Prompting Vision-Language Models For Aspect-Controlled Generation of Referring Expressions

Anonymous ACL submission

Abstract

Referring Expression Generation (REG) is the task of generating a description that unambiguously identifies a given target in the scene. Different from Image Captioning (IC), REG requires learning fine-grained characteristics of not only the scene objects but also their surrounding context. Referring expressions are usually not singular; an object can often be uniquely referenced in numerous ways, for instance, by color, by location, or by relationship with other objects. Most prior works, however, have not explored this 'aspect-based multiplicity' of referring expressions. Hence, in this work, we focus on the Aspect-Controlled REG task, which requires generating a referring expression conditioned on the input aspect(s), where an aspect captures a style of reference. By changing the input aspect such as color, location, action etc., one can generate multiple distinct expressions per target region. To solve this new task, we first modify BLIP (Li et al., 2022a) for aligning image-regions and text-expressions. We achieve this through a novel approach for feeding the input by *drawing* a bounding box around the target image-region and *prompting* the model to generate the referring expression. Our base REG model already beats all prior works in CIDEr score. To tackle Aspect-Controlled REG, we append 'aspect tokens' to the prompt and show that distinct expressions can be generated by just changing the prompt. Finally, to prove the high-quality and diversity of the data generated by our proposed aspect-controlled REG model, we also perform data-augmentation-based evaluation on the downstream Referring Expression Comprehension (REC) task. With just half of the real data augmented with the generated synthetic data, we achieve performance comparable to training with 100% of real data, using a SOTA REC model(Kamath et al., 2021).

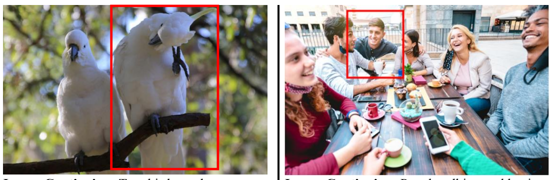


Image Captioning: Two birds on the tree
Dense Captioning: A white bird
Referring Expression Generation: Bird on the right
Referring Expression Generation (shape): The larger bird.

Image Captioning: People talking and having coffee
Dense Captioning: Man in blue
Referring Expression Generation: The man standing behind a man and woman
Referring Expression Generation (action): The man looking at another man

Figure 1: IC, DC, REG and aspect-controlled REG tasks

1 Introduction

Referring Expression Generation (REG) is the task of generating a descriptive caption that uniquely identifies a given target in the scene. REG is different from IC, which requires generating captions for the whole image (Li et al., 2022a)(Yu et al., 2022). REG is also different from the Dense Captioning (DC), which is aimed at generating detailed description for each salient region in the image but the descriptions are not required to uniquely identify a target (Yin et al., 2019) (Johnson et al., 2016). An example is shown in Figure 1. IC captures the high-level summary of the image (*"two birds on the tree"*). DC provides a brief description of the target region (*"A white bird"*). REG, on the other hand, generates a reference that allows the target to be uniquely located (*"Bird on the right"*).

Before 2020, REG in Vision-Language (VL) domain was popular (Mao et al., 2016a; Yu et al., 2016a; Liu et al., 2017; Tanaka et al., 2019). Models could refer to a target using either spatial or textual features. More recently, REG is often used as a pretraining task when learning multimodal representations (Yang et al., 2021; Lu et al., 2022; Wang et al., 2022). In order to generate unique object references, the model needs to *understand* fine-grained features of not only the object but also its situated context and ground those features in the generated expressions. This makes REG a good pretraining task for unified VL models. REG is also useful for cheaply generating synthetic training datasets for downstream tasks such as REC. This helps reduce the high cost associated with collecting and human-annotating large scale datasets. This is one of the main motivations of our work.

A distinct feature of referring expressions, which hasn't been explored much in prior work, is its aspect-based *multiplicity*. In reality, there are almost always multiple ways to refer to a target in the scene. For instance, as shown in Figure 1, the target bird can be referred by describing its appearance (*"the larger bird"*), its location (*"bird on the right"*) or its action (*"the bird cleaning its feather"*). Similarly, the man in the red box on the right-side picture can be referred in numerous ways. Each description captures a unique *aspect* of the referring expression. In this paper, we propose an aspect-controlled REG model that can generate multiple valid expressions for referring the same target region. Moreover, the style of the expressions generated is controllable by an aspect (e.g., color, location, action etc.) specified as natural language input. Aspect-Controlled REG has applications in goal-oriented dialogue systems that are now-a-days ubiquitous and allow users to complete simple tasks like restaurant reservation, flight booking, shopping etc. For instance, when a customer asks *"Can you show me a similar table but with different color?"*, the agent should focus on the color attribute and respond as *"What about this one in brown?"*, rather than talking about other aspects like material (*"the wooden one"*). In addition, an REG model capable of generating multiple aspect-controlled expressions has arguably a better *understanding* of this complex task as it has learned to cover all the unique properties of the object and capture the inherent diversity in referring expressions. This also leads to better utilization of multiple ground-truth references often available in standard REG datasets. Finally, this allows generating richer and more diverse synthetic datasets for downstream tasks. Our main contributions are:

- We explore the Aspect-Controlled REG task where an expression needs to be generated conditional on the provided aspect. By changing the input aspect, we can generate multiple expressions for the same target region.

- We modify BLIP (Li et al., 2022a) to align image-regions and corresponding text-expressions. We achieve this via a novel approach of feeding the input: by *drawing* a bounding box around the target image-region and *prompting* the model to describe the marked region. Our REG method beats all prior works in CIDEr score.

- To tackle Aspect-Controlled REG, we append 'aspect tokens' to the prompt and show that by merely changing the prompt, we can fully control the style of the generated expressions.

- Finally, we showcase the high-quality and diversity of the synthetic data generated by our proposed Aspect-Controlled REG model by evaluating on the downstream task of REC. With just 50% of real data augmented with our synthetically generated data, we achieve performance comparable to training with 100% of real data using a SOTA REC model(Kamath et al., 2021).

2 Related Works

2.1 REG and REC

Most previous REG models in the literature consist of a visual encoder and text decoder, where the focus is on REG and REC together (Mao et al., 2016a; Liu et al., 2017; Yu et al., 2017; Luo and Shakhnarovich, 2017; Liu et al., 2020). Other works in this area propose region specific modules after the vision encoder to understand the higher context between objects (Yu et al., 2016a), graphical approaches (Kim et al., 2020), reinforcement learning (Tanaka et al., 2019) to improve the diversity of generated expressions, and minimization on the semantic distance between predictions and ground truth (Panagiaris et al., 2020).

