Towards Unifying Reference Expression Generation and Comprehension

Duo Zheng[1]*, Tao Kong[2], Ya Jing[2], Jiaan Wang[3], Xiaojie Wang[1]
[1]Beijing University of Posts and Telecommunications
[2]ByteDance AI Lab, Beijing, China
[3]Soochow University, Suzhou, China
{zd, xjwang}@bupt.edu.cn
{kongtao,jingya}@bytedance.com, jawang1@stu.suda.edu.cn

Abstract

Reference Expression Generation (REG) and Comprehension (REC) are two highly correlated tasks. Modeling REG and REC simultaneously for utilizing the relation between them is a promising way to improve both. However, the problem of distinct inputs, as well as building connections between them in a single model, brings challenges to the design and training of the joint model. To address the problems, we propose a unified model for REG and REC, named UniRef. It unifies these two tasks with the carefully-designed Image-Region-Text Fusion layer (IRTF), which fuses the image, region and text via the *image cross-attention* and *region cross-attention*. Additionally, IRTF could generate pseudo input regions for the REC task to enable a uniform way for sharing the identical representation space across the REC and REG. We further propose Vision-conditioned Masked Language Modeling (VMLM) and Text-Conditioned Region Prediction (TRP) to pre-train UniRef model on multi-granular corpora. The VMLM and TRP are directly related to REG and REC, respectively, but could help each other. We conduct extensive experiments on three benchmark datasets, RefCOCO, RefCOCO+ and RefCOCOg. Experimental results show that our model outperforms previous state-of-the-art methods on both REG and REC.

1 Introduction

Reference Expression (RE), which describes an unambiguous object in a real scene, is a significant cognitive behaviour in human society. People conceive a RE for an object and recognize a referent according to a RE in daily life, which we name Reference Expression Generation (REG) and Comprehension (REC), respectively. Both tasks have attracted surging interest (Rohrbach et al., 2015; Deng et al., 2018; Yu et al., 2018; Yang et al., 2019; Kamath et al., 2021) from Natural Language Processing (NLP), Computer Vision (CV) and Human-Computer Interaction (HCI), due to their broad research prospects and actual applications.

REG and REC are the two sides to the same coin and are dependent on each other. For example, before conceiving an unambiguous description, people need to correctly locate the object according to the description in their mind. However, there is less focus on addressing the unified modeling for both REG and REC. One line of the work lies in Bayes' modeling. Mao et al. (2016) first propose a method that can generate a RE grounded on an image, and which can also locate the object described by the RE via Bayes' rule. The subsequent work (Yu et al., 2016, 2017; Luo and Shakhnarovich, 2017; Tanaka et al., 2019a; Kim et al., 2020; Liu et al., 2020) typically follows this paradigm. Another line of the work studies the parameter-shared model for the two tasks. Sun et al. (2022) propose the first parameter-shared framework PFOS. Considering the inputs for the two tasks are distinct (images and regions for REG while images and text for REC), PFOS shares the language-guide-vision module with the object-guide-context module, and the vision-guide-language module with the context-guide-object module. These modules need to handle the object and language inputs in REC and REG respectively, ignoring the modality gap between the inputs. To better share knowledge across REG and REC, we argue that it is important to coordinate the difference between their inputs for a unified modeling.

Therefore, in this paper, we propose UniRef, a unified model for REG and REC. To alleviate the issue of distinct inputs, we design the Image-Region-Text Fusion layer (IRTF), which extends the transformer encoder layer through adding the *image cross-attention* and *region cross-attention*. Specifically, the image and region information is fused by the *image cross-attention* and *region cross-*

*Work was done when Zheng was interning at ByteDance AI Lab, Beijing, China.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 6598 - 6611
December 7-11, 2022

attention, respectively. In REC, since the input region is not given, a region predictor is used to produce a region prediction as the input for the *region cross-attention*. In this manner, UniRef could share the identical representation space across different tasks. Furthermore, our UniRef is pre-trained with two objectives, Vison-conditioned Masked Language Modeling (VMLM) and Text-Conditioned Region Prediction (TRP) on corpora of different granularities ranging from object labels to object phrases, from region descriptions to RE.

We note that the emergence of Vision-Langue Pre-training (VLP) (Lu et al., 2019; Tan and Bansal, 2019; Zhou et al., 2020; Yu et al., 2020; Su et al., 2020; Cho et al., 2021; Kim et al., 2021; Wang et al., 20b; Radford et al., 2021; Huang et al., 2021) has greatly promoted the development of multi-modal tasks. And some of them (Li et al., 2020; Chen et al., 2020; Zeng et al., 2021) have significantly boosted the performance of REC and demonstrated tremendous generalization ability. Most of them focus on the alignment between either images and captions, or regions and region descriptions. To our knowledge, there is no VLP study focusing on unified modeling for both REG and REC.

To verify the effectiveness of our UniRef, we conduct extensive experiments on three benchmark datasets, RefCOCO, RefCOCO+ (Yu et al., 2016) and RefCOCOg (Mao et al., 2016) datasets. Experimental results deliver that our UniRef outperforms previous SOTA methods on REG and REC. In addition, we conduct case studies to investigate the abilities learned by our model and the challenges still remained.

Our main contributions are concluded as follows[1]:

- We propose a unified model for REG and REC, named UniRef. To alleviate the issue of distinct inputs, we design the Image-Region-Text Fusion layer (IRTF), which helps the model to share knowledge across REG and REC.
- We pre-train UniRef with two objectives, Vision-conditioned Masked Language Modeling (VMLM) and Text-Conditioned Region Prediction (TRP), to learn the abilities required by REG and REC, respectively.
- Experimental results show that our unified model UniRef surpasses previous SOTA models on both REG and REC.

[1]We release the code and model at: https://github.com/zd11024/UniRef.

2 Method

We first briefly review the task definitions of REG and REC in § 2.1. Then we introduce the architecture of our UniRef and the pre-training in § 2.2 and § 2.3, respectively. Last, we describe the fine-tuning and inference in § 2.4.

2.1 Task Definitions

Reference Expression Generation. Given an image I and a region R described by box coordinates, the REG model generates the corresponding RE text $T = \{t_1, \cdots, t_{L_T}\}$ with L_T tokens. The conditional distribution could be formalized as:

$$p_{\theta_{\mathrm{G}}}(T|I,R) = \prod_{i=1}^{L_T} p(t_i|I,R,t_{1:i-1}), \quad (1)$$

where $t_{1:i-1}$ is the previous generated tokens and θ_{G} are parameters of the REG model.

