DisCLIP: Open-Vocabulary Referring Expression Generation

Lior Bracha*[1]
brachal@biu.ac.il
Eitan Shaar*[1]
shaarei@biu.ac.il
Aviv Shamsian[1]
aviv.shamsian@live.biu.ac.il
Ethan Fetaya[1]
ethan.fetaya@biu.ac.il
Gal Chechik[1,2]
gal.chechik@biu.ac.il

[1] Bar-Ilan University
Ramat-Gan, Israel
[2] NVIDIA
Tel-Aviv, Israel

Abstract

Referring Expressions Generation (REG) aims to produce textual descriptions that unambiguously identifies specific objects within a visual scene. Traditionally, this has been achieved through supervised learning methods, which perform well on specific data distributions but often struggle to generalize to new images and concepts. To address this issue, we present a novel approach for REG, named *DisCLIP*, short for discriminative CLIP. We build on CLIP, a large-scale visual-semantic model, to guide an LLM to generate a contextual description of a target concept in an image while avoiding other distracting concepts. Notably, this optimization happens at inference time and does not require additional training or tuning of learned parameters. We measure the quality of the generated text by evaluating the capability of a receiver model to accurately identify the described object within the scene. To achieve this, we use a frozen zero-shot comprehension module as a critique of our generated referring expressions. We evaluate DisCLIP on multiple referring-expression benchmarks through human evaluation and show that it significantly outperforms previous methods on out-of-domain datasets. Our results highlight the potential of using pre-trained visual-semantic models for generating high-quality contextual descriptions in new visual domains.

1 Introduction

Referring expressions (REs) are a key component of language communication. They allow people to refer to one entity in a complex visual scene, in an unambiguous way. Comprehending and generating REs is essential for embedded agents that need to communicate with people about their environment. For instance, an autonomous vehicle may inquire a passenger about their preferences - “Should I park in the nearest spot or the shaded one?", or a robot

* Equal Contribution.

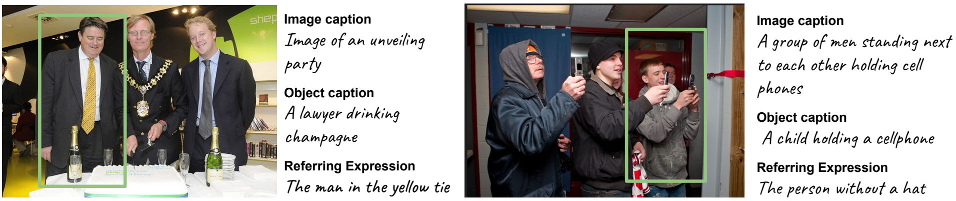


Image caption
Image of an unveiling party
Object caption
A lawyer drinking champagne
Referring Expression
The man in the yellow tie
Image caption
A group of men standing next to each other holding cell phones
Object caption
A child holding a cellphone
Referring Expression
The person without a hat

Figure 1: Referring Expressions Generation (REG) aims to generate textual descriptions that clearly identify an object in a given scene, ignoring similar distractors. REG is harder than object (dense) captioning because it must take into account the context of other objects. For instance, a REG model must be capable of identifying unique features such as the color of a tie (on the left) or general descriptions such as "*the man without the hat*" (right). The same object can have multiple distinct descriptions based on the context.

assistant may wish to clarify an instruction: “Which chair should I get you: the black one or the white one?".

Significant effort has been devoted to RE *comprehension*, namely, training agents to understand referring expressions generated in natural language. The current paper focuses on a complementary task: *referring expression generation* (REG), namely, training agents to refer to entities in a visual scene using natural language. RE generation and comprehension are complementary; they can be viewed as played by two communicating players [2, 3, 13, 28]. First, a *speaker* observes a scene that contains multiple objects and generates language that refers to a specific target object. Then, a second player, a *listener*, interprets the RE in the context of the same visual scene and selects the entity that is referred to. In this communication-as-a-game setup, the two players are cooperative and have a common objective, the speaker wishes to generate REs that are easily interpretable by the listener [17, 29].

