Supplementary Material for 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

DisCLIP is a framework that generates referring expressions for objects in a scene by leveraging an LLM and CLIP. It guides a language model by iteratively aligning the generated text with the target region in the CLIP space. This supplementary material includes additional experimental results and analysis. To facilitate comparison with previous supervised methods, we report standard language metrics. Furthermore, we conduct an ablation study to examine the impact of object representation methods and hyper-parameters on robustness. Lastly, we provide qualitative examples where human raters either succeeded or failed in a Referential Expression Comprehension (REC) task when provided with descriptions generated by our model and the baseline methods.

A Language Metrics

Commonly used evaluation metrics for referring expressions (REs) include language metrics such as BLUE [7], CIDEr [13], ROUGE-L [6], and METEOR [1]. These metrics primarily assess the agreement between generated expressions and a set of ground-truth references. However, when it comes to open-text generation, these metrics may not be suitable since language models (LLMs) produce detailed natural sentences while ground-truth expressions tend to be terse. Nonetheless, for the sake of consistency with previous studies, we provide the results of standard language metrics in Table S1.

Metrics for open-vocabulary text generation In the context of unsupervised or open-text generation settings, several metrics have been proposed [11]. These metrics are computed without relying on human annotation. One such metric is *Relatedness to the image*, which

© 2023. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.
* Equal Contribution.

	RefClef									
	Test A (Human)					Test B (Objects)				
	*BLEU*1 ↑	*BLEU*4 ↑	*METEOR* ↑	*CIDEr* ↑	*Rouge* $-L$ ↑	*BLEU*1 ↑	*BLEU*4 ↑	*METEOR* ↑	*CIDEr* ↑	*Rouge* $-L$ ↑
Schutz et al.[0]	**0.226**	0.000	0.080	**0.330**	**0.215**	**0.183**	0.000	0.066	**0.200**	**0.151**
Tanaka et al. [1]	0.022	0.000	0.020	0.095	0.057	0.020	0.003	0.019	0.077	0.049
Yu et al. [2]	0.046	0.000	0.035	0.109	0.084	0.031	**0.004**	0.028	0.085	0.062
DisCLIP (Ours)	0.123	**0.005**	**0.097**	0.088	0.158	0.120	0.000	**0.090**	0.063	0.144
DisCLIP-HPT (Ours)	0.096	0.000	0.080	0.059	0.126	0.098	0.000	0.083	0.060	0.126

	RefGTA									
	Validation					Test				
	*BLEU*1 ↑	*BLEU*4 ↑	*Meteor* ↑	*CIDEr* ↑	*Rouge* $-L$ ↑	*BLEU*1 ↑	*BLEU*4 ↑	*Meteor* ↑	*CIDEr* ↑	*Rouge* $-L$ ↑
Schutz et al.[0]	0.078	0.018	0.072	**0.161**	0.192	0.076	**0.019**	0.072	**0.164**	0.190
Tanaka et al. [1]	0.110	0.012	0.063	0.151	0.179	0.110	0.014	0.064	0.158	0.179
Yu et al. [2]	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
DisCLIP (Ours)	**0.307**	**0.018**	**0.097**	0.070	**0.234**	**0.301**	0.017	**0.096**	0.068	**0.233**
DisCLIP-HPT (Ours)	0.196	0.004	0.079	0.048	0.162	0.196	0.004	0.078	0.047	0.161

	Flickr30k Entities									
	Validation					Test				
	*BLEU*1 ↑	*BLEU*4 ↑	*Meteor* ↑	*CIDEr* ↑	*Rouge* $-L$ ↑	*BLEU*1 ↑	*BLEU*4 ↑	*Meteor* ↑	*CIDEr* ↑	*Rouge* $-L$ ↑
Schutz et al.[0]	0.140	0.008	0.078	**0.309**	0.141	**0.141**	0.000	0.079	**0.329**	0.190
Tanaka et al. [1]	0.024	0.001	0.027	0.148	0.066	0.022	0.000	0.024	0.150	0.061
Yu et al. [2]	0.028	0.000	0.030	0.141	0.068	0.024	0.000	0.027	0.139	0.063
DisCLIP (Ours)	0.098	0.003	**0.095**	0.067	**0.156**	0.098	**0.002**	**0.096**	0.068	**0.157**
DisCLIP-HPT (Ours)	0.050	0.001	0.065	0.043	0.080	0.050	0.000	0.064	0.044	0.081

Table S1: Language Metrics for REG

assesses the distance between the image and the generated text using a retrieval-based approach. *Language quality* is evaluated by estimating the *perplexity* score of the generated caption, utilizing BERT. The perplexity score, which is the negative logarithm of the likelihood, reflects the level of uncertainty in the model's text predictions. Another metric, *Diversity*, measures the vocabulary size and the percentage of novel sentences compared to the training set (%Novel).