Reference Expression Comprehension. The REC model predicts the region R with an image I and the corresponding RE text T as the input, which could be denoted as $p_{\theta_{\mathrm{C}}}(R|I,T)$. θ_{C} are parameters of the REC model.

2.2 Architecture

As depicted in Fig. 1, UniRef consists of a vision encoder, a language encoder and a fusion encoder as well as two task-specific head, i.e., a language model (LM) head and a box head.

Vision Encoder. Given an image I, the vision encoder extracts the image features. It is based on the Vision Transformer (ViT) (Li et al., 2020) and initialized with the weights of CLIP-ViT (Radford et al., 2021). It first splits the image into non-overlapping patches, and then projects them into a sequence of embeddings. After that, these embeddings are fed to stacked transformer encoder blocks and interact with each other through self-attention, resulting in the image features $V^I = \{v_1, \cdots, v_{L_I}\}$, where L_I is the number of patches.

In REC, given a region R from I, we obtain the region features $V^R = \{v_{p_1}, \cdots, v_{p_{L_R}}\}$, where $\{p_i\}$ and L_R are the indexes and the number of patches that overlaps with the region, respectively.

Language Encoder. The language encoder is based on BERT (Devlin et al., 2019). The input sentence is tokenized into WordPieces (Wu et al., 20), which are subsequently transformed into the text features $Z = \{z_{\texttt{[cls]}}, z_1, \cdots, z_{L_T}\}$ by the text encoder, where L_T is the number of tokens

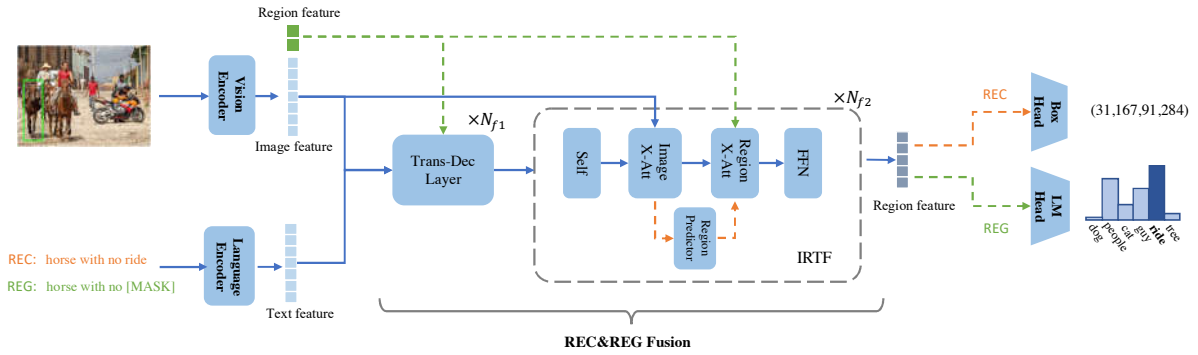


Figure 1: The architecture of UniRef. The orange and green dashed lines indicate the specific design for REC and REG respectively to enable identical representation space. "Tran-Dec Layer" mean the transformer decoder layer. "Self" means the self-attention module while "X-Att" represents the *cross-attention* module. N_{f1}, N_{f2} mean the number of layers of transformer decoder block and IRTF, respectively. $\{p_i\}$ are the indexes of patches that overlap with the input bounding box.

and $z_{\texttt{[cls]}}$ are the text features corresponding to the special token `[cls]`.

Fusion Encoder. The fusion encoder extends the transformer decoder by replacing last N_{f2} vanilla transformer decoder layers with Image-Region-Text-Fusion layers (c.f., Fig. 1), which are designed to bridge the gap between REG and REC.

The vanilla transformer decoder layer fuses region or image information via *cross-attention*, depending on the input requirement of the task.

IRTF extends the vanilla transformer encoder layer through adding the *image cross-attention* and *region cross-attention*, and fuses the image information and region information with queries. Given the input $X = \{x_{\texttt{[CLS]}}, x_1, \cdots, x_{L_T}\}$, self-attention is first applied to obtain the queries:

$$X^Q = \text{MHA}(X, X, X) + X, \tag{2}$$

where MHA is multi-head attention.

Then the *image cross-attention* and the *region cross-attention* are performed successively as follows:

$$Z^I = \text{MHA}(X^Q, V^I, V^I), \tag{3}$$
$$X^I = \text{GLU}([Z^I, X^Q]) + X^Q, \tag{4}$$
$$Z^R = \text{MHA}(X^I, V^R, V^R), \tag{5}$$
$$X^R = \text{GLU}([Z^R, X^Q]) + X^I, \tag{6}$$

where Z^I, Z^R are the intermediate representations after multi-head attention. X^I, X^R are the outputs of the *image cross-attention* and the *region cross-attention*, respectively. $[\cdot]$ means the concatenation of vectors. Following Huang et al. (2019), we adopt Gated Linear Unit (GLU) to refine the attention outputs, denoted as:

$$\text{GLU}(X) = \sigma(XW^1) \odot XW^2, \tag{7}$$

where W^1, W^2 are learnable parameters, $\sigma(\cdot)$ is the sigmoid function and $\odot$ means the element-wise multiplication.

Lastly, X^R is fed to a feed-forward network to obtain the output hidden states.

When performing REC, the region input is not available, which requires to predict the region conditioned on the image and text. To make the input of REC identical with REG, a region predictor is utilized for producing a region prediction, as the input for the *region cross-attention*. In detail, for each patch, it calculates a score α_i based on X^I_{cls} and the position embedding of i-th patch e_i. Then, it selects all patches whose scores exceed the threshold δ, constituting the predicted region presentations V^R. We formalize this procedure as:

$$\alpha_i = \text{MLP}([X^I_{cls}, e_i]), \tag{8}$$
$$V^R = \{V^I_i | \alpha_i \geq \delta\}. \tag{9}$$

LM Head&Box Head. To carry out REG, we use a LM head to predict the next token given the last hidden state of the `[MASK]` token. During performing REC, we employ a box head to regress the bounding box b conditioned on the last hidden state of the `[CLS]` token.