Most recent works do not focus solely on REG, with a few exceptions e.g., (Sun et al., 2022; Kim et al., 2021), and instead rely on expression generation as one of the many tasks in their multi-task framework, (Lu et al., 2022; Yang et al., 2021; Wang et al., 2022). REC is a foundational task for most state-of-the-art unified VL models pretrained on large datasets (Wang et al., 2022; Yang et al., 2021; Kamath et al., 2021).

2.2 CLIP and Contrastive Learning

Contrastive learning enables models to better learn multi-modal feature alignment by forcing the models to distinguish similar and different data, all in a non-supervised setting. It has been a mainstay in numerous VL models (Wang et al., 2021; Nan et al., 2021; Chen et al., 2022b,a), with increasing popularity after its usage in CLIP(Radford et al., 2021). In later work, BLIP(Li et al., 2022a) and CoCa(Yu et al., 2022) improve CLIP by applying a multitask pretraining scheme that minimizes contrastive loss and captioning loss together; GLIP(Li et al., 2022b) and (Zhang et al., 2021) align regions with object category words within the image captions.

2.3 Aspect-Controlled Generation

Controlled generation have been studied in many domains, e.g., natural language generation(Hu et al., 2017) and image generation(Karras et al., 2021). In closely related work for aspect-controlled image captioning, (Mathews et al., 2018; Guo et al., 2019) propose models to generate captions of a certain style such as positive, negative, subjective and objective; (Chen et al., 2020, 2021) propose solutions to generate captions that contain specific objects or actions. In these work, the requested control can be fed through a text encoder and combined with visual features (Mathews et al., 2018; Guo et al., 2019; Chen et al., 2021); or the request can be provided as an input graph that contains objects and relations (Chen et al., 2020). In our work, we leverage prompts on specific aspects (e.g., color, action) to achieve this control over the generated reference, where the model learns how to relate different prompts to various aspects during training.

3 Methodology

To summarize our full method pipeline, we build upon the BLIP Multimodal Mixture of Encoder-Decoder (MED) model architecture (Li et al., 2022a), adapting it for aligning image-regions and text- expressions describing those regions. We introduce a novel and intuitive approach for feeding the input; by *drawing* a bounding box around the target region in the image and *prompting* the model to describe the marked region. To reinforce that the generated descriptions are unique referring expressions, we introduce a simple technique to craft negative examples that are utilized during contrastive learning. Finally, we propose an intuitive yet novel approach for generating expressions conditioned on a given aspect (e.g. color, location etc.), by simply appending the aspect tag to the input prompt. To effectively evaluate our model, we generate synthetic data for training models for the downstream task of REC and use the REC performance as the evaluation metric. This evaluation approach allows handling multiple expressions (with various aspects) generated per target, by our REG model.

The architecture and training setup of the model is shown in Figure 2. We follow the general structure of BLIP(Li et al., 2022a) that consists of uni-modal image and text encoder, an image-grounded text encoder and an image-grounded text decoder. Our additions, here, are the modified image, the additional prompt input and new loss computations. The overall system is first pre-trained in a multitask manner, jointly minimizing region-expression contrastive loss, region-expression matching loss and expression generation loss. Following this, image encoder and text decoder are fine-tuned only with the expression generation loss on larger images. In the following sections, we detail each of these new components of our proposed system.

3.1 Region-Expression Alignment

As mentioned in Section 2.2, most prior CLIP-based models have focused on the image-caption level. (Li et al., 2022b), (Zhang et al., 2021) and (Zhong et al., 2022) are among a handful of works that learn alignment between image regions and text spans. However, their focus is on a single image object and simple expressions. For instance, matching the image-region containing *cat* to the phrase *"a photo of cat"*. In this work, we allow alignment of regions with more complex expressions involving surrounding context (e.g. *"a cat next to a dog"*) through two simple design choices:

- We draw a red rectangle on the input image marking the bounding box of the target region.
- We add the prompt "Describe the red box inside the image:" prior to generating the target expression.

As shown in the upper part of Figure 2, the modified image becomes the input to the image encoder. During pre-training, the prompt is appended before the ground-truth expression and fed to the two text encoders (uni-modal and image-grounded). The prompt is also used by the decoder to generate expressions. During inference, only the text decoder is utilized to generate an expression following the prompt. The rationale here is to provide a cue to visual encoder, image-grounded text encoder and decoder to *focus* on the target region of the image. At the same time, since the whole scene is fed