Referring Expressions Generation (REG) generated by agents should satisfy two key properties: (1) discriminative - using attributes to point clearly to a unique object in the scene, and (2) intelligible - producing language that can be easily understood by people. Recent advances in NLP using web-scale corpora have been very successful in generating natural language, but datasets available for referring expressions are significantly smaller. As a result, current visual REG models are limited and do not transfer well to images outside the narrow domain they were trained on. In contrast, Visual-Linguistic (VL) models, such as CLIP [21] and LLM, were trained on large text corpora, encompassing a wide variety of expressions. Therefore, they provide a more versatile and general-purpose framework for REG. In addition, the vast scale of foundation models enables them to generalize effectively to new data, even in zero-shot scenarios.

Building on these advances we propose an approach that builds on LLMs and large VL models. Our approach is based on two key components. First, we use a pre-trained CLIP as a listener to evaluate how well a text phrase corresponds to an object in a given scene. Second, we introduce a method for using CLIP in a discriminative manner across localized boxes. This optimization is guiding the text generation of an LLM, at inference time. As such, it does not require any further training of learned parameters. Importantly, we avoid training the listener and speaker models jointly, because such training can lead to a “runaway” drift into a specialized language that is less natural for human interpretation. Furthermore, CLIP was trained on a large text corpora, which allows for an open-vocabulary generation, that

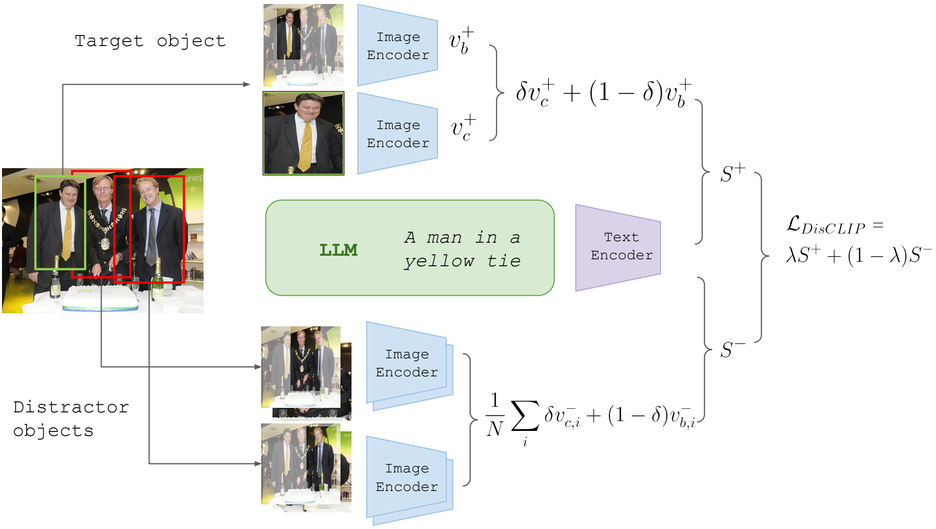

Target object
Image Encoder
v_b^+
Image Encoder
v_c^+
$\delta v_c^+ + (1-\delta)v_b^+$
S^+
LLM
A man in a yellow tie
Text Encoder
$\mathcal{L}_{DisCLIP} = \lambda S^+ + (1-\lambda)S^-$
S^-
Distractor objects
Image Encoder
Image Encoder
$\frac{1}{N}\sum_i \delta v_{c,i}^- + (1-\delta)v_{b,i}^-$

Figure 2: **DisCLIP architecture for REG.** DisCLIP score encourages the language model (green) to generate text that is semantically related to the target object. The output sequence of the LM is encoded by CLIP's text encoder (purple). CLIP image encoder (blue) is used to encode representations of the target object (v_c^+, v_b^+) as well as the set of other objects in the scene (v^-). At each timestep, we maximize CLIP similarity with the target object and minimize similarity with a set of distractors.

effectively generalizes to new vision and language domains.

Referring expressions have been traditionally separated into two types: relational ("the person on the left") and attribute-based ("the person with the hat"). This paper focuses on attribute-based REG because current VL models represent attributes much better than spatial relations.

This paper makes the following contributions. **(1)** First, we introduce the first approach to open-vocabulary visual referring expression generation named *DisCLIP* for discriminative CLIP. It allows generalizing to new data distributions and concepts, making it more versatile and adaptable to various applications. Notably, these results are achieved without any additional training or fine-tuning required. **(2)** We put forward a method that utilizes foundation models trained on image-level descriptions, for the generation of *contextual descriptions*. Such descriptions are costly to curate, and are rarely available, even in large VL corpora. **(3)** We show through extensive experiments that DisCLIP outperforms supervised methods on out-of-domain datasets in varying learning setups. Importantly, our model produces descriptions that are more natural and accurate, according to human raters.

