Visual Spatial Description: Controlled Spatial-Oriented Image-to-Text Generation

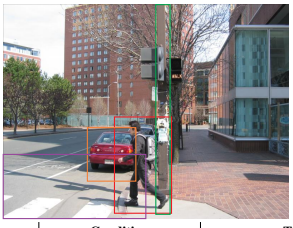
Yu Zhao[1], **Jianguo Wei**[1], **Zhichao Lin**[2], **Yueheng Sun**[1], **Meishan Zhang**[3]*, **Min Zhang**[3]
[1]College of Intelligence and Computing, Tianjin University
[2]School of New Media and Communication, Tianjin University
[3]Institute of Computing and Intelligence, Harbin Institute of Technology (Shenzhen)
{zhaoyucs,jianguo,chaosmyth,yhs}@tju.edu.cn,
{zhangmeishan,zhangmin2021}@hit.edu.cn

Abstract

Image-to-text tasks, such as open-ended image captioning and controllable image description, have received extensive attention for decades. Here, we further advance this line of work by presenting Visual Spatial Description (VSD), a new perspective for image-to-text toward spatial semantics. Given an image and two objects inside it, VSD aims to produce one description focusing on the spatial perspective between the two objects. Accordingly, we manually annotate a dataset to facilitate the investigation of the newly-introduced task and build several benchmark encoder-decoder models by using VL-BART and VL-T5 as backbones. In addition, we investigate pipeline and joint end-to-end architectures for incorporating visual spatial relationship classification (VSRC) information into our model. Finally, we conduct experiments on our benchmark dataset to evaluate all our models. Results show that our models are impressive, providing accurate and human-like spatial-oriented text descriptions. Meanwhile, VSRC has great potential for VSD, and the joint end-to-end architecture is the better choice for their integration. We make the dataset and codes public for research purposes.[1]

1 Introduction

Text generation from images is a widely-adopted means for deep understanding of cross-modal data that has received increasing interest of both computer vision (CV) and natural language processing (NLP) communities (He and Deng, 2017). Image-to-text tasks generate natural language texts to assist in understanding the scene meaning of a specific image, which might be beneficial for a variety of applications such as image retrieval (Diao et al., 2021; Ahmed et al., 2021), perception assistance (Xu et al., 2018; Shashirangana et al., 2021), pedestrian detection (Hasan et al., 2021), and medical system (Miura et al., 2021).



Task	Condition	Target Text
Image Captioning	——	A man is walking past a car.
VSR-guided Captioning	walk; ⟨Arg⟩, ⟨Loc⟩	A man is walking cross a street.
Visual Question Answering	What color is the car?	The car is red.
Our Task: VSD	⟨man, car⟩	A man is walking behind a red car from right to left.
	⟨car, pole⟩	A red car is parked to the left of a pole.

Figure 1: A comparison of three example image-to-text generation tasks and the proposed VSD in this work.

Image-to-text tasks take on various forms when serving different purposes. Figure 1 illustrates a comparison of three example tasks. First, the generic open-ended image captioning aims to provide a summarised description that describes an input image and reflects the overall understanding of the image (Lindh et al., 2020; Vinyals et al., 2015; Ji et al., 2020). Furthermore, the verb-specific semantic roles (VSR) guided captioning (Chen et al., 2021) and visual question answering (VQA) (Antol et al., 2015) are two examples of controllable image description, which produce human-like and stylized descriptions under specified conditions based on a thorough comprehension of the input image (Chen et al., 2021; Fei et al., 2021b; Mathews et al., 2018; Cornia et al., 2019; Lindh et al., 2020; Pont-Tuset et al., 2020; Deng et al., 2020; Zhong et al., 2020; Kim et al., 2019; Chen et al., 2020a; Fei et al., 2022; Jhamtani and Berg-Kirkpatrick, 2018). The VSR-guided captioning produces a description focusing on a verb with specified semantic roles, and the VQA generates a reasoning answer based on a given question.

In this work, we extend the line of controllable image description by presenting the spatial semantics of image-to-text, which is essential but has received little attention previously. Spatial seman-

*Corresponding author
[1]https://github.com/zhaoyucs/VSD

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 1437 - 1449
December 7-11, 2022

tics is a fundamental aspect of both language and image interpretation in relation to human cognition (Zlatev, 2007), and it has shown great value in spatial-based applications such as automatic navigation, personal assistance, and unmanned manipulation (Irshad et al., 2021; Raychaudhuri et al., 2021; Zeng et al., 2018). Here, we introduce a new task, Visual Spatial Description (VSD), which generates text pieces to describe the spatial semantics in the image. The task takes an image with two specified objects in it as inputs and outputs one sentence that describes the detailed spatial relation of the objects. We manually annotate a dataset for inquiry to benchmark this task.

VSD is a typical vision-language generation problem that can be addressed by multi-modal encoder-decoder modeling. Multi-modal models allow both visual and linguistic inputs and encode them to a joint representation that can learn information from both modal inputs. Moreover, recent studies show that vision-language pretraining can bring remarkable achievements in most image-to-text tasks (Lu et al., 2019; Sun et al., 2019; Tan and Bansal, 2019; Zhou et al., 2020; Li et al., 2019; Hu and Singh, 2021; Li et al., 2021; Xiao et al., 2022). Here, we follow these tasks and adopt VL-BART and VL-T5 (Cho et al., 2021) as backbones, which exhibit state-of-the-art performance in vision-language generation.

In particular, a closely-related task, visual spatial relationship classification (VSRC), which outputs the spatial relationship between two objects inside an image, might be beneficial for our proposed VSD. The predefined discrete spatial relations such as "next to" and "behind", in VSRC should be able to effectively guide the VSD generation. To this end, we first make a thorough comparison of the connections between VSD and VSRC, which can be regarded as shallow and deep analyses of spatial semantics, respectively, and further investigate the VSRC-enhanced VSD models, performing visual spatial understanding from shallow to deep. Specifically, we present two straightforward architectures to integrate VSRC into VSD, one being the pipeline strategy and the other being the end-to-end joint strategy, respectively.