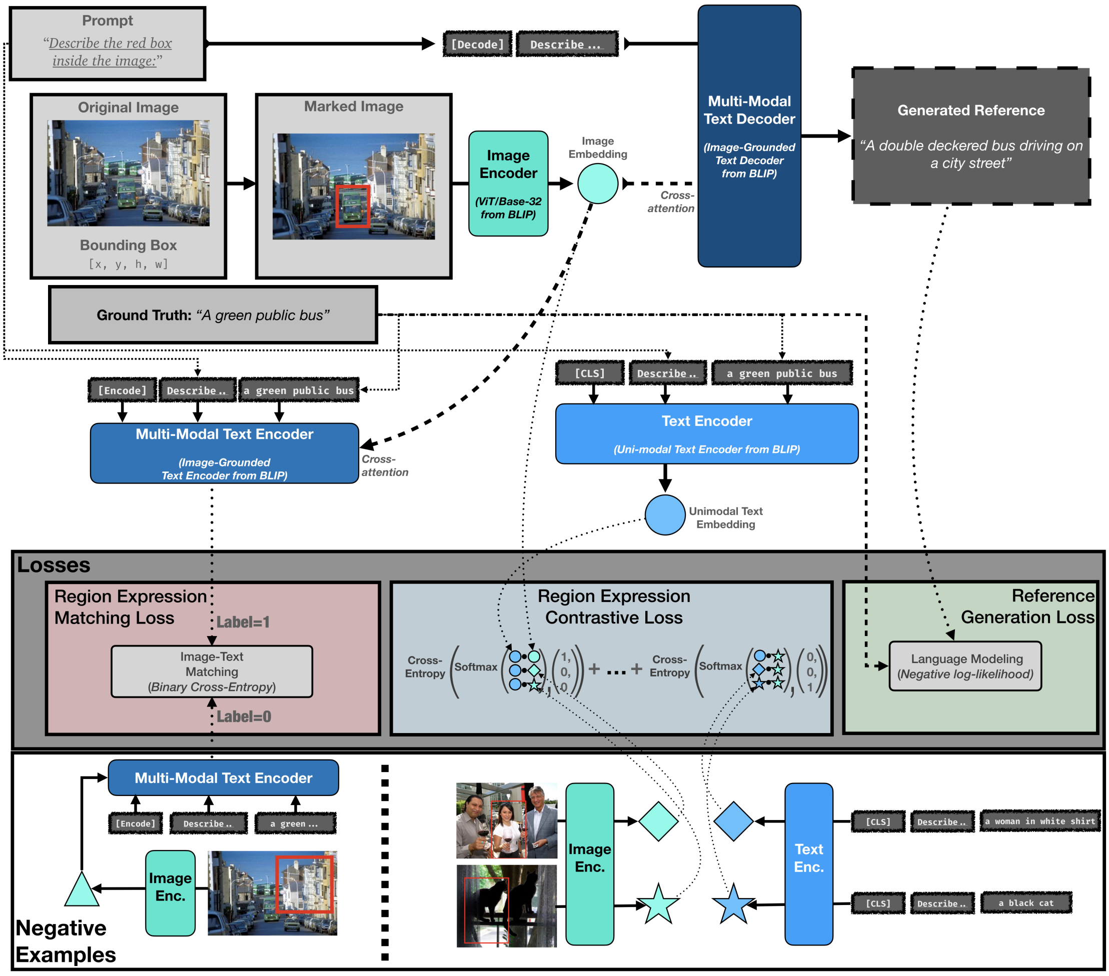


Figure 2: Pretraining model architecture and training objectives of the proposed REG system. We adapt BLIP (Li et al., 2022a) that consists of uni-modal image and text encoders, an image-grounded text encoder and an image-grounded text decoder. The target region is marked with a red bounding box in the original image and fed to image encoder. The prompt is appended to the decoder input, and fed alongside the visual embedding from the image encoder to the text decoder to generate the text. Simultaneously, we concatenate the prompt with the ground-truth expression, and feed the combined tokens to uni-modal and multi-modal text encoders. The encoders and the decoder are trained with a specific loss for each network: multi-modal text encoder is trained using image-text matching loss; uni-modal text encoder is trained via contrastive loss; and the multi-modal text decoder uses a generation loss.

as input, the model can also utilize the surrounding context to generate a unique description of the marked region. These descriptions can involve other scene objects and relationships to them as shown in Figure 1. Using text prompt as additional input provides benefit of controlling the style of generations as discussed in Section 3.3. A similar idea has been tried by (Yao et al., 2021), where a colored mask is laid on the target and aligned with a color-based text prompt. Adding a color mask, however, can distort the features of the original image and mislead the generation model. Therefore, we use a bounding box marker to keep the original image largely unchanged.

3.2 Hard Negatives Design

A referring expression needs to *uniquely* and *unambiguously* identify the target object within the image. This is a more challenging task as merely describing the target region may not be sufficient. For instance, in Figure 3, an expression such as *"a man drinking with a cup"* is not sufficient to identify the person in the center, as there are two men drinking with a cup in that image. To allow the model to learn to generate distinct expressions, we employ a contrastive learning approach. We create hard region-expression negative pairs which are utilized in the region-expression matching loss (explained in Section 3.4) during the pre-training stage.

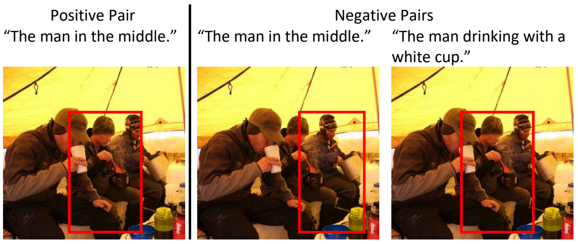


Figure 3: Hard negatives generation for contrastive learning. We create negative samples by changing either the bounding box or the reference in a positive pair, keeping the other constant.

To create a negative pair, we start with a positive pair and modify it with one of the following two strategies: (1) we update the target image-region to a randomly sampled region outside the original target in the same image, keeping the target expression as is; (2) we replace the target expression with one referring another object in the same scene, keeping the target image-region unchanged.

This approach is particularly beneficial when there are multiple objects of the same category in a scene allowing the model to learn to contrast and distinguish each object. An example is shown in Figure 3. The first negative example will allow the model to *understand* the scene layout and thereby generate unique spatial references involving object locations in the image. The second negative pair will force the model to learn to *understand* the nuanced details of the content of the image regions so as to generate discriminative references.

3.3 Aspect-Controlled Referring Expression Generation

It is always possible to refer a target object in a scene in multiple ways with different referring expressions capturing different aspects of the target object and its situated context. Such aspects can include *descriptive properties* like color, shape, pattern etc. of the target object, *spatial properties* such as the target object's location in the scene (e.g., in the middle) and *visual relationships* like spatial (e.g. on top of), action (e.g. cutting), comparative (e.g., larger than) etc. capturing its interaction with other scene objects. An expression can also capture a combination of these aspects (e.g., *"a man in white waving a bat"* as shown in Figure 4).

In this work, we propose a simple approach to control the style of generated referring expressions along these aspect dimensions. This is achieved by providing the target-aspect(s) as additional input to the model via the prompt. The target-aspect is added at the end of the default prompt i.e., *"Describe the red box inside the image by* $< aspect >$*"*. For instance, as shown in Figure 4, when the target-aspect is specified as color, the model generates the expression *"the man in white"*, while when the aspect is action, the generated expression is *"the man waving a bat"*; both expressions uniquely pointing to the hitter in the image. In order to train the system, we first annotate the aspects reflected in the training set referring expressions through rule-based heuristics. Specifically, we construct a pool of key words for each aspect and perform keyword-search on expressions. This rule-based process annotates ∼80% of the data. We then train a BERT-based classifier on this rule-labelled data and utilize it for annotating the rest of the training set. Finally, the expression generator is trained as shown in Figure 2, with annotated aspect(s) added to the end of the default prompt.

We consider four salient aspects of referring expressions in this work; color, shape, location and action and all their possible combinations. This was primarily motivated based on the structure of expressions seen in popular referring expression datasets. Our approach, however, is extensible to any number of aspect dimensions. As we show in Section 4.2, one of the main practical advantages of the proposed controlled generation approach is the capability to generate richer and more diverse synthetic dataset for downstream tasks such as referring expression comprehension, reducing the requirement of human-labeled data. Furthermore, most referring expression datasets provide multiple ground-truth expressions associated with the same target image-region. When fed as independent training examples with image and target-region being the only input, this can potentially lead to model confusion. In our approach, however, these examples will be split because different prompts will be associated with different ground-truth expressions, thereby, easing the training process.