B Ablation study

Representing objects. In our ablation study, we explored various methods for representing objects in the image. These methods include: (i) cropping the object's bounding box, (ii) blurring the entire image except the target region, and (iii) cropping with mirror padding for non-squared boxes, meant to maintain the proportions of the object.

Our findings indicate that a combination of cropping and blurring yields optimal results for this task. Our hypothesis is that this representation effectively encodes both local and global information. The act of cropping guides CLIP to focus on the target object, while the blurred surroundings provide valuable contextual cues. Moreover, since CLIP resizes the input image, cropped regions can sometimes become stretched, making it difficult to recover the object's semantics. In such cases, the blur representation proves useful as it maintains the objects' sizes and overall positioning of objects in the scene.

To balance between these two representations, we introduce the parameter δ. Table S2 compares several representation methods that we tested, using a fixed value of $\delta = 0.5$ determined through hyper-parameter tuning.

ReCLIP	RefClef - Test A	Flickr30k - Val
DisCLIP, crop-blur	**67.4**	**77.9**
DisCLIP, only blur	48.8	63.7
DisCLIP, only mirror	45.2	58.1
DisCLIP, only crop	52.2	61.7

Table S2: Accuracy of ReCLIP listener given different representations of the target object.

mDETR	RefClef - Test A	Flickr30k - Val
DisCLIP, crop-blur	**35.0**	**37.0**
DisCLIP, only blur	29.4	32.7
DisCLIP, only mirror	27.0	31.1
DisCLIP, only crop	30.4	34.2

Table S3: Accuracy of mDETR listener given different representations of the target object.

C Robustness of the speaker

In order to assess the robustness of the various speaker models, we aim to evaluate the extent of overfitting to the specific listener employed in the original paper. To achieve this, we decouple the speaker-listener pairs and measure accuracy across all possible combinations. Table S4 presents the accuracy of each speaker (REC task) when paired with different listeners. The cells highlighted in blue indicate paired speaker-listener combinations, where the listener was either trained jointly with the respective speaker or used in the original paper. It is worth noting that all supervised listeners underwent training on refCOCO+.

	mDETR		MCN		Tanaka		Yu		ReCLIP	
	Test A	Test B	Test A	Test B	Test A	Test B	Test A	Test B	Test A	Test B
Schutz et al.	34.8	26.4	28.8	19.8	31.2	36	31.6	37.6	44.6	44.2
Tanaka et al.	22.4	19.8	16.4	15.4	21.4	27.0	30.0	36.0	20.4	29
Yu et al.	27.4	22.2	20.2	14.4	**35.0**	**40.6**	38.0	41.4	22.6	28.2
DisCLIP	**36.2**	**30.8**	**24.4**	**15.6**	27.6	32.0	30.4	32.2	66.2	68.6

Table S4: **The effect of speaker-listener pairing.** Accuracy on RefClef dataset.

D Robustness to hyper-parameters

DisCLIP requires no training, but we tuned its hyperparameters δ and λ on a subset (1000 random samples from 3805) of RefCOCO+ validation split, see Figure S1. In all cases, we used 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 the original papers.

E DisCLIP Objective

The DisCLIP model consists of two branches: a language branch where a large language model (LM) generates a sequence, and a visual branch that guides the generation process

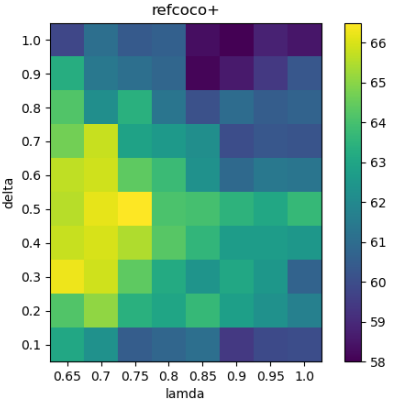


Figure S1: Sensitivity to Hyperparameters. Shown is the accuracy on RefCOCO+ as a function of various values of δ and λ. The color bar represents the Acc@0.5. Optimal values are obtained for $\lambda = 0.75$ – assigning significant weight to distractor boxes, but smaller than half, and with $\delta = 0.5$ – assigning equal weight to the local and global representation of the box in the image.

towards a target image region in a visual-semantic space. The overall objective is defined as:

$$v = \underset{v \in V^{(k)}}{\operatorname{argmax}} \Big\{ \mathcal{L}_{lang} + \beta \cdot \mathcal{L}_{DisCLIP} \Big\}. \tag{S1}$$

where v represents the next candidate token, $V^{(k)}$ denotes the set of top-k predictions from the model's probability distribution, and β is a hyper-parameter controlling the trade-off between language and vision. When $\beta = 0$, the visual controls are disabled.