Finally, we conduct experiments on our constructed dataset to evaluate all proposed models. First, we examine the two start-up models for VSD only with VL-BART and VL-T5. The results show that the two models are comparable in terms of performance, and both models can provide highly accurate and fluent human-like outputs of spatial understanding. Second, we verify the effectiveness of VSRC for VSD and find that: (1) VSRC has great potentials for VSD because gold-standard VSRC can lead to striking improvements on VSD; (2) VSD can be benefited from automatic VSRC, and the end-to-end joint framework is slightly better. We further perform several analyses to intensively understand VSD and the proposed models.

2 Related Work

Image-to-text has been intensively investigated with the support of neural networks in the past years (He and Deng, 2017). The encoder-decoder architecture is an often considered framework, where the encoder extracts visual features from the image and the decoder generates text for specific tasks. Early works employ a convolutional neural network (CNN) as the visual encoder and a recurrent neural network (RNN) as the text decoder (Vinyals et al., 2015; Rennie et al., 2017). Recently, the Transformer neural network (Vaswani et al., 2017), which is impressively powerful in feature representation learning on both vision and language, has gained increasing interest. The Transformer-based encoder-decoder models have been adopted in a wide range of image-to-text tasks (Cornia et al., 2020; Herdade et al., 2019; Fei et al., 2021a). These models coupled with visual-language pretraining have achieved the top performance for these tasks (Lu et al., 2019; Sun et al., 2019; Tan and Bansal, 2019; Zhou et al., 2020; Li et al., 2019; Hu and Singh, 2021; Li et al., 2021). In this work, we exploit the Transformer-based architecture and two pretrained visual-language models: VL-BART and VL-T5 (Cho et al., 2021), reaching several strong benchmark models for our task.

Image-to-text can be varied depending on the objective of the visual description. Image captioning is the most well-studied task, which aims to summarize a given image or to describe a particular region in it (Karpathy and Fei-Fei, 2015; Vinyals et al., 2015). Several subsequent studies have attempted to produce captions with specified patterns and styles (Cornia et al., 2019; Kim et al., 2019; Deng et al., 2020; Zhong et al., 2020; Zheng et al., 2019). For example, VQA and visual reasoning can be regarded as such attempts, which are conditioned by a specific question directed at the input image (Antol et al., 2015; Agrawal et al., 2018; Hudson

and Manning, 2019; Johnson et al., 2017). The VSR-guided image captioning (Chen et al., 2021) is the most close to our work, which generates a sentence for a particular event in the image with well-specified semantic roles. Here we focus on spatial semantics instead, generating a description based on the spatial relationship.

Spatial semantics is an important topic in both language and visual analysis. Kordjamshidi et al. (2011) propose an preliminary study on text-based spatial role labeling. Later, spatial element extraction and relation extraction from texts are investigated by (Nichols and Botros, 2015). Pustejovsky et al. (2015) present a fine-grained spatial semantic analysis in texts with rich spatial roles. Based on the image input, Yang et al. (2019) propose VSRC and benchmark it with a manually-crafted dataset. The VSRC is actually a shallow task for visual spatial analysis based on a closed relationship set and by using a simple classification schema. Following, Chiou et al. (2021) build a much stronger model on the dataset. Many studies have exploited spatial semantics to assist other image understanding tasks (Kim et al., 2021; Wu et al., 2021; Collell et al., 2021; Xiao et al., 2021; Pierrard et al., 2021). In addition, learning spatial representations from multiple modalities also receives particular attention (Collell and Moens, 2018; Dan et al., 2020). In this work, we extend image-to-text and propose VSD, which aims for the spatial understanding of the image.

3 Visual Spatial Description

3.1 Task Description

Formally, we define the task of VSD as follows: given an image I and an object pair $\langle O_1, O_2 \rangle$ inside I, the VSD aims to output a word sequence $S = \{w_1, ..., w_n\}$ to describe the spatial semantics between O_1 and O_2. The provided O_1 and O_2 include both the object tags and their bounding boxes. In Figure 1, we would receive "A man is walking behind a red car from right to left." for $\langle man, car \rangle$ and "A red car is parked to the left of a pole." for $\langle car, pole \rangle$ based on the same input image. The generated sentences of VSD must encode the spatial semantics between the given two objects, which differs from conventional image-to-text generation.

3.2 Compared with VSRC

Noticeably, VSRC is another representative task of visual spatial understanding that decides the spatial relation of two objects in an image. The relation is chosen from a closed set which is manually predefined. We can regard VSRC as a shallow analysis task for spatial semantics understanding, while the VSD task can offer a deeper spatial analysis by using the much more flexible output.

In particular, compared with VSRC, VSD has three major advantages. First, VSD can offer richer semantics which could be necessary for spatial understanding. Meanwhile, VSRC only outputs a spatial relation from a closed set in general. VSD can raise other semantic roles to deepen the spatial understanding beyond the relations, such as predicates and object attributes. Second, the spatial relations might be overlapped. For example, the two relationships, "behind" and "to the right of" might be both correct for VSRC given the "man" and "car" in Figure 1. The newly proposed task VSD can more accurately describe the multiple spatial semantics. Third, from the viewpoint of downstream tasks, especially the systems that require automatic content-based image indexing or visual dialogue, VSD is more straightforward and adequate to support them.

3.3 Data Collection

We build an initial dataset To benchmark the VSD task. The constructed dataset is extended from a VSRC dataset to facilitate the investigation between VSD and VSRC. Thus, our final corpus includes both VSRC and VSD annotations.

Our VSRC dataset is sourced from two existing datasets: SpatialSense (Yang et al., 2019) and VisualGenome (Krishna et al., 2017). SpatialSense is a dataset initially constructed for VSRC with nine well-defined spatial relations, namely, "on", "in", "next to", "under", "above", "behind", "in front of", "to the left of", and "to the right of". The only disadvantage of SpatialSense is its relatively-small scale. Consequently, we enlarge the corpus with the help of VisualGenome a widely adopted dataset for scene graph generation with annotations in the form of ⟨subject, predicate, object⟩. We add the triplets in VisualGenome, whose predicates can be easily aligned with the nine spatial relations in SpatialSense.[2] Accordingly, we can obtain a larger dataset of VSRC.

We develop a simple visualization tool to facilitate the VSD annotation. The system randomly as-

[2]The alignment is achieved by a map, which will be released along with the dataset.