3.4 Multitask Pretraining

We adapt the multitask training scheme in BLIP(Li et al., 2022a) and COCA(Yu et al., 2022). Our model is trained to minimize three losses:

- **Region-Expression Contrastive Loss**: It is the middle part of the loss block in Figure 2. This loss is computed on the outputs of uni-modal

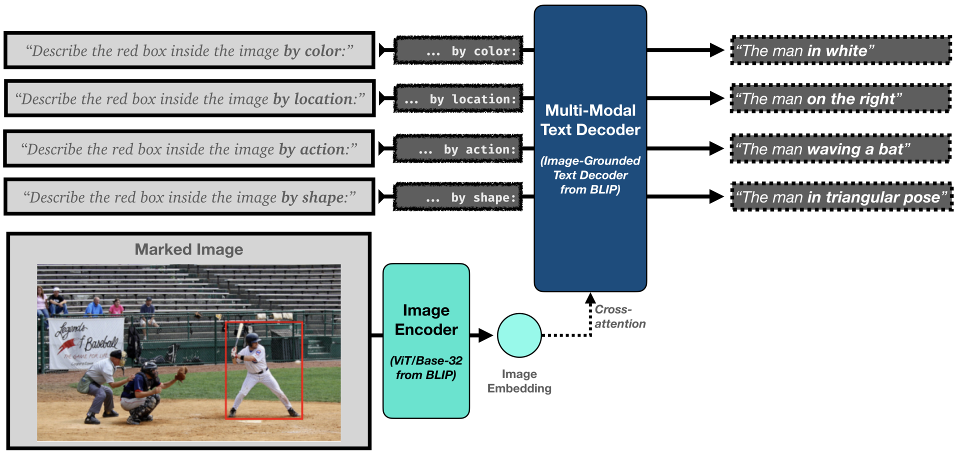

“Describe the red box inside the image by color:”
“Describe the red box inside the image by location:”
“Describe the red box inside the image by action:”
“Describe the red box inside the image by shape:”
… by color:
… by location:
… by action:
… by shape:
Multi-Modal Text Decoder
(Image-Grounded Text Decoder from BLIP)
“The man in white”
“The man on the right”
“The man waving a bat”
“The man in triangular pose”
Marked Image
Image Encoder
(ViT/Base-32 from BLIP)
Image Embedding
Cross-attention

Figure 4: Decoding process of our aspect-controlled REG model. Any combination of the 4 aspects (color, location, action, shape) can be utilized to construct the prompt. The target region is marked in the image with a box and fed to image encoder to get visual embedding. The constructed prompt is fed to the text decoder along with this embedding to generate the corresponding style of reference. The encoder and decoder are the same as Fig 2.

encoders to maximize the alignment between encoded image-region and text expression features of positive pairs while minimizing it for negative pairs. We use the Image-Text Contrastive (ITC) loss from (Li et al., 2021).

- **Region-Expression Matching Loss**: It is the left part of the loss block in Figure 2. This is the binary classification loss computed on the output of image-grounded text encoder.

- **Expression Generation Loss**: It is the right part of the loss block in Figure 2. As found by COCA(Yu et al., 2022), pre-training with captioning task helps the model learn fine-grained region-level features. We, therefore, also add the expression generation task during pre-training.

An important modification in our setup is that we conduct contrastive learning at two levels; inter-image and intra-image. For the region-expression contrastive loss, we create region-expression negative pairs across different images. For the region-expression matching loss, negative pairs are created from the same image as discussed in Section 3.2. The first task is relatively easier because region features from different images usually vary widely. It allows the uni-modal encoders to train fast, capture and align higher-level features of the image and expression (e.g., differentiating a cat from a car). The second task is harder because features from the same image will often be similar, e.g., having the same environment or belonging to the same object category. This task, therefore, enables the models to learn more detailed multimodal representations to distinguish between closely matching inputs.

We use the above three losses for pretraining on smaller-size images. Then we fine-tune the models on larger images using only the generation loss.

3.5 REG Evaluation via Data Augmentation for Referring Expression Comprehension

In order to evaluate the generative models, a common practice is to compute n-gram overlap metrics such as CIDEr(Vedantam et al., 2015). These metrics measure similarity between predicted and ground-truth text sequences. However, these only capture similarity to a single ground-truth expression and are not well-suited to evaluate the diversity inherent in our proposed aspect-controlled REG task. Furthermore, it is not possible to determine which aspect(s) of the expression is present in the test set for any given example, without looking at the labels. Therefore, to show the full potential of our approach, besides intrinsic evaluation with the above mentioned automatic metrics, we also perform extrinsic evaluation on the downstream task of REC. We first generate synthetic data with the proposed REG model, then train SOTA REC models (such as MDETR(Kamath et al., 2021)) using the generated data, and finally evaluate the REC model w.r.t. accuracy on standard expression comprehension benchmarks. This approach allows us to utilize multiple expressions generated by our REG model and the computed REC accuracy is comparable with those reported in prior works.

	RefCOCOg	RefCOCO+		RefCOCO	
	test	testA	testB	testA	testB
(Liu et al., 2017)	0.639	0.512	0.704	0.710	1.257
(Yu et al., 2017)	0.742	0.579	0.798	0.804	1.358
(Tanaka et al., 2019)	0.763	0.663	0.812	0.859	1.375
(Liu et al., 2020)	0.645	0.585	0.692	0.802	1.301
(Sun et al., 2022)	0.749	0.722	0.758	0.877	1.333
Ours	**1.069**	**1.039**	**0.966**	**1.119**	**1.527**

Table 1: Comparison of our proposed model with SOTA REG models on CIDEr metric. We do not apply any aspect-control here and use the default prompt.

4 Experiments

For intrinsic evaluation, we train and test our model on RefCOCO (Yu et al., 2016b), RefCOCO+ (Yu et al., 2016b) and RefCOCOg (Mao et al., 2016b) separately. We use CIDEr as metric. For extrinsic evaluation, we train existing REC models on references generated by our REG model, and examine the REC performance on RefCOCO/g/+ test sets using Acc@0.5 as the metric. We use our REG model trained on RefCOCOg and select MSCOCO images that *do not* overlap with any of the RefCOCO/g/+ datasets to generate the synthetic data for training comprehension models.

4.1 Intrinsic Evaluation

We use AdamW optimizer. The whole system is first pre-trained at 1e-5 learning rate. Then, the image encoder and text decoder are fine-tuned with 1e-6 learning rate. The image size is 224 × 224 for pretraining and 384 × 384 for fine-tuning. Note that, we use the term 'default prompt' to refer to the prompt - (*"Describe the red box inside the image"*), where no aspect is specified.