2.3 Pre-training

2.3.1 Pre-training Objectives

To learn the abilities of language modeling and visual grounding, we pre-train UniRef with two

objectives, Vision-conditioned Masked Language Modeling (VMLM) and Text-Conditioned Region Prediction (TRP), which are corresponding to REG and REC, respectively.

Vision-Conditioned Masked Language Modeling. Given an image-region-text triplet (I, R, T), we follow X-VLM (Zeng et al., 2021) to mask 25% of tokens in text sequences. The task aims to predict the unseen tokens based on the visible text, region and image. Note that VMLM is similar to REG, but with differences of decoding order and attention masks. The loss function is defined as:

$$\mathcal{L}_{\text{VMLM}} = -\mathbb{E}_{(I,R,T)} \log p_{\theta_{\text{G}}}(\hat{T}|I, R, \tilde{T}), \quad (10)$$

where $\hat{T}$ and $\tilde{T}$ represent the masked and unmasked tokens, respectively.

Text-Conditioned Region Prediction. Given an image-text pair (I, T), the goal of TRP is to predict the bounding box of the region or object described by the text. The loss is the summation of the generalized Intersection over Union (gIoU) (Rezatofighi et al., 2019) and the l_1 distance:

$$\mathcal{L}_{\text{bbox}} = \mathbb{E}_{(I,T)}\, \mathcal{L}_{\text{gIoU}}(\hat{b}, b) + \|\hat{b} - b\|_1, \quad (11)$$

where $\hat{b}, b$ represent the bounding boxes of the ground truth and prediction, respectively.

In TRP, each IRTF produces a region prediction as the input for the *region cross-attention*. The supervised signal comes from the patch-level binary cross-entropy between the prediction and the ground truth, formulated as:

$$\mathcal{L}_{\text{pred}} = \mathbb{E}_{(I,T)} \sum_i H(\hat{m}, m_i), \quad (12)$$

where $\hat{m}, m_i$ mean the region mask of the ground truth and the region mask predicted by the i-th IRTF, respectively.

The final loss for TRP is summed by:

$$\mathcal{L}_{TRP} = \mathcal{L}_{\text{bbox}} + \mathcal{L}_{\text{pred}}. \quad (13)$$

2.3.2 Perspective from Probability

We explain how our UniRef share the identical representation space across tasks in pre-training from a probability perspective. We factorize the objectives of VMLM and TRP as follows:

$$p_{\theta_{\text{G}}}(\hat{T}|I, R, \tilde{T}) = p_{\theta_{\text{F}}}(H|I, R, \tilde{T})\, p_{\theta_{\text{LM}}}(\hat{T}|H), \quad (14)$$

Pre-training Dataset	# Images	# Text	Avg. Tok
COCO Object Labels	112k	434k	1.20
VG Phrases	104k	2M	1.24
VG Region Descriptions	105k	360k	5.40
RefCOCO-MERGE	24k	287k	5.07

Table 1: The statistics of the pre-training datasets. "# Images "and "# Text" represent the number of images and text descriptions, "Avg. Tok" indicates the average number of tokens in descriptions.

$$p_{\theta_{\text{C}}}(R|I, T) = p_{\theta_{\text{F}}}(H|I, R^{'}, T)\, p_{\theta_{\text{Box}}}(R|H) \quad p_{\theta_{\text{P}}}(R^{'}|I, T), \quad (15)$$

where $\theta_{\text{LM}}, \theta_{\text{Box}}, \theta_{\text{F}}, \theta_{\text{P}}$ mean the parameters of the LM head, box head, fusion encoder and predictor, respectively. H are the last hidden states. With the help of the predictor, both VMLM and TRP aim to align the region with text ($(R, \hat{T})$ and $(R^{'}, T)$) conditioned on the image I.

2.3.3 Pre-training Datasets

We collect four pre-training datasets of different granularities ranging from object labels to pharases, from region descriptions to RE: (1) COCO object labels (Lin et al., 2014). Each object corresponds to a label in 80 pre-defined categories. (2) Visual Genome (VG) phrases (Krishna et al., 2017). We concatenate the attribute and object of an object to form a phrase. There are over 75k unique objects and 50k unique attributes, leading to more combinations of objects and attributes. (3) Visual Genome region descriptions. The region descriptions could be either a phrase or a sentence. (4) RefCOCO-MERGE. We merge RefCOCO, RefCOCO+ and RefCOCOg together. For the above datasets, we filter out the data whose image appears in the val and test set of RefCOCO, RefCOCO+, RefCOCOg according to COCO id. Tab. 1 lists the statistics of the pre-training datasets.

2.4 Fine-tuning and Inference.

Following Li et al. (2020), we fine-tune UniRef for REG on RefCOCO, RefCOCO+ and RefCOCOg separately. In detail, 25% of the tokens are randomly masked and the model recovers them with a unidirectional attention mask instead of a bidirectional one. During inference, at each step, a `[MASK]` token is appended to the end of current generation, with a subsequent forward-pass to generate the next token. The process terminates until a `[SEP]` token is produced. For REC, the procedure is same to TRP.

Method	# Pre-train Images	RefCOCO				RefCOCO+				RefCOCOg			
		testA		testB		testA		testB		val		test	
		M	C	M	C	M	C	M	C	M	C	M	C
SR (2019b)	-	0.301	0.866	0.341	1.389	0.243	0.672	0.222	0.831	0.160	0.741	0.160	0.727
SR-rerank (2019b)	-	0.310	0.842	0.348	1.356	0.241	0.656	0.219	0.782	0.165	0.756	0.164	0.764
CoNAN (2020)	-	0.330	0.915	0.354	1.410	0.288	0.761	0.250	0.876	-	-	-	-
VL-T5 (2021)	180k	0.334	0.978	0.347	1.427	0.288	0.828	0.245	0.852	0.189	0.873	0.189	0.881
UniRef	180k	**0.347**	**1.049**	**0.374**	**1.549**	**0.311**	**0.916**	**0.266**	**0.972**	**0.197**	**1.033**	**0.195**	**1.017**

Table 2: The performance on REG. "M" and "C" indicate Meteor and CIDEr, respectively. "-" means that the details are not reported. "# Pre-train Images" means the number of images in pre-training datasets.