2 Related work

REG task is often considered a proxy of pragmatic reasoning and naturally falls under the paradigm of a dialog. Effective communication and contextual language are further explored in the Rational Speech Act (RSA) framework [7]. Accordingly, a common architecture is a speaker and a listener, performing complementary tasks: REG and Referring Expressions

Comprehension (REC). A broad class of REG methods [9, 15, 16, 18, 30] rely on joint optimization of the speaker, and the listener. The risk in such a pipeline is creating a "secret" language [29] the speaker-listener architecture tends to overfit and struggle to generalize across domains. In contrast, our method does not require any training and entirely depends on inference time decoding. Other notable works in that field include [17], which designs a loss optimized to stir image captioning towards describing the differences between two images. [28] suggest a transmitter-emitter (ES) architecture. [2, 6, 28] generates pragmatically informative captions, by decoding general captioning models, at testing time, to produce captions that discriminate target images from a given set of distractor images.

Zero-Shot Image Captioning. RE methods can be viewed as a special case of image captioning methods. Recent work on open-world image captioning combines the abilities of two large pre-trained models: CLIP and GPT. [23] suggests regularizing sequences produced by GPT2 to be semantically related to a given image with a CLIP-induced score. Concurrently, [26] proposed a similar technique that relies on gradient update and optimization over context cache. This improves accuracy but dramatically slows down inference. In later work, [27] improves efficiency and inference speed by updating the context of a pseudo-token over different iterations in which the model generates full sentences. [19] suggests producing meaningful captions by initializing GPT2 with visual prefix embeddings, which is learned by employing a simple mapping network from CLIP embedding space to GPT2 space. [8] proposed to formalize the CLIP score as a new standard metric for image captioning.

Referring Expressions Generation (REG). As far as we know, the current state-of-the-art approach in REG is presented in [22]. Their method involves incorporating pragmatic reasoning into context-agnostic generation models during inference. To generate pragmatically informative captions, they decode general captioning models at test time, producing captions that discriminate the target image from a set of distractor images. Decoding methods include criteria such as likelihood (Beam Search) and informativity (RSA decoding) [6]. Although some models achieve state-of-the-art (SotA) results on certain subsets of RefCOCO/+/g, there is no single model that consistently outperforms others across the board. A key difference is that our approach is designed to perform on objects within the same visual scene, rather than a curated set of distractors.

[25] studies REs from the standpoint of object saliency. It has been observed that salient objects can be referred to using short and simple phrases, whereas less salient objects require more complex descriptions that often involve relationships with other objects within the scene. While our work does not specifically target this aspect, we draw comparisons with this baseline due to its zero-shot REG framework. Another work by [14, 15] explicitly learns visual attributes and uses them as a supervision signal for REG-REC modules. Recent work [10] achieves impressive results, but their approach is supervised and works in domain. Our focus is on out-of-domain generalization.

3 Workflows for REs generation and comprehension

Visual referring expression involves two complementary tasks. First, *generation*, where a speaker module is given an image and a bounding box of a target object and has to create a natural language expression that refers to that object. Second, *comprehension* where a listener module parses the RE, with the goal of selecting the correct object in a given image.

There are two main strategies for training these modules. The first strategy is to pretrain a listener, freeze it, and then use it as a frozen evaluator to measure the quality of predicted REs (e.g., [16, 22]). Here, since the listener is fixed, it is used to calculate a static score, like box-selection accuracy, which can be readily used for training the speaker. The second approach is to train both modules jointly [10, 29]. This raises two main difficulties. First, since the generated language is discrete, passing gradients from the listener to the speaker is non-trivial and involves approximated optimization like using a Gumbel softmax or straight-through [29]. Second, unless restricted, the two modules tend to drift away from natural language and pass information that is unintelligible to people [29]. To alleviate this issue, some researchers use language quality metrics like BLEU against a ground truth set. Unfortunately, these measures tend to be highly insufficient [1, 4]. In all these cases, methods are trained on paired data of images, boxes, and ground-truth referring expressions collected from human raters.