Table. 1 shows the performance of our expression generation model in comparison to prior works on RefCOCO/g/+ test sets. For fair comparison, we use only the default prompt in this experiment and generate only one expression per input region. Our proposed system outperforms all previous works by a large margin on CIDEr score.

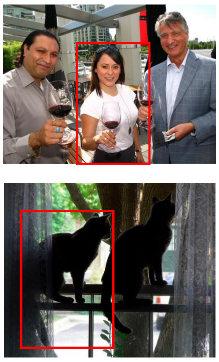
Next, we apply aspect-controlled prompts. Table 2 shows the results under different prompt setups. We experiment with 3 main settings: (1) default prompt at both training and testing, (2) prompt with annotated aspect(s) at training and a fixed-aspect prompt during testing, and (3) prompts with all aspects (*"Describe the red box inside the image by location, color, shape and action"*) at both training and testing. Because the ground-truth expressions and their aspects are unknown at test time, we experiment with feeding prompts with each aspect, one at a time. In setting 3, we provide prompts with all aspects to the model. As shown in the table, setting 1 and 3 have higher CIDEr scores compared to any experiment under setting 2. This is because, in these two settings, the training and testing prompts are consistent, unlike in setting 2 where the fixed test prompts may not match the training prompts. For setting 2, we find that changing the aspect in prompt largely does not affect the score. This is likely because n-gram overlap metrics like CIDEr do not capture the nuances in different styles of generated expressions, reinforcing our strategy to further evaluate on downstream REC task as explained in Sec. 3.5. In Fig. 5, we show our model's predictions on two examples. In both cases, changing the aspect(s) leads to corresponding change in the style of the generated expression. More examples can be found in Appendix.

Train Prompt	Test Prompt	CIDEr
Default	Default	1.069
Annotated	Default	0.917
Annotated	Action	0.898
Annotated	Color	0.946
Annotated	Location	0.971
Annotated	Shape	0.985
All	All	1.039

Table 2: Comparison of different prompt selection strategies at training and testing. Experiments are on RefCOCOg dataset. 'Annotated' means the prompts are constructed by the rule + BERT classifier. 'Default' refers to the prompt *"Describe the red box inside the image:"*, 'All' refers to the prompt *"Describe the red box inside the image by color, location, action and shape:"*.

	RefCOCOg	RefCOCO+		RefCOCO	
	test	testA	testB	testA	testB
MDETR	80.89	84.09	70.62	89.58	81.41
Only Syn	76.38	77.49	61.36	82.94	72.88
+10%Real	78.86	81.44	66.17	87.50	78.49
+30%Real	80.43	82.04	68.38	87.89	79.84
+50%Real	80.54	83.08	69.97	88.63	80.33

Table 3: Impact of replacing real training data with synthetic data generated by our Aspect-Controlled REG model for REC. We use MDETR(Kamath et al., 2021) as the REC model and Acc@50 as the metric. The first row is MDETR trained on all real data. Only Syn refers to only using synthetic data. +x% refers to additional x% real data.



Default: A woman in a white shirt.
By color: A woman in a white shirt.
By location: The woman in the middle.
By action: A woman holding a glass of wine.
By action and color: A woman in a white shirt holding a glass of wine.
Ground truth: This is a woman holding a wineglass and is wearing a white tshirt. / A woman in a white blouse holding a glass of wine.

Default: The cat on the left.
By color: A black cat.
By location: The cat on the left.
By action: Cat looking at another cat.
By action and color: A black cat looking at another cat.
Ground truth: Shorter cat on left side. / A cat whose tail is hiding behind the curtain.

Figure 5: Qualitative examples showing the behavior of prompts with different aspects for REG.

Lastly, we conduct a preliminary study on the faithfulness of our aspect-controlled generation. We report these results in the Appendix.

4.2 Extrinsic Evaluation

We select MDETR(Kamath et al., 2021) to evaluate the quality of the data generated by our REG model. It is has SOTA REC performance on RefCOCO/g/+ datasets. We use the same setting as provided in its paper. We select 158,367 annotations from 47,801 MSCOCO images which *do not* belong to any of RefCOCO/g/+ datasets to generate our synthetic data. For each annotation, we randomly sample a set of aspects (from the four categories) to construct the prompt and then generate an expression for it. We first train the model only on synthetic data of the same size as the training sets, then add real data.

Table 3 shows the results on the three datasets. Row 1 reports numbers when MDTER is fine-tuned on real RefCOCO/g/+ training sets and tested on corresponding test set. We report these numbers directly from original paper. In subsequent rows, we fine-tune MDTER on only synthetic data and synthetic data mixed with varying proportions of real data from the corresponding training set. Note that, our synthetic sets neither contain images from the original RefCOCO/g/+ datasets, nor any human-written references. As seen in the table, trained purely on this generated data, MDETR already achieves performance close to its original reported value that used 100% human-annotated data. As we add real data ranging from 10% to 50% to the synthetic dataset, the performance quickly approaches to that with 100% real data. Consequently, our proposed REG model can be used to significantly cut down annotation budget. Lastly, we use up all real and synthetic data, the performance is further improved, reported in Appendix.

Our controlled expression generation approach provides greater benefit for downstream tasks because it produces a more diverse set of references compared to traditional beam search method, given the same amount of data. In Appendix, we compare our approach with beam searching and the result shows our method generates expressions of various styles while beam search generates highly-similar ones.

Lastly, in Appendix, we conduct the aforementioned experiments using VL-T5 (Cho et al., 2021), another joint VL model published in 2021. We observe similar results as MDETR. We also generate synthetic data using (Tanaka et al., 2019) and compare with ours. Our model shows better performance. Besides, we include an analysis of comprehension errors in the Appendix.

5 Conclusion

We present a model to generate referring expressions for a given object in arbitrary ways, where we use a prompt to guide our decoder. Our approach, compared to traditional beam search, provides synthetic data of higher quality as evidenced in its diversity and ability to achieve higher accuracy with the same amount of training data.

6 Limitations

A limitation of our method would be the use of red box. It may fail in specific images such as the images that have red boxes inside.

Moreover, our study only covers 4 aspects, while more aspects could be included.

By now, there is not a dataset to test the performance of aspect-controlled generation directly. In the future, it would be good to build such a dataset that measures if models can generate references following the prompts.

References

Cheng Chen, Zhenshan Tan, et al. 2022a. Utc: A unified transformer with inter-task contrastive learning for visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18103–18112.