Method	# Params	# Pre-train Images	RefCOCO		RefCOCO+		RefCOCOg	
			testA	testB	testA	testB	val	test
MattNet (2018)	-	-	81.14	69.99	71.62	56.02	66.58	67.27
ViLBERT (2019)	-	3.3M	-	-	78.52	62.61	-	-
VL-BERT$_{\text{large}}$ (2020)	-	3.3M	-	-	78.57	62.30	-	-
UNITER$_{\text{large}}$ (2020)	300M	3.3M	-	-	78.57	62.30	-	-
MDETR (2021)	-	200k	90.42	83.06	85.05	71.88	83.44	83.93
X-VLM (2021)	240M	4M	-	-	86.36	71.00	-	-
OFA (2022)	180M	14.7M	90.67	83.30	87.15	74.29	82.29	82.31
UniRef	227M	180k	**91.21**	**83.87**	**87.74**	**75.45**	**85.62**	**84.92**

Table 3: The accuracy (%) on REC. "-" means that the details are not reported. "# Pre-train Images" means the number of images in pre-training datasets. The comparing models are base-size unless otherwise specified.

3 Experiments

3.1 Datasets and Metrics

Datasets. We evaluate our model on three widely-used benchmark datasets, i.e., RefCOCO, RefCOCO+ (Yu et al., 2016) and RefCOCOg (Mao et al., 2016), which are based on COCO (Lin et al., 2014) images.

RefCOCO contains 142,209 reference expressions for 50,000 objects on 19,994 images, while RefCOCO+ consists of 141,564 descriptions for 50,000 objects on 19,992 images. Their test sets are split into testA and testB by "People vs. Object". The main difference is that position words are prohibited in RefCOCO+, leading to more appearance-centric descriptions.

ReCOCOg contains 54,822 objects on 26,711 images with 104,560 expressions, which are longer and more informative than that of RefCOCO/RefCOCO+. For RefCOCOg, most methods evaluate on Google split in REG, and on UMD split in REC. In this paper, we reproduce some representative REG methods on UMD split and report the corresponding results.

Metrics. We evaluate the performance of REG with two automatic metrics, i.e., CIDEr (Vedantam et al., 2015) and Meteor (Lavie and Denkowski, 2009). In REC, we report the accuracy of bounding box prediction. A prediction is correct if its IoU with the ground truth is greater than 0.5.

3.2 Implementation Details

The vision encoder of UniRef is initialized with weights of CLIP-ViT/B-16[2]. The text and fusion encoder is initialized with weights of the first six and last six layers of BERT$_{\text{base}}$, respectively. The extra parameters of the fusion encoder, including the *cross-attention* and predictor, are randomly initialized. For the fusion encoder, we adopt vanilla transformer decoder layers as the first five layers and IRTF as the last layer.

We implement our method with Pytorch and perform all experiments on NVIDIA Tesla A100 GPU. We pre-train UniRef for 200k steps with a batch size of 1024. The learning rate is warmed-up from 1e-5 to 1e-4, with a subsequent decay to 1e-5. In the fine-tuning stage, we train REG and REC models for 20 epochs with a batch size of 40. Following Zeng et al. (2021), the image resolution is set to 224 in pre-training while 384 in fine-tuning.

3.3 Comparing Models

In this section, we compare UniRef with the SOTA models of REG and REC, respectively.

[2] https://huggingface.co/openai/clip-vit-base-patch16

#		REG							REC						
		Avg.	RefCOCO		RefCOCO+		RefCOCOg		Avg.	RefCOCO		RefCOCO+		RefCOCOg	
			testA	testB	testA	testB	val	test		testA	testB	testA	testB	val	test
1	UniRef (no IRTF)	1.075	1.041	1.513	0.895	0.977	1.011	1.010	83.99	90.29	83.74	86.38	75.55	83.84	84.11
2	UniRef (IRTF in L4,5,6)	1.088	1.063	1.540	0.910	0.988	1.015	1.012	84.44	90.79	84.07	86.74	74.45	85.27	85.30
3	UniRef (IRTF in L5,6)	1.083	1.031	1.505	0.912	0.981	1.037	1.033	<u>84.68</u>	91.04	84.44	87.08	75.53	85.01	84.95
4	UniRef (IRTF in L6)	<u>1.089</u>	1.049	1.549	0.916	0.972	1.033	1.017	**84.80**	91.21	83.87	87.74	75.45	85.62	84.92
5	*w/o.* GLU	1.080	1.054	1.511	0.899	0.985	1.015	1.014	84.65	90.93	84.81	86.78	75.72	85.44	84.20
6	*w/o.* VMLM	0.760	0.818	1.183	0.645	0.738	0.591	0.585	82.46	89.50	82.99	84.51	72.51	82.68	82.56
7	*w/o.* TRP	1.060	1.025	1.492	0.889	0.962	1.003	0.989	61.39	75.06	63.96	65.81	48.98	57.84	56.67
8	*w/o.* RefCOCO-MERGE	**1.098**	1.063	1.540	0.910	0.988	1.043	1.043	82.31	89.52	82.68	84.23	71.01	83.35	83.07

Table 4: The ablation studies of fusion encoder and pre-training. We report CIDEr for REG and accuracy for REC. Avg. means the average of CIDEr/accuracy on REG/REC. "UniRef (IRTF in LX)" means that layers X are IRTF while others are transformer decoder layers, and "UniRef (no IRTF)" indicates the fusion encoder only contains transformer decoder layers. The **bold** and <u>underline</u> denote the best and the second performances, respectively.

REG. (1) SR (Tanaka et al., 2019b) extends the speaker-listener-reinforcer framework (Yu et al., 2017) with a well-designed attention mechanism. (2) SR-rerank picks the expression through reranking a set of generated sentences. (3) CoNAN (Kim et al., 2020) introduces an attentional ranking module to obtain complementary neighbor features. (4) VL-T5 (Cho et al., 2021) unifies many tasks into a sequence-to-sequence framework via instruction learning. To adapt VL-T5 to REG, we append the region features at the fixed position of the input.

REC. (1) MattNet (Yu et al., 2018) is a representative two-stage method. (2) ViLBERT (Lu et al., 2019), (3) VL-BERT$_{\text{large}}$ (Su et al., 2020) and (4) UNITER$_{\text{large}}$ (Chen et al., 2020) are VLP models with region features. (5) MDETR (Kamath et al., 2021) is a pre-trained model that takes DETR (Carion et al., 2020) as the backbone. Additionally, (6) X-VLM (Zeng et al., 2021) and (7) OFA (Wang et al., 2022) are pre-trained on much larger datasets and show marvelous generalization ability. Note that X-VLM and OFA also utilize fine-grained labeled data, thus the comparison is fair.