Object Pair: ⟨man, woman⟩

1: Description Writing
The man in a white shirt is standing on the right of the woman laughing.
The man in white and the woman laughing are standing opposite each other.
The man is standing on the right of the woman.
The man is sitting next to the woman laughing.
The man in white is standing in front of the woman laughing.
2: Description Checking
The man in a white shirt is standing on the right of the woman laughing. ✓
The man in white and the woman laughing are standing opposite each other. ✓
The man is standing on the right of the woman. ✗
The man is sitting next to the woman laughing. ✗
The man in white is standing in front of the woman laughing. ✗

Figure 2: The data annotation flow.

signs the instances to the annotators. Each instance contains one image and two objects inside it. The annotators are asked to write text descriptions for the given instance. We also set up another interface for experts to check the correctness of all annotated sentences and to ensure the quality of these written descriptions. In the description checking step, the given instances include the image inputs, paired objects, and the written descriptions by the first step. The annotator mainly checks whether the description is valid. The annotation flow is shown in Figure 2.

All annotators we recruited are college students who are native English speakers. During the preparation, we train the annotators with a guideline and perform two pre-annotation tests from easy to difficult. In the first test, the annotators are asked to participate in the checking interface, where several well-written descriptions are prepared in advance by experts and various pseudo-ill-conditioned descriptions by intentional word substitutions. Thereafter, we start the second test to let annotators write the real spatial descriptions. The annotators are allowed for official annotations only when both tests are passed. All annotators are properly paid under the open market competition.

Specifically, we have three basic principles mainly in our data annotation guideline as follows:

- The sentence must correctly describe the spatial semantics of the given object pair.
- The descriptions can help us correctly locate the exacted objects in the image.
- The length of each text description should be limited to no more than 40 tokens.

The annotation submissions with excess invalid annotations (more than 4/100) according the above principles would be returned to the annotators to rework until it reaches the standard. Figure 2 shows some examples of invalid annotations. There might be several exceptions, such as, spelling mistakes or mismatches between the image and object tag inputs. In these cases, annotators should skip and report these instances, leaving them for further discussions by expert. In the expert-checking step, the remaining invalid and controversial annotations would be be discussed and then finalized.

Finally, we annotate a total of 29K images with 143K descriptions, wherein 6,591 images are from the SpatialSense with 9,744 descriptions, and the remaining images and descriptions are sourced from VisualGenome. Furthermore, we randomly split the whole annotated VSD dataset by a ratio of 7:1:2 as training / validation / testing sections. The statistics of the dataset are shown in Table 1.

	Input		Output	
Sect.	#Img	#OBJ-TAG	#SENT	AvgLEN
Train	20,490	4,506	116,791	7.35
Dev	2,927	1,416	16,823	7.33
Test	5,855	2,104	10,038	8.04

Table 1: The statistics of our constructed VSD dataset.

4 Model

We exploit the Transformer-based encoder-decoder architecture to accomplish our VSD goal. The architecture can be partially pretrained from ultra-large-scale self-supervised datasets, which makes it capable of obtaining the top performance on a range of image-to-text generation tasks (Hu and Singh, 2021; Li et al., 2021; Tan and Bansal, 2019; Chen et al., 2020b). In this section, we first briefly summarize the adopted model architecture and describe the two well-pretrained models exploited as backbones.

4.1 Model Architecture

The encoder-decoder architecture contains a vision-language (V&L) encoder and a text decoder. The encoder takes the combined V&L inputs to learn

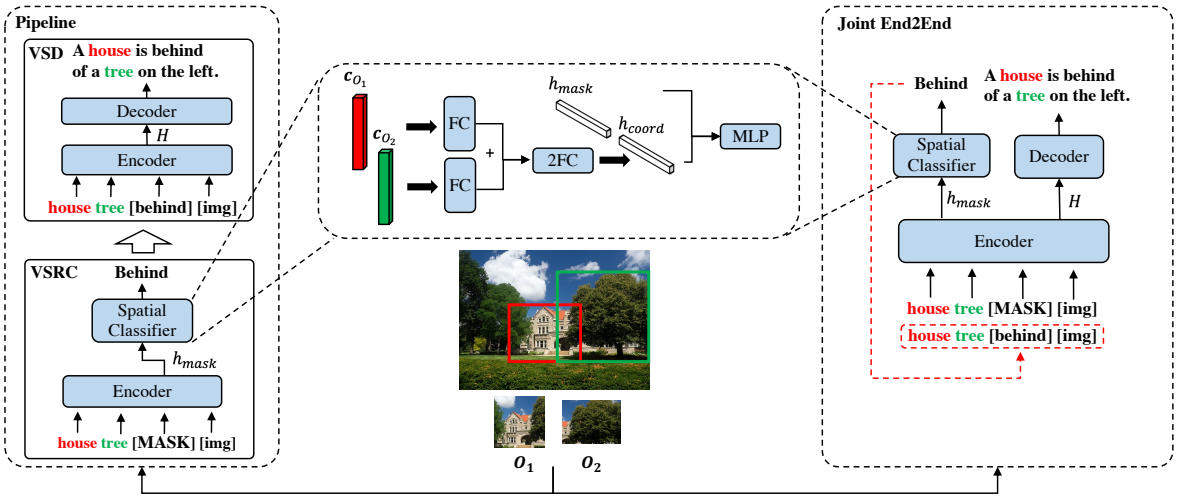


Figure 3: Overview of our pipeline and end-to-end models with VSRC, where FC denotes fully-connected network.

a joint feature representation, and the text decoder generates the output sentential words incrementally with the joint representation.

Formally, the VSD input includes: (1) one image I and (2) the inside object pair $\langle O_1, O_2 \rangle$. First, we obtain a sequence of visual features by:

$$\begin{aligned} \boldsymbol{F}^{\mathrm{V}} &= \mathrm{VisionExtractor}(I), \\ \boldsymbol{E}^{\mathrm{V}} &= \mathrm{FC}(\boldsymbol{F}^{\mathrm{V}}), \end{aligned} \tag{1}$$

where $\boldsymbol{F}^{\mathrm{V}}$ is obtained by a Faster R-CNN (Ren et al., 2015). Then a fully-connected (FC) linear transformation layer is used to align the vectorial dimensions between vision and language, leading to $\boldsymbol{E}^{\mathrm{V}}$.