Feilong Chen, Xiuyi Chen, et al. 2022b. Improving cross-modal understanding in visual dialog via contrastive learning. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7937–7941.

Long Chen, Zhihong Jiang, et al. 2021. Human-like controllable image captioning with verb-specific semantic roles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16846–16856.

S. Chen, Q. Jin, et al. 2020. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9959–9968, Los Alamitos, CA, USA. IEEE Computer Society.

Jaemin Cho, Jie Lei, et al. 2021. Unifying vision-and-language tasks via text generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1931–1942. PMLR.

Longteng Guo, Jing Liu, et al. 2019. Mscap: Multi-style image captioning with unpaired stylized text. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4199–4208.

Zhiting Hu, Zichao Yang, et al. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596. PMLR.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574.

Aishwarya Kamath, Mannat Singh, et al. 2021. Mdetr - modulated detection for end-to-end multi-modal understanding. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1760–1770.

T. Karras, S. Laine, and T. Aila. 2021. A style-based generator architecture for generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4217–4228.

Jung-Jun Kim, Dong-Gyu Lee, et al. 2021. Visual question answering based on local-scene-aware referring expression generation. *Neural Netw.*, 139(C):158–167.

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.

Junnan Li, Dongxu Li, et al. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Junnan Li, Ramprasaath Selvaraju, et al. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705. Curran Associates, Inc.

Liunian Harold Li, Pengchuan Zhang, et al. 2022b. Grounded language-image pre-training. In *CVPR*.

Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. 2017. Referring expression generation and comprehension via attributes. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4866–4874.

Jingyu Liu, Wei Wang, et al. 2020. Attribute-guided attention for referring expression generation and comprehension. *IEEE Transactions on Image Processing*, 29:5244–5258.

Jiasen Lu, Christopher Clark, et al. 2022. Unified-io: A unified model for vision, language, and multi-modal tasks. *ArXiv*, abs/2206.08916.

R. Luo and G. Shakhnarovich. 2017. Comprehension-guided referring expressions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3125–3134, Los Alamitos, CA, USA. IEEE Computer Society.

Junhua Mao, Jonathan Huang, et al. 2016a. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20.

Junhua Mao, Jonathan Huang, et al. 2016b. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20.

A. Mathews, L. Xie, and X. He. 2018. Semstyle: Learning to generate stylised image captions using unaligned text. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8591–8600, Los Alamitos, CA, USA. IEEE Computer Society.

Guoshun Nan, Rui Qiao, et al. 2021. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2765–2775.

Nikolaos Panagiaris, Emma Hart, et al. 2020. Improving the naturalness and diversity of referring expression generation models using minimum risk training. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 41–51, Dublin, Ireland. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, et al. 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.

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

Mikihiro Tanaka, Takayuki Itamochi, et al. 2019. Generating easy-to-understand referring expressions for target identifications. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5793–5802.

R. Vedantam, C. Zitnick, and D. Parikh. 2015. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, Los Alamitos, CA, USA. IEEE Computer Society.

Liwei Wang, Jing Huang, et al. 2021. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14090–14100.

Peng Wang, An Yang, et al. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR*, abs/2202.03052.

Zhengyuan Yang, Zhe Gan, et al. 2021. UniTAB: Unifying Text and Box Outputs for Grounded Vision-Language Modeling. *ArXiv*, abs/2111.12085.

Yuan Yao, Ao Zhang, et al. 2021. Cpt: Colorful prompt tuning for pre-trained vision-language models. *ArXiv*, abs/2109.11797.

Guojun Yin, Lu Sheng, et al. 2019. Context and attribute grounded dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiahui Yu, Zirui Wang, et al. 2022. Coca: Contrastive captioners are image-text foundation models. *ArXiv*, abs/2205.01917.

Licheng Yu, Patrick Poirson, et al. 2016a. Modeling context in referring expressions. In *Computer Vision – ECCV 2016*, pages 69–85, Cham. Springer International Publishing.

Licheng Yu, Patrick Poirson, et al. 2016b. Modeling context in referring expressions. In *Computer Vision – ECCV 2016*, pages 69–85, Cham. Springer International Publishing.

Licheng Yu, Hao Tan, et al. 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.

Yizhen Zhang, Minkyu Choi, et al. 2021. Explainable semantic space by grounding language to vision with cross-modal contrastive learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 18513–18526. Curran Associates, Inc.

Yiwu Zhong, Jianwei Yang, et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803.

A Appendix

A.1 Experiment Settings

We adapt the same model settings and training settings of BLIP for our generation model. 4 Nividia V100 GPUs are used. The training time in total is around 24 hours. For MDETR and VL-T5, we follow their original settings.

A.2 Ablation Study

To quantify the importance of our design choices, we perform ablation study on RefCOCOg test set and report the results in Table 4. First, we directly utilize vanilla BLIP for REG. Low scores in row 1 clearly indicate that the original design of BLIP is not suitable for REG. Next, we study the effect of incorporating different loss functions in 3.4. For practical reasons, we perform these comparisons at the pretraining stage. As seen in row 2-4, notable increase on CIDEr is gained by adding each loss. Finally, the last row shows that fine-tuning our model with generation loss provides further improvement over the model pretrained with all 3 losses.

Experiment Settings	Stage	CIDEr
Vanilla BLIP	Pretrain	0.584
Gen loss	Pretrain	0.946
Gen + Ctr loss	Pretrain	0.963
Gen + Ctr + Mtc loss	Pretrain	0.999
Gen loss	Fine-tune	1.069

Table 4: Ablation study on model design and training strategies. Experiments are on RefCOCOg dataset. Row 1: Vanila BLIP applied for REG. Row 2-4: Our model pretrained with various losses (Gen: generation loss, Ctr: contrastive loss, Mtc: matching loss). Row 5: Our model fine-tuned on gen loss (Pretrained with all losses).

A.3 REG Faithfulness

We performed a preliminary study on the faithfulness of our aspect-control strategy. For RefCOCOg test set, we generated expressions conditioned on a single aspect and computed the % of expressions containing that aspect. The result is included in Table 5. Note that, an aspect (in isolation) may not be sufficient to uniquely refer every target region. This is, however, not accounted in this analysis as we run the model on every example. In addition, sometimes an aspect cannot be used to describe an object, e.g., the action of a chair, the shape of a zebra. They are likely the reasons for low faithfulness w.r.t *action* and *shape* coupled with the fact that there are very few ground-truth expressions with that aspect.

color	location	action	shape
87%	79%	49%	5%

Table 5: Faithfulness of the proposed aspect-control REG model.