A potential issue with these workflows is that they tend to be limited to the distribution of the data they are trained on. Indeed, our experiments below show that when tested on new data, they may collapse and yield very low accuracy.

How can we progress towards open-world referring expression generation? We wish to provide dataset-agnostic models that can provide referring expressions that are both natural and informative even for images outside the training distribution. To this end, we propose to use large pre-trained image captioning models [21]. These models are trained on massive web datasets and, as such, cover the long tail of visual and semantic content.

4 Model

DisCLIP model is composed of two branches (Fig. 2): a language branch where a Large Language-Model (LLM) generates a sequence of words (Fig. 2 green box), and a visual branch that guides generation to be close to an input image in a visual-semantic space. In an iterative process, we maximize the similarity [8] between the generated sequence at a timestep $x_{<t}$ to *the target region* in the image, and minimize the similarity to a set of distractors regions (namely, other objects). Our work is closely related to [23], who put forward a similar process for zero-shot image captioning.

Let $x_{<t}$ be a sequence generated by an LLM at time t. Given input image $\mathcal{I}$, and $V^{(k)}$ (top k) candidate tokens from the language model (LM), the probability of candidate token v is computed as

$$f(v|\mathcal{I},x_{<t},V^{(k)}) = \frac{e^{CLIP(I,[x_{<t}:v])}}{\sum_{z\in V^{(k)}} e^{CLIP(I,[x_{<t}:z])}} \quad , \tag{1}$$

[:] denotes the concatenation operation, s.t. $x_{<t}:v$ represents the generated sequence so far, together with the current token. $CLIP(I,[x_{<t}:v])$ is the CLIP similarity score of an image I and text $[x_{<t}:v]$. In our case, given an image $\mathcal{I}$ containing n objects $O=\{o_1,\dots,o_n\}$, we require that the generated sequence maximize the CLIP similarity with a target object O^+, while minimizing CLIP similarity to a set of distractors $O^- = \{o_1^-,\dots,o_{n-1}^-\}$. The total score is defined as,

$$\mathcal{L}_{DisCLIP} = \lambda \left(\overbrace{CLIP(O^+, [x_{<t} : v])}^{S^+} \right) + (1-\lambda) \left(\frac{-1}{N} \sum_i \overbrace{CLIP(O_i^-, [x_{<t} : v])}^{S_i^-} \right). \tag{2}$$

The hyper-parameter $\lambda \in [0,1]$ controls how strongly the negative set affects generation. When $\lambda = 1$, negatives have no effect at all, and smaller values are expected to create increasingly discriminative text (see also [5]).

$$v = \text{argmax} \left\{ \mathcal{L}_{lang} + \beta \cdot \mathcal{L}_{disCLIP} \right\}. \tag{3}$$

The full objective includes two other terms designed for maintaining language fluency and consistency with the context tokens. For clarity, we refer to it in short as $\mathcal{L}_{lang}$, and describe it in detail in the Sec. E of the supplementary material. Finally, the hyperparameter β controls the tradeoff between language weight and the DisCLIP vision score.

Representing boxes. In contrast with standard captioning, the RE text-generating task has to: (i) describe a specific object in the scene instead of the entire image. This is challenging since CLIP was trained on image-level descriptions. (ii) The generated text should be contextual, which requires gathering information about the rest of the objects in the scene. This impacts how we experiment with the visual representation of the objects.

To capture both local and global information we create different representations for each object in the scene. The first is simply a crop of the object's box, and the second is a blurred version of the image, except for the target region. We discuss other representations in supplementary Section B.

Object representations are passed to the CLIP image encoder and used to compute the similarity to the generated text at time t. For the set of negatives, we sum over the similarity scores S_i,

$$S_i = \delta \cdot Blur(O_i) + (1-\delta) \cdot Crop(O_i) \quad , \tag{4}$$

where δ controls the trade-off between the two representations, as illustrated in Fig 2.

5 Experiments

We evaluate DisCLIP and the baselines in several experimental setups. We begin by showing out-of-domain performance measured using a pre-trained listener on three datasets: Flickr30k-Entities, RefCLEF, and RefGTA. Next, we put forward human evaluation results on our generated REs compared to the baselines. We also show that DisCLIP performs reasonably well in in-domain benchmarks compared to supervised methods. To encourage future research and reproducibility, we make our code publicly available at https://github.com/dekelio/disclip-open-vocab-reg.