3.4 Main Results

In REG and REC, our UniRef delivers better results than previous SOTA results, which cannot be simultaneously achieved by previous methods.

Performance on REG. As shown in Tab. 2, our UniRef outperforms previous SOTA methods on three datasets. Specifically, UniRef achieves 1.049/1.549 on RefCOCO testA/testB, 0.916/0.972 on RefCOCO+ testA/testB, and 1.033/1.017 on RefCOCOg val/test, in terms of CIDEr. Furthermore, it has the most prominent improvement on RefCOCOg, with CIDEr lift rate of 18.3% and 15.4% on val and test respectively, compared with VL-T5. This demonstrates that our model can better handle the expression with more details.

Performance on REC. As shown in Tab. 3, our UniRef outperforms SOTA models on all benchmark datasets. Specifically, it outperforms MDETR by 0.79/0.81% on RefCOCO, 2.69/3.57% on RefCOCO+ and 2.18/0.99% on RefCOCOg. Even compared to OFA pre-trained on 14.7M images, our model still shows its superiority, especially on RefCOCOg.

3.5 Ablation Study.

To investigate the effects of the fusion encoder and pre-training, we conduct ablation studies (Tab. 4).

ITRF Boosts the Results on REG and REC. Comparing rows 1 and 4, it can be seen that the UniRef with IRTF in 6-th layer outperforms the counterpart without IRTF, validating the effectiveness of IRTF. IRTF decouples the *cross-attention* into image and *region cross-attention*, and takes image, region and text as the identical inputs, resulting in better interaction between them. Furthermore, GLU slightly boost the performance for it could refine the attention outputs via non-linear transformation (row 4 vs. row 5).

UniRef with IRTF in 6-th Layer Outperforms Other Counterparts. Comparing rows 2, 3 and 4, UniRef with IRTF in 6-th layer achieves the best performance. With the increase of the number of IRTF, REC performance shows a downward trend, possibly due to the error accumulation of predicted regions generated by IRTF.

VMLM and TRP Benefit the Pre-training. Comparing rows 4, 6 and 7, our model outperforms the variant removing either pre-training task. The performance of REG/REC noticeably drops without

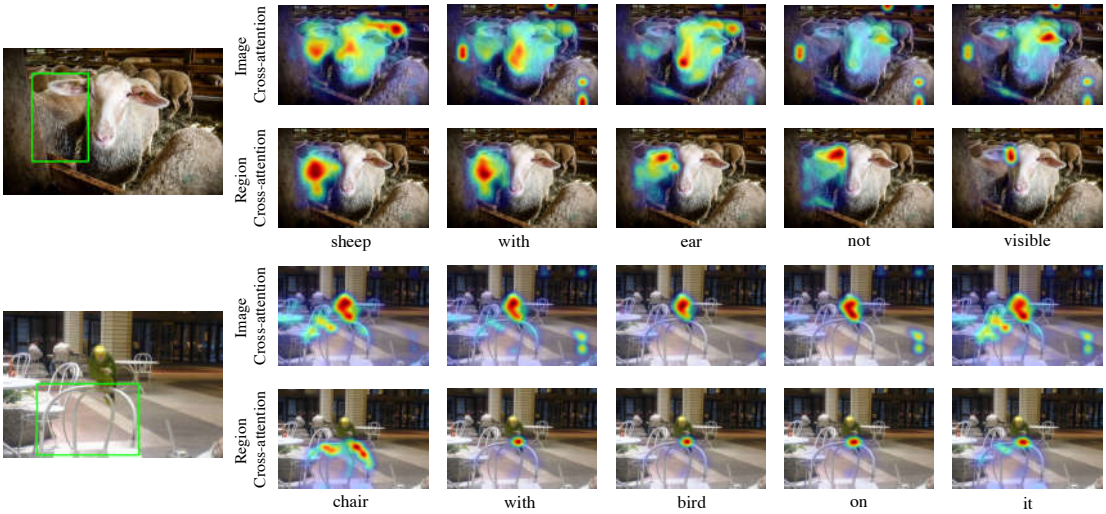


Figure 2: The visualization of the autoregressive generation process of REG. The model successively pays attention to the whole image and region, with a subsequent generated token. For each example, the first and second row respectively give the attention maps of *image cross-attention* and the *region cross-attention* when generating the next token, which is below the corresponding *cross-attention* maps. The input region is marked in the green box.

VMLM/TRP, illustrating the effectiveness of the pre-training tasks.

Pre-training on In-domain Data Significantly Improves REC but Slightly Damages REG. Furthermore, with pre-training on refCOCO-MERGE, UniRef suffers a significant increase in REC, from 82.31% to 84.72% on the average accuracy (row 8 vs. row 4). However, the average CIDEr slightly decreases in REG. We speculate that it is caused by the unbalanced sampling on the collected pre-training datasets, leading to overfitting to RefCOCO-MERGE.

3.6 Case Study.

In this section, we conduct case studies to provide a deeper understanding for UniRef. More examples are given in Appendix A.

How UniRef Utilizes Image and Region Information in REG? As shown in Fig. 2, we give visualization on the *cross-attention* maps, including *image cross-attention* and *region cross-attention*, across the process of autoregressive generation. Through observing cases, we discover two phenomena: 1) The *image cross-attention* could pay attention to other objects in the image that are indistinguishable from the target object, thereby assisting the model to generate more discriminative descriptions. For example, in the first instance, the ears of sheep are attended by *image cross-attention* while the sheep with ear not visible is attended by the *region cross-attention*, resulting in the description "sheep with ear not visible". 2) Through attending to the object related to the target object, the model could generate descriptions with relationships, e.g., spatial relationships. In the second example, the model unambiguously describes the chair in green box by the spatial relationship between it and the bird, which is not in green box.

The Ability that UniRef Learns in REC. We give examples of bounding box predictions in Fig. 3. UniRef is able to handle descriptions with various properties, e.g., comparisons (Fig. 3 (a)), attribute recognition (Fig. 3 (b),(c)), spatial relationships (Fig. 3 (j),(k)) and counting (Fig. 3 (d)-(f)).