Second, a textual embedding layer is used to represent O_1 and O_2 by their respective textual tags (i.e., T_{O_1} and T_{O_2}):

$$\boldsymbol{E}^{\mathrm{T}} = \mathrm{TextEmbed}([T_{O_1}, T_{O_2}]), \tag{2}$$

where $\boldsymbol{E}^{\mathrm{T}}(|\boldsymbol{E}^{\mathrm{T}}| = |T_{O_1}| + |T_{O_2}|)$ ($|\cdot|$ indicates the sequence length) is the textual representation of the two objects.

Thereafter that, a Transformer is exploited to produce the final encoder output by the following expression:

$$\boldsymbol{H} = \mathrm{Transformer}([\boldsymbol{E}^{\mathrm{T}}, \boldsymbol{E}^{\mathrm{V}}]), \tag{3}$$

where $\boldsymbol{E}^{\mathrm{T}}$ and $\boldsymbol{E}^{\mathrm{V}}$ are concatenated and then fed into the Transformer, resulting in $\boldsymbol{H}$ which is a sequence of the high-level joint V&L representations.

We generate a sequence of words incrementally for the decoder, wherein one word is predicted each step based on the previous context:

$$\boldsymbol{o}_j(\boldsymbol{y}|\boldsymbol{y}_{i<j}) = \mathrm{FC}\big(\mathrm{Transformer}(\boldsymbol{y}_{i<j}, \boldsymbol{H})\big), \tag{4}$$

where $\boldsymbol{y}_i < j$ denotes the previously generated tokens and $\boldsymbol{H}$ represents the encoder outputs. The decoder is also dominated by Transformer. Thereafter, an FC layer is used to score all candidate words.

We exploit the cross-entropy as objective loss to train the model, following the majority of sentence generation models (Lewis et al., 2020; Raffel et al., 2020). During the decoding, we can apply the beam search algorithm to obtain better results.

4.2 VL-BART and VL-T5

VL-BART is a well-pretrained model that can be directly applied to our VSD model with an initializing-then-fine-tuning mode. VL-BART is a standard Transformer-based model similar to our VSD model with a bidirectional joint V&L encoder and an autoregressive text decoder, which is extended from BART (Lewis et al., 2020) by importing an extra visual embedding module for the joint encoding. Before pretraining, VL-BART is partially initialized with BART on the shared parameters, which is trained on the text-only corpus by corrupting documents and optimizing the model by a reconstruction loss.

VL-T5 is similar to VL-BART on model architecture but differs in that it extends from T5 (Raffel et al., 2020). The T5 model uses relative position embeddings on text representation and is trained on a very different text-only corpus with a span-based reconstruction process.

5 Enhancing with VSRC

Our VSD task aims to control image-to-text generation by the aspect of spatial semantics. If we know the explicit spatial relation by VSRC in advance for the given two objects, then the description generation could be more instructional. In this section, we introduce two architectures of integrating VSRC into the above-mentioned VSD models.

5.1 Pipeline

The pipeline architecture includes two stages. In the first stage, VSRC is executed to extract spatial relations between the two given objects of the VSD input. In the second stage, our VSD model adds the spatial relation as one additional textual input to enhance the encoder. We illustrate the architecture in the left portion of Figure 3.

Our VSRC model takes the same input as VSD, an image and two objects inside it. Accordingly, our encoder can be highly similar to that of the VSD model: VisionEmbed and TextEmbed, followed by the Transformer as mentioned in Equations 2 and 3. Here, we make a slight modification to adapt the VSRC task. Specifically, a special [MASK] token is added inside the TextEmbed module:

$$\boldsymbol{E}^{\mathrm{T}} = \mathrm{TextEmbed}([T_{O_1}, T_{O_2}, \mathrm{MASK}]), \quad (5)$$

where the updated TextEmbed has been illustrated in Figure 3 by the input depiction of the VSRC. We only use one vector $\boldsymbol{h}_{\mathrm{MASK}}$ from the sequential encoder output $\boldsymbol{H}$ for relation classification, which is exactly corresponding to the position of the special [MASK] token.

Before the final-step classification, we follow (Chiou et al., 2021) to add the bounding box coordinates of the two objects for geometric information. Each bounding box is converted into a 4-dimensional vector, thus we have $\boldsymbol{c}_{O_1}$ and $\boldsymbol{c}_{O_2}$ for the two objects, respectively. Then, we use the following fully-connected (FC) networks sequentially to reach a bounding box representation:

$$\begin{aligned} \tilde{\boldsymbol{h}}_{\mathrm{coord}} &= \mathrm{FC}(\boldsymbol{c}_{O_1}) + \mathrm{FC}(\boldsymbol{c}_{O_2}), \\ \boldsymbol{h}_{\mathrm{coord}} &= \mathrm{FC}(\mathrm{FC}(\tilde{\boldsymbol{h}}_{\mathrm{coord}})), \end{aligned} \quad (6)$$

where $\boldsymbol{h}_{\mathrm{coord}}$ is the desired bounding box representation. Finally, we concatenate $\boldsymbol{h}_{\mathrm{coord}}$ and $\boldsymbol{h}_{\mathrm{MASK}}$ to score candidate spatial relations:

$$\boldsymbol{o}^{\mathrm{VSRC}} = \mathrm{MLP}^{\mathrm{VSRC}}([\boldsymbol{h}_{\mathrm{MASK}}, \boldsymbol{h}_{\mathrm{coord}}]), \quad (7)$$

where $\mathrm{MLP}^{\mathrm{VSRC}}$ is the classifier for VSRC. In this way, the VSRC task is accomplished. The middle part of Figure 3 shows the detailed network operation of the classification.

Our VSD task receives three types of inputs from the VSRC output, with additional spatial relation as one supplement compared with the original VSD. Considering the textual property of the spatial relation, we add this information to the textual embedding of the original VSD encoder:

$$\boldsymbol{E}^{\mathrm{T}} = \mathrm{TextEmbed}([T_{O_1}, T_{O_2}, r_{O_1,O_2}]), \quad (8)$$

where r_{O_1,O_2} is the textual expression of the spatial relation between the given objects O_1 and O_2. This distinction is the only difference between the VSRC-enhanced and the original VSD models, and the other parts remain the same.