Data. We used the following datasets in our experiments. **(1) RefCOCO** [13] contains 142,209 referring expressions for 50,000 objects in 19,994 images. **(2) RefCOCO+** [13] contains 141,564 referring expressions for 49,856 objects in 19,992 images. This dataset focuses on objects' appearance, rather than spatial relations. In both RefCOCO and RefCOCO+, Test A contains references to humans, and Test B references to other object types. **(3) RefCOCOg (Google RefExp)** [18] contains

85,474 referring expressions for 54,822 objects in 26,711 images and contains longer and more complex expressions. **(4) RefCLEF** (ReferIt) [12] A dataset containing *complex* photographs of real-world cluttered scenes. 10K test images, with $\sim 60K$ references in the train/val set and 60,105 in the test set. **(5) RefGTA** [25], contain synthetic images from the Grand Theft Auto (GTA) videogame. 6504 test images. All REs correspond to people, focusing on relations expressions. **(6) Flickr30k-Entities** [20], provides a comprehensive ground-truth correspondence between regions in images and phrases in captions. It contains 244K coreference chains, with 275K corresponding bounding boxes. We excluded "group" references (e.g. *People are outside waving flags*), resulting in 1966 images and 4597 references in the validation set and 4601 in the test. Since our method does not require training, we only used the validation and text splits in evaluations.

Baselines. We compared our approach with the following baselines with their model that trained on RefCOCO+: **Schutz et al. 2021** [22] adopts an Emitter-Suppressor (ES) framework of [28]. A speaker (E) models a caption for a target image I_t in conjunction with a listener function (S) that rates how discriminative is the utterance with regard to a distractor image, λ is a parameter that weighs the suppressor. We compare with $\lambda = 0.5$ for his best model. **Tanaka et al. 2019 [25]** suggested an end-to-end training for encoder decoder. Based on low-level visual representations as the input, various aspects of the task are modeled jointly, e.g. lexicalization and content selection. **Licheng Yu et al. 2017 [30]** proposed an end-to-end trained listener-speaker for RE task. He also added a discriminative reward-based module (reinforcer) to guide the sampling of more discriminative expressions and further improve his final model.

Method Efficiency. In our experiments, inference in all methods was on the same order of magnitude, with our method being ~ 0.4 seconds per object, with a batch size of 1. Our key advantage in terms of efficiency is the time it takes to train the model. DisCLIP requires no training, for comparison, the number of trainable parameters of the baseline methods is Schuz et al. $\sim 65M$, Yu et al. $\sim 20M$, and Tanaka et al. $\sim 23M$.

Evaluation metrics. Standard evaluation metrics for REs like BLUE or CIDER [22, 23] focus on agreement with ground-truth expressions. In the case of open-text generation, these metrics do not reflect true performance because LLMs produce rich natural sentences whereas GT phrases tend to be terse. To address this, we use two evaluation approaches: human raters and a frozen REC model – a "listener". We follow the protocol in [11, 24] and measure listener accuracy as the percentage of instances for which the predicted box whose IoU with the ground-truth box is at least 0.5, a standard metric used to evaluate RE methods. For consistency with previous works in the field, we also report standard language metrics, provided in the supplementary material Table S1.

Failure modes. In cases of significant overlap between bounding boxes, DisCLIP struggles, since its optimization method builds on contrasting them. This issue can be mitigated by employing simple heuristics like limiting the set of negatives by IOU, sampling, or by narrowing it to same-category objects. Another issue that is related to the use of bounding boxes, is that information about other

		RefClef		RefGTA		Flickr30K-Entities	
	trained on	Test A	Test B	Val	Test	Val	Test
SUPERVISED METHODS							
Schutz et al.[22]	RefCOCO+	26.0	18.2	11.5	11.8	31.7	32.0
Tanaka et al. [25]	RefCOCO+	27.0	33.4	52.5	53.2	34.6	39.6
Licheng Yu et al. [30]	RefCOCO+	38.0	41.4	31.2	31.8	50.9	49.0
OPEN-VOCABULARY							
DisCLIP (Ours) + ReCLIP [24]		66.2	68.6	58.0	56.9	77.9	78.8
DisCLIP-HPT (Ours) + ReCLIP [24]		**83.4**	**85.4**	**73.4**	**73.6**	**89.2**	**91.2**

Table 1: **Out-of-domain generalization.** Listener accuracy on three different datasets, RefClef, RefGTA, and Flickr30K-Entities.

objects within the target box can sometimes inadvertently affect the generated description. However, when studying the failure modes of the different methods, we observed that upon failing, the current methods tend to generate entirely unintelligible text. Our model, on the other hand, often preserves information that is helpful for recognizing the target object. For example, the attributes of the target objects or actions (See Table S5 of the supp.).