A.4 REC Error Analysis

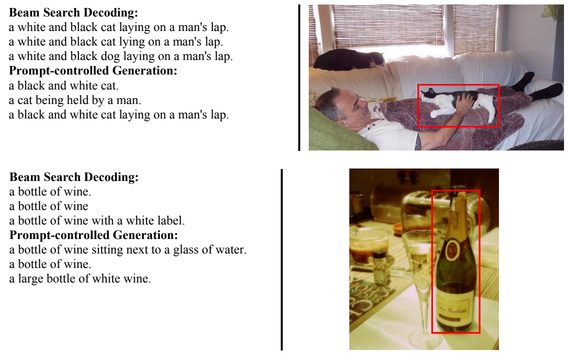
We examine the test errors for the task of REC with MDETR trained on 100% synthetic data. We collect the error cases and categorize them across the 4 types of references. The statistics is shown in Table 6. Location based references are the hardest ones. They account for ∼45% of all errors. This is expected because resolving a reference by location requires the model to *understand* the relationship between the target and its environment, while for other types of references, the model mostly needs to *look* at the features of the target. In addition, the proportion of the 4 types of references in our synthetic data is almost equal as the prompts are randomly sampled. However, in real data, the distribution may not be uniform. For instance, in RefCOCOg training set, there are ∼38.5% ground-truth references by location but only 6% by shape. This observation suggests that a better sampling strategy can be employed such that difficult examples (e.g. references by location) are generated more frequently so as to create a better synthetic training set for downstream tasks. We leave further investigation in this direction for future work.

	Loc	Color	Action	Shape
% of errors	44.53%	36.68%	13.32%	5.86%

Table 6: Error distribution on RefCOCOg test set with MDETR trained on 100% synthetic data.

A.5 REC with All Data

We use up all our 158,637 synthetic data and real training data to train MDETR and test its performance. As shown in Table 7, by using up all data, the performance of MDETR exceeds the one using all real data. Given that our model can generate synthetic data with low cost, one may expect that in the future the performance can be further improved



Beam Search Decoding:
a white and black cat laying on a man's lap.
a white and black cat lying on a man's lap.
a white and black dog laying on a man's lap.
Prompt-controlled Generation:
a black and white cat.
a cat being held by a man.
a black and white cat laying on a man's lap.

Beam Search Decoding:
a bottle of wine.
a bottle of wine
a bottle of wine with a white label.
Prompt-controlled Generation:
a bottle of wine sitting next to a glass of water.
a bottle of wine.
a large bottle of white wine.

Figure 6: Examples of referring expression variations generated from controlling aspect via prompt vs beam search decoding (k=3). The control words are randomly sampled.

by including more synthetic data from external raw images.

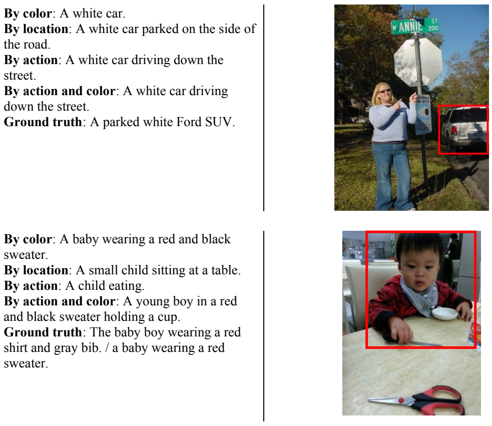
A.6 Comparison with Beam Searching

We hypothesize that our controlled expression generation approach will provide greater benefit for downstream tasks because it produces a more diverse set of references compared to traditional beam search method, given the same amount of data. To test this hypothesis, we run the following experiment. Starting with a fixed number of images, we generate expressions 1) using our generation model with default prompt and beam search decoding with beam size = 3 and 2) using the aspect-controlled variant of our generation model with 3 randomly sampled prompts utilized during decoding. These two settings lead to the same amount of synthetic data. With the two datasets, we train MDETR model and test on RefCOCOg test set. We run experiments varying the number of images used for training. The results are shown in Table 8. Our prompt-controlled generation has obvious advantage over the beam search decoding, especially when the number of input images is small. As the model trained on more images, the gap becomes narrower.

Figure 6 shows two sets of references; one generated by beam search and the other by prompt control. The result from the top three beams are quite similar to each other. On the other hand, the results generated by varying prompts are more diverse and can refer the target in different ways.

A.7 REC with VL-T5

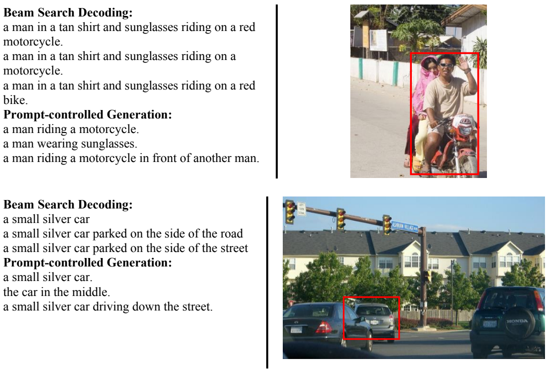
As mentioned in Sec 4.2, we perform the same REC experiments on VL-T5. The results are in Table 9 and 10. Similar to MDETR, trained purely on our generated data, it already achieves performance close to its original reported value that used 100% real training data. As we add annotated data ranging from 10% to 50% to the synthetic dataset, the performance readily approaches the value with 100% real data. Lastly, when using up all data, it outperforms the original VL-T5 by a large gap.



By color: A white car.
By location: A white car parked on the side of the road.
By action: A white car driving down the street.
By action and color: A white car driving down the street.
Ground truth: A parked white Ford SUV.

By color: A baby wearing a red and black sweater.
By location: A small child sitting at a table.
By action: A child eating.
By action and color: A young boy in a red and black sweater holding a cup.
Ground truth: The baby boy wearing a red shirt and gray bib. / a baby wearing a red sweater.

Figure 7: More examples showing the behavior of default prompt (*"Describe the red box inside the image:"*) and prompts appended with different aspects like color, location, action for generating referring expressions.



Beam Search Decoding:
a man in a tan shirt and sunglasses riding on a red motorcycle.
a man in a tan shirt and sunglasses riding on a motorcycle.
a man in a tan shirt and sunglasses riding on a red bike.
Prompt-controlled Generation:
a man riding a motorcycle.
a man wearing sunglasses.
a man riding a motorcycle in front of another man.