5.2 End to End

The end-to-end model for joint VSRC and VSD is not only more elegant in form, allowing their full natural interactions, but also can avoid the error propagation problem where the VSRC errors may result in further degraded VSD performance.

We adopt multi-task learning (MTL) to achieve the end-to-end goal with a single model. Figure 3 shows the detailed structure by the right part. The joint encoder is directly borrowed from the individual VSRC model, resulting in the encoder output $\boldsymbol{H}$. Thus, the input of the joint model is the same as the original VSD model and the VSRC model. Then, we execute the decoders of VSRC and VSD, achieving the goal of joint learning.

During the training, given the VSRC input (also the joint input) and the VSRC and VSD outputs, we optimize the end-to-end model by the joint loss, which is a weighted addition of the VSRC and VSD losses. During the inference, we have two strategies for our VSD task. First, we can use the end-to-end joint model to simultaneously obtain the VSRC and VSD results under the MTL architecture (Figure 3 End2End without the red dashed line). Second, we can execute the end-to-end model by two rounds, where the first round outputs the VSRC result, and the second round uses the VSRC result to substitute the [MASK] part of the joint encoder, and then executes the VSD part only. The second strategy is similar to the pipeline architecture, but only a single model is involved.

	VSD					VSRC
	BLEU-4	METEOR	ROUGE	CIDEr	SPICE	Acc(%)
VL-BART	52.71	41.96	77.57	471.21	67.83	-
VL-BART+VSRC-pipeline	53.49	42.14	77.79	474.34	67.97	53.32
VL-BART+VSRC-end2end	**53.60**	**42.45**	**78.15**	**476.47**	**68.18**	**54.53**
VL-BART+VSRC-golden	72.30	50.90	87.44	578.27	76.59	golden
VL-T5	52.58	41.94	77.63	472.24	67.90	-
VL-T5+VSRC-pipeline	53.71	42.56	78.33	480.32	68.72	53.50
VL-T5+VSRC-end2end	**54.31**	**42.63**	**78.38**	**481.13**	**68.74**	**56.36**
VL-T5+VSRC-golden	72.12	50.95	87.54	579.41	77.29	golden
OSCAR$^+$	37.17	35.06	66.47	427.21	67.41	-
OSCAR$^+$+VSRC-end2end	**38.70**	**35.81**	**67.89**	**438.28**	**67.54**	**57.90**

Table 2: The main results of our proposed models on the VSD test dataset, where we implement three types of baseline models (i.e., VL-BART, VL-T5 and OSCAR$^+$), and the ones equipped with VSRC supporting.

6 Experiments

6.1 Setup

Implementation Details We initialize our encoder-decoder backbone with two pretrained models VL-BART and VL-T5, and follow (Anderson et al., 2018) to obtain visual region features from Faster R-CNN. We use the two-round strategy as default for the decoding of the end-to-end models with VSRC, . We present more model details and hyper-parameters in Appendix A.

Evaluation We report five standard evaluation metrics of the text generation for the VSD task, including BLEU-4 (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016). In this work, we use BLEU-4 and SPICE as the primary metrics to evaluate our models, where the former can measure the syntactic quality of the generated descriptions, and the latter emphasizes the consistency with the input scene graphs. Although CIDEr has been widely-adopted for image-to-text generation as the major metric, it might be unsuitable for our VSD task because it can lower the importance of frequently occurring words closely related to spatial relations by the IDF values. We conduct each experiment by five times and report the average number.

6.2 Main Results

Table 2 shows the main results on the test dataset. The model results based on VL-BART and VL-T5 are reported in two different regions. The first row of each region shows the performance of our original models. The base VL-BART and VL-T5 models can achieve impressive performance as a whole, and the two models are generally comparable. The rows with "+VSRC–*" stand for the results of our VSD models with the support of spatial relation. Meanwhile, the "VL-T5-*" models demonstrate better performance under this setting.

To show the potential of VSRC for VSD, we first examine the oracle performance with gold-standard spatial relations as input. The results are highly exciting, as shown by "+VSRC–golden" with the gray numbers. We can obtain very large improvements over all evaluation metrics based on both VL-BART and VL-T5. The observation indicates that spatial relation is very useful to our VSD task. However, using gold-standard spatial relations in real scenarios is impractical. Thus, it is deserved to investigate the benefits of spatial relations outputted from a VSRC model.

Spatial relations from a VSRC model can be incorporated in two ways, as shown by "+VSRC-pipeline" and "+VSRC-end2end" in Table 2. The two types of models show significant performance decreases compared with that of "+VSRC-golden". Nonetheless, these models can still lead to positive gains on the VSD task by comparing their performance with the basic models without spatial relation information. In addition, our end-to-end joint models (i.e., "+VSRC-end2end") outperform their corresponding pipelines. If the gain on VSRC is larger, then the increase on VSD is also more significant, indicating that the VSRC performance is the key.

Furthermore, we compare our VL-BART and VL-T5 models with another representative image-

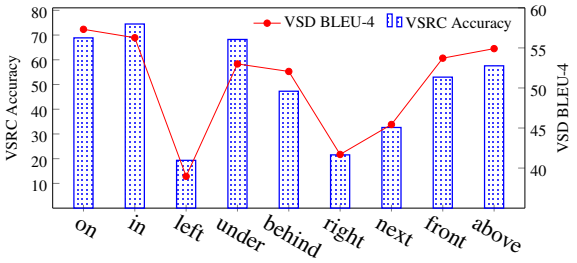
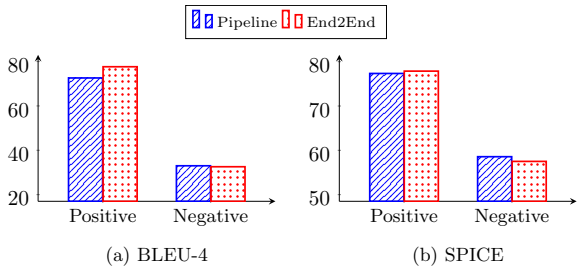


Figure 4: Fine-grained results of the VL-T5+VSRC-end2end model in terms of spatial relations.