6 Results

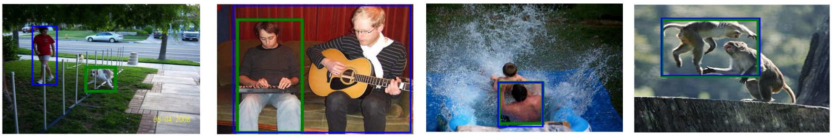
Out-of-domain generalization. We now evaluate all methods in an out-of-domain setup. We trained the baseline methods on RefCOCO+, which captures attribute-based referrals. DisCLIP requires no training, but we tuned its hyperparameters δ and λ on a subset of 200 random samples from the validation split, see Figure S1. In the following tables, "HPT" stands for Hyper Parameter Tuning. When we mention "DisCLIP-HPT," it pertains to the model in which we individually adjust the β and λ parameters for each dataset. In the evaluations below, we use the "natural" listener that is "paired" with the speaker in the sense that the listener was used either when training or evaluating the speaker in their original papers.

The results presented in Table 1 show that for the out-of-domain datasets RefCLEF, RefGTA, and Flickr30k-Entities, DisCLIP significantly outperforms the baselines methods. Qualitative examples are shown in Table. 5.

Independent pre-trained listener. The performance degradation of baselines observed in Table 1 might result from the domain shift that the listener (the REC model) experiences, rather than the speaker – which is our prime interest. We further test a single pre-trained REC model as a common listener to evaluate all different "speakers", in an identical way. For that listener, we choose mDETR [11]; an end-to-end modulated detector that detects objects in an image conditioned on a raw text query.

		RefClef		RefGTA		Flickr30K-Entities	
	trained on	Test A	Test B	Val	Test	Val	Test
GT RefExp		65.5	64.4	40.3	40.6	72.6	73.9
Schutz et al.[22]	RefCOCO+	34.8	26.4	**40.8**	**40.9**	**40.7**	**40.6**
Tanaka et al. [25]	RefCOCO+	22.8	20.4	38.9	40.2	32.0	31.1
Licheng Yu et al. [30]	RefCOCO+	27.6	22.0	24.8	25.2	31.8	31.1
DisCLIP (Ours)		35.0	29.8	33.0	32.6	37.0	36.7
DisCLIP-HPT (Ours)		**36.2**	**30.8**	33.9	33.3	36.3	35.9

Table 2: OOD Evaluation with an independent listener REC module (mDETR).



DisCLIP (Ours): **White dog** running at small bird stand that shows blue head and eyes.

DisCLIP (Ours): Person playing computer **keyboard** sitting on large metal tray.

Schutz et al.: Brown and white horse.

Schutz et al.: Dog.

Figure 3: An mDETR listener fails more with long natural sentences. In many cases, mDETR predicts a box (blue) that has a high overlap with GT box (green), even when the captions are completely unrelated to the image like in the two examples on the right. On the other hand, it often misses valid clues in the textual descriptions.

		RefClef	RefGTA	Flickr30K-Entities
	trained on	Test A	Val	Test
Schutz et al.[22]	RefCOCO+	31.0	36.0	43.1
Tanaka et al. [25]	RefCOCO+	6.0	46.0	19.0
Licheng Yu et al. [30]	RefCOCO+	14.0	26.0	29.0
DisCLIP (Ours)		**46.3**	**49.0**	**45.7**

Table 3: Human evaluation on the out-of-domain datasets. On average, DisCLIP outperforms the baseline methods by a margin of 29.3% on RefClef, 13% on RefGTA, and 15.3% on Flickr30K-Entities.

The results are presented in Table 2. DisCLIP outperforms the baselines on the RefClef dataset and is competitive on RefGTA and Flickr30K-Entities.