In the main text, we discuss the vision part $\mathcal{L}_{DisCLIP}$, which maximizes the similarity [2] between the generated sequence and a specific *region* in the image, while minimizing the similarity to a set of distractor regions.

The complete objective also includes two additional terms designed to ensure language fluency and consistency with the context tokens, referred to as $\mathcal{L}_{lang}$. These terms were defined in the optimization procedure proposed in [10]:

$$\mathcal{L}_{lang} = (1-\alpha) \cdot \overbrace{p_\theta(v|x_{<t})}^{\text{model confidence}} - \alpha \cdot \overbrace{(max\{s(h_v, h_{x_j}) : 1 \leq j \leq t-1\})}^{\text{degeneration penalty}}$$

The last term was originally suggested in [10] and addresses the degeneration problem in text generation by language models. This problem refers to the generation of dull and repetitive text at different levels (e.g., token-, phrase-, and sentence-level). The authors propose the use of contrastive learning to calibrate the representation space of the language model. h_v represents the CLIP embedding of the sequence so far $x_{<t}$ and the current token v. We incorporate the degeneration penalty, along with the model confidence, into the score to guide the model toward likely outputs while avoiding the problem of model degeneration.

F Human Evaluation

We utilized the Amazon Mechanical Turk (AMT) platform to conduct a direct comparison of the generated referring expressions (REs) in a REC task performed by human participants. In this task, participants were presented with multiple candidate boxes (n) and asked to select the one that best corresponds to the provided textual description. The textual descriptions were generated by our system as well as the baseline methods. The task layout is depicted in Figure S2. To ensure robust evaluation, we collected evaluations for 100 randomly selected samples from each of the three out-of-domain datasets. Each sample was independently evaluated by three distinct annotators.

F.1 Naturalness

DisCLIP outperforms the baseline methods by generating more diverse and natural phrases, as demonstrated in the Flickr30k-Entities dataset. The wins and losses of our model compared to the baselines are presented in Table S5 and Table S6, respectively.

In Table S5a, for example, two baselines produce accurate but non-discriminative captions such as "*woman*". In contrast, our model specifies "*woman jumping*" uniquely identifying the target object. Additionally, it provides context (e.g., "*volleyball*") and information about attributes (e.g., "*black*"). In Table S5c, all models struggle. However, the raters found the descriptions of all other baselines unintelligible, except for ours. DisCLIP offers enough clues regarding visual attributes like "*yellow*" "*white*" "*green*" and "*ball*" aiding the raters in correctly identifying the colorful ball.

Table S6 presents examples where our method did not exhibit a clear advantage. In Table S6a, individuals correctly identified the target object in all cases, but DisCLIP provided more information than necessary. Table S6b illustrates a common issue where DisCLIP captures unwanted contextual information in cases of overlapping boxes. In the given example, the target object is the white shirt, but DisCLIP focuses on the woman and entirely omits the shirt. We anticipate that utilizing segmentation masks instead of boxes will help mitigate these problems.

F.2 Diversity

Models based on open text generation naturally have a larger vocabulary compared to supervised methods, which are trained on a limited, predetermined set of categories. In the case of the Flickr30k-test dataset, the baselines achieve a maximum vocabulary of 519 words, while our models cover a much larger vocabulary of 4279 words, which is more than **eight times** the size.

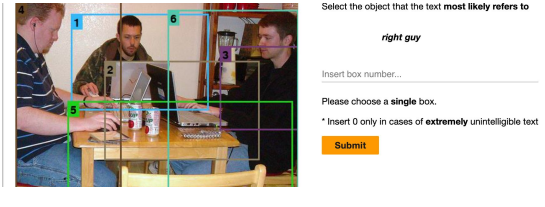


Figure S2: mTurk REC task. Given a sentence, generated by our model or the baselines, we ask raters to select the box the text most likely refers to.

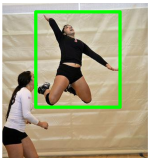
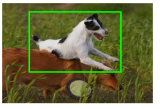
	(a) 	(b) 	(c) 
Schutz et al.[9]	Skateboarder	black dog	number 1 meter
Tanaka et al. [14]	Woman	barber barber barber barber	barber barber barber barber barber
Yu et al. [14]	Woman	camper goo goo mix raspberries goo mix raspberries	curlys curlys curlys curlys
DISCLIP-HPT (OURS)	**Woman jumping** as part of volleyball swing using **black** object.	**White dog** riding horse during a dog fight.	The baseballs of teams one **yellow white green orange**

Table S5: **Win cases from Flickr30k Entities.** In all these examples, people successfully chose the target object given our caption, but failed to do so, given the captions produced by the baselines.