(a) BLEU-4 (b) SPICE

Figure 5: VSD results of VL-T5+VSRC-end2end by Positive and Negative relations predicted from VSRC.

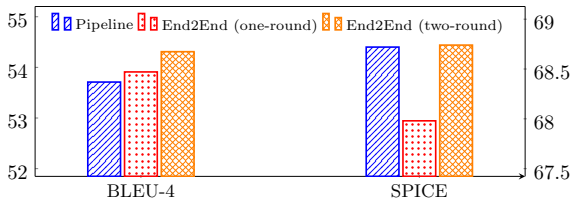


Figure 6: A comparison of the VL-T5+VSRC-end2end model by using one-round and two-round decodings.

to-text model, namely OSCAR$^+$ (Zhang et al., 2021). The major difference between our models and OSCAR$^+$ is that OSCAR$^+$ exploits VL-BERT as the backbone, which only contains an encoder. The spatial relation can effectively improve the OSCAR$^+$ model as well. Notice that VL-BERT excels at understanding tasks because of its discriminative pretraining benefiting based on sole encoder learning, so we can find that Oscar$^+$+VSRC-end2end can achieve the best VSRC accuracy. Overall, the OSCAR$^+$ models still obtain lower VSD performance than our suggested VL-BART and VL-T5 models, demonstrating the advantage of the encoder-decoder pretraining on the VSD task.

6.3 Discussion

Fine-grained Performance by Spatial Relations The performance differences among various spatial relations are interesting. Several particular relations may be more difficult to comprehend within the images. Figure 4 shows the BLEU-4 results across different spatial relations by the VL-T5+VSRC-end2end model, where the VSRC precisions are also shown for comparison. Overall, one approximative correction exists between the VSD and VSRC performance by fine-grained evaluation. Additionally, spatial relations such as "to the left of" and "to the right of" show significantly lower performance than the others. The two possible reasons are as follows: (1) These relations (e.g., including ambiguities by compounds) are visually not as clear as the others, such as "on", "under", and "in". (2) The distribution of spatial relations is unbalanced. Although we have paid particular attention to this issue while building our dataset, this problem is still challenging to handle due to the natural characteristic of spatial semantics.

Pipeline v.s. End-to-End To further understand the disparity between the pipeline and the end-to-end models, we divide the model outputs by the VSRC correctness into two categories and then evaluate the VSD results on them separately. Concretely, if the VSRC output is correct, then we regard the instance as positive; otherwise, it is negative. Figure 5 shows the BLEU-4 and SPICE results by the VL-T5+VSRC-end2end model. The end-to-end model outperforms the pipeline model on the positive samples while doing the opposite on negative samples. The obversion is reasonable because the end-to-end model tends to trust its VSRC output by its overall positive influence, thus resulting in downgraded performance when the VSRC outputs are incorrect. According to the finding, we can see that the VSRC accuracy is vital for the final VSD performance.

Decoding in End-to-end: One Round or Two As mentioned in Section 5.2, we have two strategies for the decoding of the end-to-end models. The two-round strategy is selected by default. Here, we compare the two decoding strategies based on the VL-T5+VSRC-end2end model. Figure 6 shows the results, where the pipeline results are also shown for reference. The two-round decoding is highly critical for the end2end model, without which the model can even be inferior to the pipeline one. The possible reason might be that the simple one-round decoding is unable to leverage this advantage even though our MTL architecture for the end2end learning can effectively learn the interactions between the two tasks, .

Human Evaluation We perform a human evalua-

Model	Spatial	Fluency	Location	Avg
VL-T5(Base)	93.2	93.8	96.1	94.4
+VSRC-pipeline	93.9	94.0	96.3	94.7
+VSRC-end2end	94.9	**95.2**	**96.6**	95.6
+VSRC-golden	**99.3**	95.0	**96.6**	**97.0**

Table 3: Results of human evaluation.

tion to better compare the results of our proposed VSD models. We focus on models with VL-T5 backbone, and randomly sample 100 test instances of each model for evaluation. The VSD outputs of each model are scored with the following three measurements:

- **Spatial** Correctness: whether the spatial semantics of the generated text is consistent with the image?
- **Fluency**: whether the generated text is readable and not different from human sentence-making?
- **Location** Correctness: whether the input objects can be identified from the image according to the generated text?

Each question will be answered by a number from 0 to 1, indicating terrible to perfect. We let three annotators participate in a model-blind evaluation. Table 3 shows the accumulation scores of over the 100 instances with one decimal place retained. Noticeably, the Spatial Correctness is different from the VSRC accuracy in Table 2, where the former is for human judgement of VSD descriptions and the latter is for a nine-way classification. Generally, the VSRC accuracy can only evaluate one of multiple reasonable spatial relations of the given two objects while the human evaluation is more tolerant and reasonable. The tendency in performance is consistent with the automatic evaluation, where VSRC can help VSD because it can offer more spatial information, and the end-to-end model is better in utilizing automatic VSRC. The model with golden VSRC achieves a very high score of 99.3, which is reasonable due to the golden VSRC information of inputs.

7 Conclusion

In this work, we introduced a novel image-to-text generation task, namely VSD, aiming to generate text descriptions containing spatial semantics of two objects in an image, and constructed a dataset to benchmark this task. We adopted the models with Transformer-based encoder-decoder architectures (i.e., VL-BART and VL-T5) for our task to obtain the baseline results. Moreover, we proposed to integrate VSRC into our models by pipeline and end-to-end architectures, enhancing VSD with the support of spatial relations. The experimental results show that the VSRC-enhanced approach achieves significant progress over our initial models. Moreover, the end-to-end models outperform the pipeline ones due to joint learning.

Limitations

This work has two major limitations. The first limitation lies in our dataset. Our annotated dataset is built on SpatialSence and VG-Relation, aiming to study the relationship between VSRC and VSD. Under this setting, the variety of the spatial relations is limited to only nine. In addition, we only annotate one sentence for each instance, which limits the diversity of the description styles. We plan to continuously improve our dataset with more spatial relations and descriptions as the future work to improve this condition. The second limitation is that our base models only focus on single spatial relations in this work, ignoring the compound relations such as "left" and "behind" concurrently occurring. To solve this issue, we need to explore more methods to model multiple relations to generate descriptions with richer semantics. We also leave this aspect to future in-depth studies.