To understand this difference, we noted that mDETR was fine-tuned on RefCOCO/+/g, presumably making it tuned to sentences that are short and perform worse on rich natural sentences. Indeed, from a qualitative error analysis, we find that mDETR makes more mistakes with long sentences, potentially causing a bias against DisCLIP and favoring the baselines. See qualitative examples in Fig. 3.

Human evaluations on OOD REs. Given the above limitations of out-of-the-box listeners as evaluators, as well as traditional language metrics, we follow up with evaluation by human raters. We generated REs for 100 random samples from three out-of-domain datasets and sent each RE to three unique raters. Given a textual description (generated by us or the baselines), participants are asked to choose one out of n candidate boxes that best matches the RE (details in Supp. F). Table 3 shows that human raters prefer phrases generated by DisCLIP model, across all out-of-domain datasets by a large margin.

Naturalness and Diversity. Further analysis shows that phrases generated by current methods are very limited in terms of vocabulary and naturalness. In contrast, text produced by DisCLIP is $\times 8$ more diverse than the baselines (See sec. F.2 and Fig. S3 in the supp. material), which makes it better equipped to handle non-generic or complex scenes.

In-domain Referring Expressions. DisCLIP is designed for out-domain and open-vocabulary setup. For completeness, we also tested its accuracy on in-domain datasets RefCOCO/+/g in Table 4. Baseline models had both their listener and speaker trained on RefCOCO+ (attribute-based REs). There are also versions of the baseline models that were trained on RefCOCO, but since it is focused on

		In domain			GT Label shift				
		RefCOCO+			RefCOCO			RefCOCOg	
	trained on	Val	Test A	Test B	Val	Test A	Test B	Val	Test
SUPERVISED METHODS									
Schutz et al. [22]	RefCOCO+	58.3	68.4	48.2	58.2	68.4	48.1	62.1	62.5
Tanaka et al. [25]	RefCOCO+	65.8	70.9	62.5	65.8	70.9	62.2	72.0	71.4
Licheng Yu et al. [30]	RefCOCO+	**79.2**	**82.9**	75.0	**79.1**	**82.9**	**74.6**	**86.1**	85.7
OPEN-VOCABULARY									
DisCLIP (Ours) + ReCLIP [24]		67.3	70.1	64.7	67.2	70.1	64.5	72.8	74.9
DisCLIP-HPT (Ours) + ReCLIP [24]		78.6	80.2	**77.2**	76.5	80.2	73.7	85.1	**86.5**

Table 4: **In-domain accuracy** of models tested on three variants of RefCOCO. Each method uses a paired (jointly trained) speaker and listener. All datasets have the same distribution of images, but GT labels shift between datasets.

spatial phrases, this comparison is less relevant to our task, which focuses on attribute-based REs. Table 4 shows that DisCLIP stays competitive with the supervised baselines on all the in-domain datasets. Qualitative examples from both in and out-of-domain are shown in Table. 5.

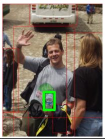
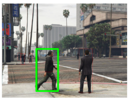
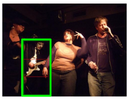
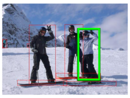
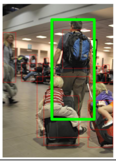
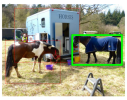
			Ground Truth	Schutz et al.[22]	Tanaka et al. [25]	Yu et al. [30]	DisCLIP (Ours)
Out-of-Domain	RefClef		**Water bottle** man is holding	Blue shirt	Barber barber barber barber barber	Curlys curlys curlys curlys curlys	**A blue water bottle** and clear wooden pot
	RefGTA		Man in black jacket and green pants **walking**	Skateboarder	Person in black	Curlys overlapping seed seed seed seed seed	**Man walking** away from area waiting for vehicle loading event
	Flickr30K		A man	Woman	Barber barber barber barber barber barber	Curlys curlys goo goo goo loops curlys goo goo	Young adult dressed as Prince **playing guitar** standing behind suit vest.
In-Domain	RefCOCO+		**Person** in **white**, shortest person in white, man with both hands up	Blue jacket	White jacket	Man in white jacket	**Person** skiing with **white vest** showing shirt standing in black foreground.
	RefCOCO		Man with blue **backpack**, guy dragging kids, black and blue backpack	Man in black	Man	Man in black jacket	Person carrying **backpack** standing talking on side.
	RefCOCOg		A **black horse with a blue cover** tied to a horse trailer	Blue jacket	Blue shirt	Brown horse	**Black horse** dressed down **wearing blue robe** posing for camera.