The Challenges still Remain in REC. By analysing bad cases, we conclude some difficulties faced by our model: (1) Short path. The model correctly localizes the plant (Fig. 3 (m)) while fails to ground to the flowerpot (Fig. 3 (n)). It first locates the flowers on the wall, and then regards this wall as flowerpot. It shows that the model does not really understand what is flowerpot, but learns short paths through flowers; (2) Small objects. We discover that the model is not very good for small objects (Fig. 3 (i) and (r)).

4 Related Work

Reference Expression (RE). To study the RE, many datasets have been introduced, including RefCOCO (Yu et al., 2016), RefCOCO+ (Yu et al., 2016) and RefCOCOg (Mao et al., 2016). The first two are collected in a two-palyer cooperative game, namely ReferIt (Kazemzadeh et al., 2014), while the last one is annotated in a non-interactive setting.

Figure 3: Examples of the predicted bounding box in REC. The green and orange boxes indicate the ground truth and predicted boxes, respectively. The images are from RefCOCO+ while the texts are constructed.

The early work focuses on the CNN-LSTM framework, which could be applied to REG, as well as REC via Bayes' rule. Specifically, it first models $P(T|I,R)$, then obtains $P(R|I,T)$ by Bayes' rule, where I, R, T represent the image, the reigon and the text, repectively. Mao et al. (2016) first introduce this approach and propose a maximum mutual information method, which penalizes the likelihood of the RE to wrong objects in an image. Following this method, Yu et al. (2016) propose a visual comparative method, VisDiff, which uses the image, target object and visual difference information for generating unambiguous descriptions. Further, Yu et al. (2017) extend VisDiff to a speaker-listener-reinforcer model, in which the speaker, listener and reinforcer interact with each other.

Thanks to the success of object detection, REC attracts more attention and many endeavors have been devoted to it, ranging from two-stage to one-stage approaches. The two-stage methods (Yu et al., 2018; Deng et al., 2018; Wang et al., 2019) first extract region proposals with a object detector such as faster-RCNN (Ren et al., 2015), then select a region conditioned on the input text. In contrast, the one-stage methods (Yang et al., 2019, 2020; Li and Sigal, 2021) directly predict the bounding box given the image and the text, obtaining improvement of performance from end-to-end training.

Vision-Language Pre-training (VLP). VLP, motivated by the pre-trained language models in NLP, aims at learning generic representations from abundant image-text data, advancing many vision-language tasks, e.g., VQA (Antol et al., 2015), image captioning and visual dialog (de Vries et al., 2017; Das et al., 2017). ViLBERT (Lu et al., 2019) pioneers the adaption of pre-trained models for this field. Then, VL-BERT (Su et al., 2020) and LXMERT (Tan and Bansal, 2019) use a two-stream architecture for fusing information from different modality. Subsequently, Li et al. (2020) propose OSCAR, which takes object labels as anchors for aligning objects and text. More recently, Zeng et al. (2021) adopt vision transformers to extract visual features and design the task of region prediction to model the fine-grained alignment between regions and descriptions.

Moreover, various technologies are applied in VLP, ranging from contrastive learning (Li et al., 2021b; Radford et al., 2021) to knowledge distillation (Li et al., 2021a), from stage-wise pre-training (Liu et al., 2021; Wang et al., 20a) to prompt learning (Tsimpoukelli et al., 2021; Wang et al., 2022; Jin et al., 2022). Standing on the shoulders of giants, we step forward with the purpose of building more advanced models for REG and REC.

5 Conclusions

In this paper, we propose a unified model for reference expression generation and comprehension, named UniRef. To alleviate the issue of distinct inputs for the tasks, we design the Image-Region-Text Fusion layer (IRTF) to handle the difference between the distinct inputs. In addition, UniRef is pre-trained with two objectives, Vision-conditioned Masked Language Modeling (VMLM) and Text-Conditioned Region Prediction (TRP), on multi-granular corpora. Experimental results show that our UniRef outperforms previous state-of-the-art

methods on both REG and REC.

Ethical Considerations

In this section, we consider potential ethical issues of our model. In this paper, we propose UniRef, whose vision encoder and language encoder are initialized with the weights of CLIP-ViT (Radford et al., 2021) and BERT (Devlin et al., 2019), respectively. The pre-training datasets are collected from COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017), RefCOCO (Yu et al., 2016), RefCOCO+ (Yu et al., 2016) and RefCOCOg (Mao et al., 2016). Therefore, UniRef might involve the same biases and toxic behaviors exhibited by the pre-trained models and pre-training datasets.

Limitations

Our work has several limitations that can be further explored. (1) The size of the model and pre-training datasets could be scaled up. Since our model is designed for REG and REC, it requires carefully modification for the model architecture to adapt to massive image-text pairs. (2) We do not perform any optimization approaches for the REG model, such as self-critical sequence training and reinforcement learning. These approaches are proved to be beneficial in previous work (Yu et al., 2017; Huang et al., 2019; Cornia et al., 2020). (3) It is feasible to adapt our model to other related downstream tasks, e.g., phrase grounding (Plummer et al., 2015), reference expression segmentation (Wu et al., 2020) and dense captioning (Johnson et al., 2016), through elaborating task-specific designs. (4) It is worth more exploration on multi-task fine-tuning with REG and REC. We have done experiments that jointly fine-tune one model for both REG and REC. The performance on REG and REC is on par with or slightly worse than the separated UniRef.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. *european conference on computer vision*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Learning universal image-text representations. *european conference on computer vision*.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 1931–1942. PMLR.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10575–10584. IEEE.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1080–1089. IEEE Computer Society.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4466–4475. IEEE Computer Society.

Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. 2018. Visual grounding via accumulated attention. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7746–7755. IEEE Computer Society.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lun Huang, Wenmin Wang, Jie Chen, and Xiaoyong Wei. 2019. Attention on attention for image captioning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4633–4642. IEEE.

Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. Seeing out of the box: End-to-end pre-training for vision-language representation learning. *computer vision and pattern recognition*.

Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. 2022. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2763–2775, Dublin, Ireland. Association for Computational Linguistics.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4565–4574. IEEE Computer Society.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. 2021. Mdetr – modulated detection for end-to-end multi-modal understanding. *international conference on computer vision*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Jungjun Kim, Hanbin Ko, and Jialin Wu. 2020. CoNAN: A complementary neighboring-based attention network for referring expression generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1952–1962, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

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. *International Journal of Computer Vision*.