Ethical Considerations

We construct a new large-scale image-to-text generation dataset with crowd annotations. All the images of our dataset are sourced from two existing public datasets, SpatialSense and VisualGenome, which are open-access. All the annotators were voluntary participants and can quit at any time. They were informed of the study's goals before giving their express consent. All annotators were properly paid by their actual efforts and there is no information related to annotator privacy in the dataset.

Acknowledgement

This work is supported by grants from the National Natural Science Foundation of China (No. 62176180).

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, pages 4971–4980.

Khawaja Tehseen Ahmed, Sumaira Aslam, Humaira Afzal, Sajid Iqbal, Arif Mehmood, and Gyu Sang Choi. 2021. Symmetric image contents analysis and retrieval using decimation, pattern analysis, orientation, and features fusion. *IEEE Access*, 9:57215–57242.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In *ECCV*, volume 9909 of *Lecture Notes in Computer Science*, pages 382–398.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *ICCV*, pages 2425–2433.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of ACL*, pages 65–72.

Long Chen, Zhihong Jiang, Jun Xiao, and Wei Liu. 2021. Human-like controllable image captioning with verb-specific semantic roles. In *CVPR*, pages 16846–16856.

Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. 2020a. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *CVPR*, pages 9959–9968.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020b. UNITER: universal image-text representation learning. In *ECCV*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120.

Meng-Jiun Chiou, Roger Zimmermann, and Jiashi Feng. 2021. Visual relationship detection with visual-linguistic knowledge from multimodal representations. *IEEE Access*, 9:50441–50451.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *Proceedings of ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 1931–1942.

Guillem Collell, Thierry Deruyttere, and Marie-Francine Moens. 2021. Probing spatial clues: Canonical spatial templates for object relationship understanding. *IEEE Access*, 9:134298–134318.

Guillem Collell and Marie-Francine Moens. 2018. Learning representations specialized in spatial knowledge: Leveraging language and vision. *Trans. Assoc. Comput. Linguistics*, 6:133–144.

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2019. Show, control and tell: A framework for generating controllable and grounded captions. In *CVPR*, pages 8307–8316.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *CVPR*, pages 10575–10584.

Soham Dan, Hangfeng He, and Dan Roth. 2020. Understanding spatial relations through multiple modalities. In *LREC*, pages 2368–2372.

Chaorui Deng, Ning Ding, Mingkui Tan, and Qi Wu. 2020. Length-controllable image captioning. In *ECCV*, volume 12358 of *Lecture Notes in Computer Science*, pages 712–729.

Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. 2021. Similarity reasoning and filtration for image-text matching. In *AAAI*, pages 1218–1226.

Hao Fei, Yafeng Ren, Shengqiong Wu, Bobo Li, and Donghong Ji. 2021a. Latent target-opinion as prior for document-level sentiment classification: A variational approach from fine-grained perspective. In *WWW*, pages 553–564.

Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. 2021b. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021*, pages 549–559.

Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. 2022. Matching structure for dual learning. In *ICML*, pages 6373–6391.

Irtiza Hasan, Shengcai Liao, Jinpeng Li, Saad Ullah Akram, and Ling Shao. 2021. Generalizable pedestrian detection: The elephant in the room. In *CVPR*, pages 11328–11337.

Xiaodong He and Li Deng. 2017. Deep learning for image-to-text generation: A technical overview. *IEEE Signal Process. Mag.*, 34(6):109–116.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image captioning: Transforming objects into words. In *NeurIPS*, pages 11135–11145.

Ronghang Hu and Amanpreet Singh. 2021. Transformer is all you need: Multimodal multitask learning with a unified transformer. *CoRR*, abs/2102.10772.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709.

Muhammad Zubair Irshad, Niluthpol Chowdhury Mithun, Zachary Seymour, Han-Pang Chiu, Supun Samarasekera, and Rakesh Kumar. 2021. SASRA: semantically-aware spatio-temporal reasoning agent for vision-and-language navigation in continuous environments. *CoRR*, abs/2108.11945.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to describe differences between pairs of similar images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4024–4034. Association for Computational Linguistics.

Wei Ji, Xi Li, Lina Wei, Fei Wu, and Yueting Zhuang. 2020. Context-aware graph label propagation network for saliency detection. *IEEE Transactions on Image Processing*, 29:8177–8186.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pages 1988–1997.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137.

Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. 2019. Dense relational captioning: Triple-stream networks for relationship-based captioning. In *CVPR*, pages 6271–6280.

Geonuk Kim, Honggyu Jung, and Seong-Whan Lee. 2021. Spatial reasoning for few-shot object detection. *Pattern Recognit.*, 120:108118.

Parisa Kordjamshidi, Martijn van Otterlo, and Marie-Francine Moens. 2011. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Trans. Speech Lang. Process.*, 8(3):4:1–4:36.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pages 7871–7880.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021. UNIMO: towards unified-modal understanding and generation via cross-modal contrastive learning. In *ACL/IJCNLP*, pages 2592–2607.

Chin-Yew Lin. 2004. Rouge: a package for automatic evaluation of summaries. In *Workshop of ACL*, pages 74–81.

Annika Lindh, Robert J. Ross, and John D. Kelleher. 2020. Language-driven region pointer advancement for controllable image captioning. In *Proceedings of COLING*, pages 1922–1935.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pages 13–23.

Alexander Patrick Mathews, Lexing Xie, and Xuming He. 2018. Semstyle: Learning to generate stylised image captions using unaligned text. In *CVPR*, pages 8591–8600.

Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. Improving factual completeness and consistency of image-to-text radiology report generation. In *NAACL-HLT*, pages 5288–5304.

Eric Nichols and Fadi Botros. 2015. Sprl-cww: Spatial relation classification with independent multi-class models. In *SemEval@NAACL-HLT*, pages 895–901.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Régis Pierrard, Jean-Philippe Poli, and Céline Hudelot. 2021. Spatial relation learning for explainable image classification and annotation in critical applications. *Artif. Intell.*, 292:103434.

Jordi Pont-Tuset, Jasper R. R. Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *ECCV*, volume 12350 of *Lecture Notes in Computer Science*, pages 647–664.