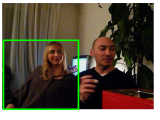
	(a) 	(b) 	(c) 
Schutz et al.[9]	younger woman	white shirt	white pants
Tanaka et al. [14]	woman	white shirt	barber barber barber barber barber
Yu et al. [14]	woman	white shirt	partial partial partial barely partial barely fingers curlys submerged
DISCLIP-HPT (OURS)	Woman smiling posing with multiple pairs of black neck shirts.	Woman standing eating	Person dressed shirt pants leg boots while holding phone.

Table S6: **Lose cases** from Flickr30K-Entities.

Schutz et al.[9]	shirt (782), man (635), red (503), white (475), black (374), blue (306), woman (298), green (197), blurry (166), dog (146)
Tanaka et al. [14]	barber (25081), shirt (456), man (366), white (222), red (149), black (140), closest (136), blue (128), woman (110), barely (109)
Yu et al. [14]	curlys (9191), loops (2257), goo (1815), raisins (1745), almonds (775), seed (691), blurry (686), shirt (499), man (424), mix (398)
DISCLIP (OURS)	white (591), black (591), holding (516), red (502), person (498), woman (487), large (423), man (416), young (371), close (360)

Table S7: Top 10 words in the generated REs for Flickr30K-Entities test split. (Overall 4601 Refs in 979 images)

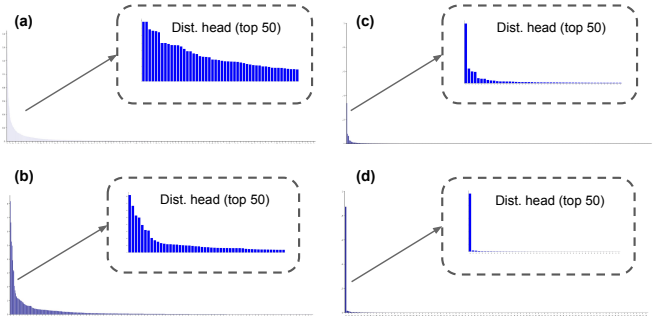


Figure S3: Distribution of words in the generated referring expression in Flickr30k-Entities by the different models: (a) DisCLIP (b) Schuz (c) Yu (d) Tanaka

G Datasets for Referring Expressions Generation

Below, we provide some more details about the datasets that were used in the scope of this paper. **(1) RefCOCO** [4] contains 142,209 referring expressions for 50,000 objects in 19,994 images. **(2) RefCOCO+** [4] contains 141,564 referring expressions for 49,856 objects in 19,992 images. This dataset focuses more on the appearance of objects. In both RefCOCO and RefCOCO+, Test A contains references to humans, and Test B references to other object types. **(3) RefCOCOg (Google RefExp)** [6] contains 85,474 referring expressions for 54,822 objects in 26,711 images and contains longer and more complex expressions.

(4) RefCLEF (ReferIt) [3] 10,000 images for training/validation and 10,000 for test, with 59,976 references in the train/val set and 60,105 in the test set. RefCLEF dataset is larger and more varied and is curated specifically *complex* photographs of real-world cluttered scenes. **(5) RefGTA** [12], contain 6563 samples in train/val and 6504 in test. Synthetic images are from the Grand Theft Auto (GTA) videogame. All referring expressions in RefGTA correspond to people. The focus is on relations expressions, since for a salient target, a brief description suffices. while less salient targets, require utilizing relationships with salient contexts around them to help tell their location. **(6) Flickr30k-Entities** [8], 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.

Figure S4: Images in RefCOCO/+/g are from MS-COCO dataset, but their textual annotations were designed to capture different types of referring expressions. RefCOCO is focused on spatial phrases, RefCOCO+ is attribute-based, and RefCOCOg provides long, rich, and diverse text.

References

[1] Desmond Elliott and Frank Keller. Image description using visual dependency representations. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1292–1302, 2013.

[2] 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.

[3] 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.

[4] 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.

[5] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, 2004.

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

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[8] 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.

[9] 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.

[10] 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.

[11] Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. A contrastive framework for neural text generation. *arXiv preprint arXiv:2202.06417*, 2022.

[12] 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.

[13] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[14] 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.