Alon Lavie and Michael Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine Translation*.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven C. H. Hoi. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. *neural information processing systems*.

Muchen Li and Leonid Sigal. 2021. Referring transformer: A one-step approach to multi-task visual grounding. *neural information processing systems*.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021b. 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 and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2592–2607, Online. Association for Computational Linguistics.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. *european conference on computer vision*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. *european conference on computer vision*.

Jingyu Liu, Wei Wang, Liang Wang, and Ming-Hsuan Yang. 2020. Attribute-guided attention for referring expression generation and comprehension. *IEEE Transactions on Image Processing*.

Tongtong Liu, Fangxiang Feng, and Xiaojie Wang. 2021. Multi-stage pre-training over simplified multimodal pre-training models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2556–2565, Online. Association for Computational Linguistics.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Ruotian Luo and Gregory Shakhnarovich. 2017. Comprehension-guided referring expressions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3125–3134. IEEE Computer Society.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 11–20. IEEE Computer Society.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649. IEEE Computer Society.

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

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

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 658–666. Computer Vision Foundation / IEEE.

Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2015. Grounding of textual phrases in images by reconstruction. *european conference on computer vision*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Mengyang Sun, Wei Suo, Peng Wang, Yanning Zhang, and Qi Wu. 2022. A proposal-free one-stage framework for referring expression comprehension and generation via dense cross-attention. *IEEE Transactions on Multimedia*, pages 1–1.

Hao Tan and Mohit Bansal. 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)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Mikihiro Tanaka, Takayuki Itamochi, Kenichi Narioka, Ikuro Sato, Yoshitaka Ushiku, and Tatsuya Harada. 2019a. Generating easy-to-understand referring expressions for target identifications. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5793–5802. IEEE.

Mikihiro Tanaka, Takayuki Itamochi, Kenichi Narioka, Ikuro Sato, Yoshitaka Ushiku, and Tatsuya Harada. 2019b. Generating easy-to-understand referring expressions for target identifications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *neural information processing systems*.

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

Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. 2019. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1960–1968. Computer Vision Foundation / IEEE.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, Hongxia Yang, and Chang Zhou<ericzhou. 2022. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework.

Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. 20a. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *ArXiv preprint*, abs/.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 20b. Simvlm: Simple visual language model pretraining with weak supervision. *ArXiv preprint*, abs/.

Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. 2020. Phrasecut: Language-based image segmentation in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10213–10222. IEEE.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason A. Smith, Jason Riesa, Alex Rudnick, Oriol

Vinyals, Greg S. Corrado, Macduff Hughes, and Jeffrey Dean. 20. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv preprint*, abs/.

Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. 2020. Improving one-stage visual grounding by recursive sub-query construction. *european conference on computer vision*.

Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. 2019. A fast and accurate one-stage approach to visual grounding. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4682–4692. IEEE.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. *national conference on artificial intelligence*.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. 2018. Mattnet: Modular attention network for referring expression comprehension. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1307–1315. IEEE Computer Society.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. *european conference on computer vision*.

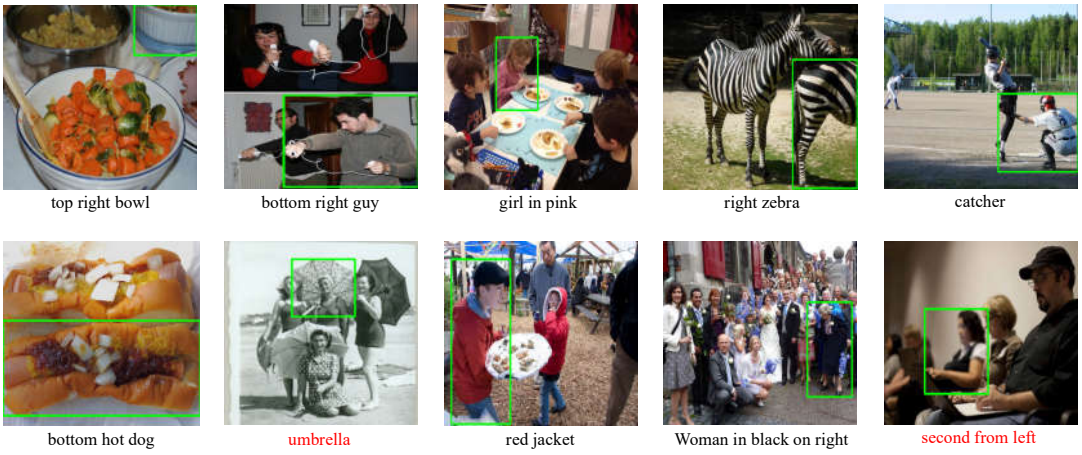
Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L. Berg. 2017. A joint speaker-listener-reinforcer model for referring expressions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3521–3529. IEEE Computer Society.

Yan Zeng, Xinsong Zhang, and Hang Li. 2021. Multi-grained vision language pre-training: Aligning texts with visual concepts. *ArXiv preprint*, abs/.

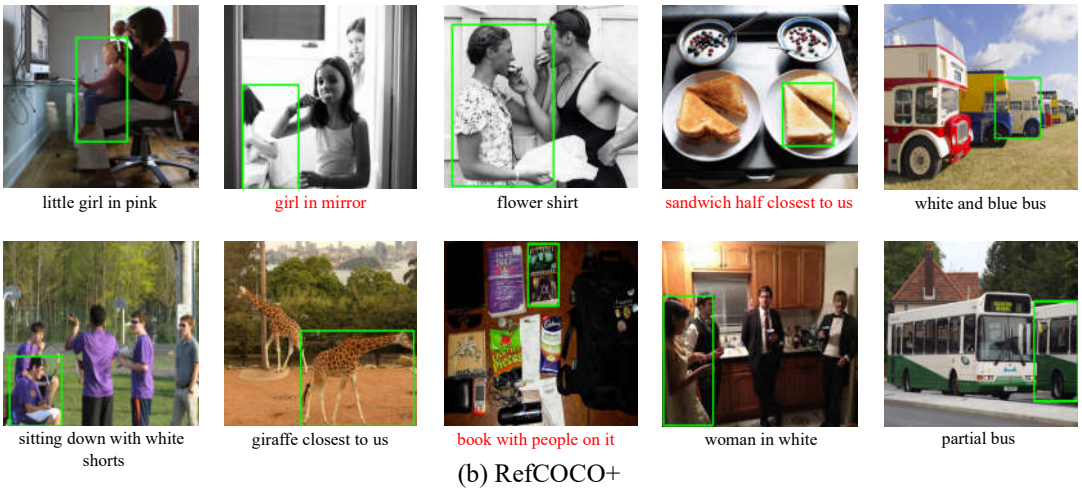
Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and VQA. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13041–13049. AAAI Press.