James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. 2015. Semeval-2015 task 8: Spaceeval. In *SemEval@NAACL-HLT*, pages 884–894.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Sonia Raychaudhuri, Saim Wani, Shivansh Patel, Unnat Jain, and Angel X. Chang. 2021. Language-aligned waypoint (LAW) supervision for vision-and-language navigation in continuous environments. *CoRR*, abs/2109.15207.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *CVPR*, pages 1179–1195.

Jithmi Shashirangana, Heshan Padmasiri, Dulani Meedeniya, and Charith Perera. 2021. Automated license plate recognition: A survey on methods and techniques. *IEEE Access*, 9:11203–11225.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *ICCV*, pages 7463–7472.

Hao Tan and Mohit Bansal. 2019. LXMERT: learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*, pages 5099–5110.

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 *NeurIPS*, pages 5998–6008.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164.

Xinxiao Wu, Ruiqi Wang, Jingyi Hou, Hanxi Lin, and Jiebo Luo. 2021. Spatial-temporal relation reasoning for action prediction in videos. *Int. J. Comput. Vis.*, 129(5):1484–1505.

Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. 2022. Video as conditional graph hierarchy for multi-granular question answering. In *AAAI*, pages 2804–2812.

Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. 2021. Boundary proposal network for two-stage natural language video localization. In *AAAI*, pages 2986–2994.

Zhenbo Xu, Wei Yang, Ajin Meng, Nanxue Lu, Huan Huang, Changchun Ying, and Liusheng Huang. 2018. Towards end-to-end license plate detection and recognition: A large dataset and baseline. In *ECCV*, volume 11217 of *Lecture Notes in Computer Science*, pages 261–277.

Kaiyu Yang, Olga Russakovsky, and Jia Deng. 2019. Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. In *ICCV*, pages 2051–2060.

Zhen Zeng, Zheming Zhou, Zhiqiang Sui, and Odest Chadwicke Jenkins. 2018. Semantic robot programming for goal-directed manipulation in cluttered scenes. In *ICRA*, pages 7462–7469.

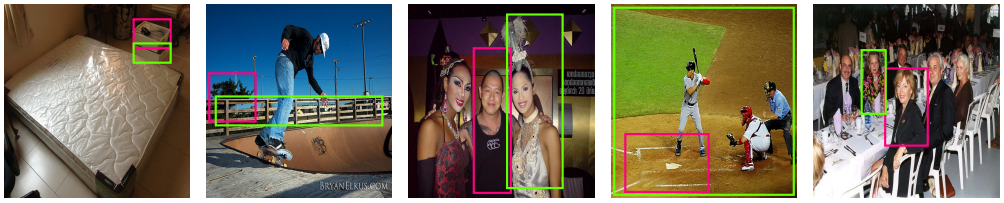
Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, pages 5579–5588.

Yue Zheng, Yali Li, and Shengjin Wang. 2019. Intention oriented image captions with guiding objects. In *CVPR*, pages 8395–8404.

Yiwu Zhong, Liwei Wang, Jianshu Chen, Dong Yu, and Yin Li. 2020. Comprehensive image captioning via scene graph decomposition. In *ECCV*, volume 12359 of *Lecture Notes in Computer Science*, pages 211–229.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and VQA. In *AAAI*, pages 13041–13049.

Jordan Zlatev. 2007. Spatial semantics. *The Oxford handbook of cognitive linguistics*, pages 318–350.



Input					
VL-T5(Base)	**A bin is above another bin.**	A tree is in front of a fence.	A man is next to a woman with red dresses.	A white line is on the dirt and several people.	Two women sit close to each other around a table.
VL-T5+VSRC-pipeline	**A bin is on another bin.**	**There is a tree behind a fence.**	A man is on the left of a woman dressed with a white hat.	Some white lines is in the dirt.	One women is sitting close to another woman.
VL-T5+VSRC-end2end	**A bin is put on another bin.**	**A tree grows behind a fence.**	**A man stands on the right of a woman.**	There are several lines above the dirt region.	One women dressed in black is sitting in front of another woman.
VL-T5+VSRC-golden	**A bin is put on another bin.**	**A tree grows behind a fence.**	**A man is standing on the right of a woman.**	**Some white lines are on the dirt region of a sports ground.**	One women dressed in black is sitting in front of another woman.
Human	**Two bins are stacked together, where one is put on the other.**	**In the distance, there is a tree behind a fence.**	**A man stands on the right of a woman who wears a white hat.**	**There are several white lines painted on the dirt infield of the baseball park.**	**One women dressed in black sits in front of another woman who is at the other side of a dining-table.**

Figure 7: Case studies, where the object in an image marked by the red box is the first object of VSD input, and the bold orange descriptions are regarded as relatively acceptable.

A Detailed Experiment Settings

We adopt the default settings of VL-BART and VL-T5 backbones (Cho et al., 2021). In VSRC, the dimension size of the bounding box coordinate features ($\boldsymbol{c}_{O_1}$ and $\boldsymbol{c}_{O_2}$ in Equation (6)) is 64 and the dimension of the fully connected layers is set to 1024. For hyper-parameters, we detail them in Appendix A. We train our models by using the AdamW optimizer (Loshchilov and Hutter, 2017), setting the initial learning rate to $5e^{-4}$ and weight decay to 0.01. We apply the gradient clipping mechanism by a maximum value of 5.0 to avoid gradient explosion. The training batch size is 16 and the max epoch number is 40.

B Case Study

We show five case studies in Figure 7 to extensively understand the model outputs. In the first case, all four models (human is the golden answer) are able to output good descriptions because the relation in the image is simple and easy to understand. In the second case, the VL-T5 (base) model is unable to provide a correct answer, while the other models are all correct due to the benefit from VSRC. In the third case, the VL-T5+VSRC-end2end and VL-T5+VSRC-golden models output acceptable results, while the other two models fail. The reason might be that the two models can identify the correct or more important spatial relation between the two objects. In the fourth case, we can only obtain a correct description by VL-T5+VSRC-golden because the spatial relation is very difficult to recognize by automatic VSRC. Finally, our VSRC-enhanced VSD model fails in the fifth case even with the golden spatial relation. The reason might be the extreme complexity of this particular input image.