Table 5: **Qualitative results** on out-of-domain (top), and in-domain datasets (bottom).

Limitations. DisCLIP is successful, but it is also important to address its limitations. First, CLIP has notorious poor sensitivity to spatial relations. As a result, the expressions generated by our model use attribute-based REs rather than relation-based REs, like “bike on right”. Second, Our language generation is very simple, generating the expression token by token. It is appealing that smarter models for expression generation may improve performance. Given that DisCLIP does not rely on any training or fine-tuning procedures, using better foundation models in the future is expected to yield better REG using similar DisCLIP inference.

7 Conclusion

In this work, we present a novel method, named DisCLIP, to generate discriminative referring expressions in an open-world setting. Instead of training a model for one specific dataset, we leverage large pre-trained foundation models (CLIP, GPT2). DisCLIP achieves significant improvement over baseline models trained on different datasets, showing robustness to the domain shift occurring across datasets.

Acknowledgements

This study was funded by a grant to GC from the Israel Science Foundation (ISF 737/2018), and by an equipment grant to GC and Bar-Ilan University from the Israel Science Foundation (ISF 2332/18). Lior Bracha is supported by a Ph.D. fellowship in data science from the Israeli National Council of Higher Education.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016.

[2] Jacob Andreas and Dan Klein. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1125. URL https://aclanthology.org/D16-1125.

[3] Jacob Andreas and Dan Klein. Reasoning about pragmatics with neural listeners and speakers. *arXiv preprint arXiv:1604.00562*, 2016.

[4] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of bleu in machine translation research. In *11th conference of the european chapter of the association for computational linguistics*, pages 249–256, 2006.

[5] Gal Chechik and Naftali Tishby. Extracting relevant structures with side information. *Advances in Neural Information Processing Systems*, 15, 2002.

[6] Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. Pragmatically informative image captioning with character-level inference. *arXiv preprint arXiv:1804.05417*, 2018.

[7] Michael C Frank and Noah D Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012.

[8] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[9] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4555–4564, 2016.

[10] Shijia Huang, Feng Li, Hao Zhang, Shilong Liu, Lei Zhang, and Liwei Wang. A unified mutual supervision framework for referring expression segmentation and generation. *arXiv preprint arXiv:2211.07919*, 2022.

[11] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.

[12] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 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, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1086. URL https://aclanthology.org/D14-1086.

[13] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 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, 2014.

[14] Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. Referring expression generation and comprehension via attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4856–4864, 2017.

[15] Jingyu Liu, Wei Wang, Liang Wang, and Ming-Hsuan Yang. Attribute-guided attention for referring expression generation and comprehension. *IEEE Transactions on Image Processing*, 29: 5244–5258, 2020.

[16] Ruotian Luo and Gregory Shakhnarovich. Comprehension-guided referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7102–7111, 2017.

[17] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2018.

[18] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016.

[19] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[20] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[22] Simeon Schüz and Sina Zarrieß. Decoupling pragmatics: Discriminative decoding for referring expression generation. In *Proceedings of the Reasoning and Interaction Conference (ReInAct 2021)*, pages 47–52, 2021.

[23] Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*, 2022.

[24] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5198–5215, 2022.

[25] Mikihiro Tanaka, Takayuki Itamochi, Kenichi Narioka, Ikuro Sato, Yoshitaka Ushiku, and Tatsuya Harada. Generating easy-to-understand referring expressions for target identifications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5794–5803, 2019.

[26] Yoad Tewel, Yoav Shalev, Roy Nadler, Idan Schwartz, and Lior Wolf. Zero-shot video captioning with evolving pseudo-tokens. *arXiv preprint arXiv:2207.11100*, 2022.

[27] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17928, 2022.

[28] Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 251–260, 2017.

[29] Gilad Vered, Gal Oren, Yuval Atzmon, and Gal Chechik. Joint optimization for cooperative image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8898–8907, 2019.

[30] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speaker-listener-reinforcer model for referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7282–7290, 2017.