A Examples on REG and REC

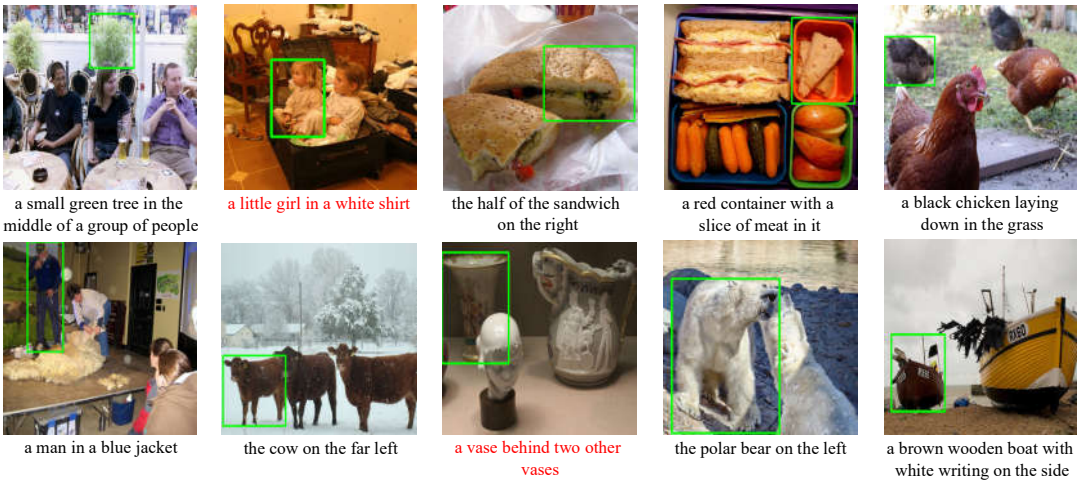
We give more uncurated examples on REG and REC in Fig. 4 and 5, respectively.



(a) RefCOCO

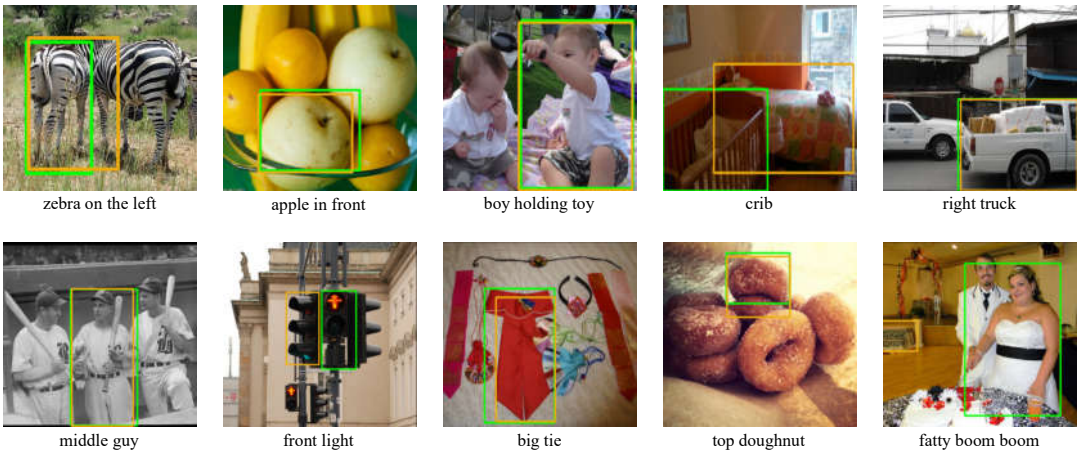


(b) RefCOCO+



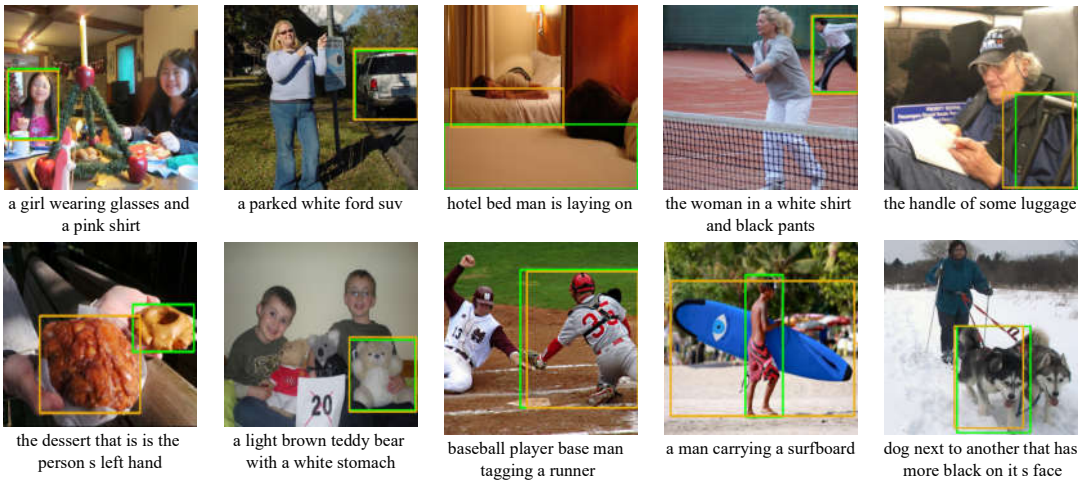
(c) RefCOCOg

Figure 4: Uncurated examples of the generated text in REG. The inaccurate or ambiguous text is marked in red.



(a) RefCOCO

(b) RefCOCO+



(c) RefCOCOg

Figure 5: Uncurated examples of the predicted bounding box in REC. The green and orange boxes indicate the ground truth and prediction, respectively.