Beam Search Decoding:
a small silver car
a small silver car parked on the side of the road
a small silver car parked on the side of the street
Prompt-controlled Generation:
a small silver car.
the car in the middle.
a small silver car driving down the street.

Figure 8: More examples comparing referring expressions generated using beam search decoding vs by varying prompt. The 'aspect tokens' in the prompt are randomly sampled.

A.8 Comparison with Other REG Models for Synthetic Data Generation

We also use another REG model (Tanaka et al., 2019), to generate synthetic data, and conduct our extrinsic evaluation on RefCOCO/g/+ datasets. We only test the accuracy of MDETR where 100% synthetic data is used for training. The result is shown in Table 11. Our model outperforms (Tanaka et al., 2019) by a large gap.

	RefCOCOg	RefCOCO+		RefCOCO	
	test	testA	testB	testA	testB
MDETR	80.89	84.09	70.62	89.58	81.41
All	81.50	84.46	71.86	89.85	82.39

Table 7: MDETR trained on all our synthetic data and real data. The first row is MDETR trained on all real data. All means using up all real and full set of 158,367 synthetic examples from non-overlapping COCO images.

# of Images	# of References	Beam Approach	Aspect Prompt
3,000	25,013	72.68	76.58 (+3.90)
6,000	49,806	74.65	77.61 (+2.96)
12,000	99,876	75.23	76.85 (+1.62)

Table 8: Comparison of synthetic data quality generated with beam search decoding vs aspect-controlled REG. Using the synthetic data, we train MDTER model for REC and evaluate on the RefCOCOg test set by Acc@50.

	RefCOCOg	RefCOCO+		RefCOCO	
	test	testA	testB	testA	testB
Reference	71.2	76.09	59.21	85.89	72.68
Only Syn	67.39	65.12	49.76	72.88	59.92
+10%Real	69.27	73.21	54.65	80.94	68.97
+30%Real	70.07	75.20	57.19	83.24	70.87
+50%Real	71.14	76.02	57.58	84.15	71.89
All	73.22	79.39	62.36	85.93	73.03

Table 9: Impact of replacing real training data with varying amounts of synthetic data generated by our Aspect-Controlled REG model for the task of REC. We use VL-T5 as the REC model and Acc@50 as the metric. 'Reference' is the accuracy reported in the VL-T5 paper. Only Syn refers to model trained purely on synthetic data. +x% refers to additional x% real training data. All means using up all real and synthetic data. VL-T5 does not report its results on RefCOCO and RefCOCO+. We compute those numbers ourselves.

# of Images	# of References	Beam Approach	Aspect Prompt
3,000	25,013	66.56	67.20
6,000	49,806	66.50	66.65
12,000	99,876	66.64	67.52

Table 10: Comparison between beam search decoding and prompt-controlled generation in terms of Acc@50 on RefCOCOg test set for the task of REC using VL-T5 model.

	RefCOCOg	RefCOCO+		RefCOCO	
	test	testA	testB	testA	testB
(Tanaka et al., 2019)	69.48	69.73	56.06	71.82	58.53
Our REG	76.38	77.49	61.36	82.94	72.88

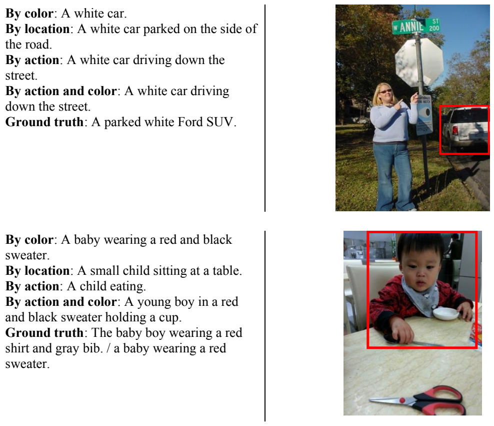
Table 11: Performance of MDETR trained on synthetic data generated from (Tanaka et al., 2019) and our REG.

A.9 Aspect Annotation

Table 12 shows statistics on annotated aspects. As seen from the table, this approach labels majority of the data leaving only ∼2% as unlabeled. Most of the unlabeled expressions are brief phrases such as *"A refrigerator"*.

Location	Color	Action	Shape	Unlabeled
47,083 58.48%	35,733 44.38%	14,187 17.62%	5,871 7.29%	1,865 2.32%

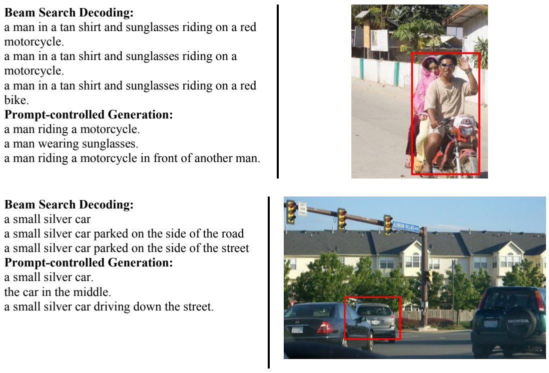
Table 12: A summary of aspect-class distribution in training data. Note that a reference can belong to multiple aspect classes.



By color: A white car.
By location: A white car parked on the side of the road.
By action: A white car driving down the street.
By action and color: A white car driving down the street.
Ground truth: A parked white Ford SUV.

By color: A baby wearing a red and black sweater.
By location: A small child sitting at a table.
By action: A child eating.
By action and color: A young boy in a red and black sweater holding a cup.
Ground truth: The baby boy wearing a red shirt and gray bib. / a baby wearing a red sweater.

Figure 9: More examples showing the behavior of default prompt (*“Describe the red box inside the image:"*) and prompts appended with different aspects like color, location, action for generating referring expressions.



Beam Search Decoding:
a man in a tan shirt and sunglasses riding on a red motorcycle.
a man in a tan shirt and sunglasses riding on a motorcycle.
a man in a tan shirt and sunglasses riding on a red bike.
Prompt-controlled Generation:
a man riding a motorcycle.
a man wearing sunglasses.
a man riding a motorcycle in front of another man.

Beam Search Decoding:
a small silver car
a small silver car parked on the side of the road
a small silver car parked on the side of the street
Prompt-controlled Generation:
a small silver car.
the car in the middle.
a small silver car driving down the street.

Figure 10: More examples comparing referring expressions generated using beam search decoding vs by varying prompt. The ‘aspect tokens’ in the prompt are randomly sampled.

A.10 Other Examples

Figure 9 shows two more examples on our aspect-controlled generation. Figure 10 shows two more examples that compare the references generated by beam-searching approach and prompt control.