; Geng, R.; Ren, Q.; Li, B.; Li, Y.; Sun, J.; Huang, F.; Si, L.; Zhu, P.; and Zhu, X. 2021. Dynamic Hybrid Relation Exploration Network for Cross-Domain Context-Dependent Semantic Parsing. In *AAAI*.

Hui, B.; Geng, R.; Wang, L.; Qin, B.; Li, Y.; Li, B.; Sun, J.; and Li, Y. 2022. S^2SQL: Injecting Syntax to Question-Schema Interaction Graph Encoder for Text-to-SQL Parsers. In *Findings of the Association for Computational Linguistics: ACL 2022*, 1254–1262. Dublin, Ireland: Association for Computational Linguistics.

Jeong, Y.; Lee, S. J.; Ko, Y.; and Seo, J. 2021. Tom: End-to-end task-oriented multimodal dialog system with gpt-2. In *DSTC9 challenge wokrshop at AAAI*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Kim, B.; Lee, I.; Jeong, Y.; Youngjoong, K.; Koo, M.-W.; and Seo, J. 2021. Improving multimodal api prediction via adding dialog state and various multimodal gates. In *DSTC9 challenge wokrshop at AAAI*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Kottur, S.; Moon, S.; Geramifard, A.; and Damavandi, B. 2021a. Overview of Situated and Interactive Multimodal Conversations (SIMMC) 2.0 Track at DSTC 10. In *DSTC10 challenge wokrshop at AAAI*.

Kottur, S.; Moon, S.; Geramifard, A.; and Damavandi, B. 2021b. SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Le, H.; Sahoo, D.; Chen, N.; and Hoi, S. 2019. Multimodal transformer networks for end-to-end video-grounded dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (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. Learning to Embed Multi-Modal Contexts for Situated Conversational Agents. In *Findings of the Association for Computational Linguistics (NAACL)*.

Lee, J.; and Han, K. 2021. Multimodal Interactions Using Pretrained Unimodal Models for SIMMC 2.0. In *DSTC10 challenge wokrshop at AAAI*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Li, W.; Gao, C.; Niu, G.; Xiao, X.; Liu, H.; Liu, J.; Wu, H.; and Wang, H. 2021. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Liao, L.; Long, L. H.; Ma, Y.; Lei, W.; and Chua, T.-S. 2021. Dialogue State Tracking with Incremental Reasoning. *Transactions of the Association for Computational Linguistics*, 9: 557–569.

Lin, C.; and Hovy, E. H. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the Advances in neural information processing systems (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. Situated and Interactive Multimodal Conversations. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*.

Nguyen, T.-T.; Shi, W.; Jiang, R.; and jae Kim, J. 2021. Multimodal and Joint Learning Generation Models for SIMMC 2.0. In *DSTC10 challenge wokrshop at AAAI*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Qiu, L.; Liang, Y.; Zhao, Y.; Lu, P.; Peng, B.; Yu, Z.; Wu, Y. N.; and Zhu, S. 2021. SocAoG: Incremental Graph Parsing for Social Relation Inference in Dialogues. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP*.

Qiu, L.; Zhao, Y.; Liang, Y.; Lu, P.; Shi, W.; Yu, Z.; and Zhu, S.-C. 2022. Towards Socially Intelligent Agents with Mental State Transition and Human Value. In *SIGDIAL*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. In *OpenAI Blog*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the Advances in neural information processing systems (NIPS)*.

Saha, A.; Khapra, M.; and Sankaranarayanan, K. 2018. Towards building large scale multimodal domain-aware conversation systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Proceedings of the Advances in neural information processing systems (NIPS)*.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. CIDEr: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, L.; Qin, B.; Hui, B.; Li, B.; Yang, M.; Wang, B.; Li, B.; Huang, F.; Si, L.; and Li, Y. 2022a. Proton: Probing Schema Linking Information from Pre-trained Language Models for Text-to-SQL Parsing. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022b. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.