RECANTFormer: Referring Expression Comprehension with Varying Numbers of Targets

Bhathiya Hemanthage[1,2] **Hakan Bilen**[2] **Christian Dondrup**[1] **Phil Bartie**[1] **Oliver Lemon**[1]
[1]Heriot-Watt University [2]University of Edinburgh
{hsb2000, c.dondrup, phil.bartie, o.lemon}@hw.ac.uk {h.bilen}@ed.ac.uk

Abstract

The Generalized Referring Expression Comprehension (GREC) task extends classic REC by generating image bounding boxes for objects referred to in natural language expressions, which may indicate zero, one, or multiple targets. This generalization enhances the practicality of REC models for diverse real-world applications. However, the presence of varying numbers of targets in samples makes GREC a more complex task, both in terms of training supervision and final prediction selection strategy. Addressing these challenges, we introduce RECANTFormer, a one-stage method for GREC that combines a decoder-free (encoder-only) transformer architecture with DETR-like Hungarian matching. Our approach consistently outperforms baselines by significant margins in three GREC datasets.

1 Introduction

Referring expression comprehension (REC) focuses on generating an image bounding box tightly encompassing a region referred to by a natural language query. This is a core task in multi-modal information processing with potential to influence a wide range of applications including instruction following robots (Padmakumar et al., 2022; Gao et al., 2023), situated multi-modal dialogues (Kottur et al., 2021), and interactive photo editing (Jiang et al., 2021; Sharma et al., 2018). Despite advances in REC on datasets like RefCOCO/+/g (Yu et al., 2016; Nagaraja et al., 2016) current methods assume that a single referring expression always refers to a single object instance in the image. This simplification limits their real-world applicability, as they cannot handle expressions with no or multiple matching instances.

Several datasets; Generalized REC (He et al., 2023), Visual Query Detection (Acharya et al., 2019), and REF-ZOM (Hu et al., 2023) have been proposed to bridge this gap between real-world and classic REC datasets. Expressions in these datasets may refer to zero, one, or many instances in an image. (In this work, we adopt the term 'GREC' coined by He et al. (2023) to refer to this family of tasks). Despite the existence of suggested datasets and their corresponding baselines based on state-of-the-art classic REC models, no models have been developed specifically addressing REC with varying numbers of targets, including zero targets.

The GREC task is more challenging than classic REC, where an image with n referable objects results in n distinct targets in classic REC, as opposed to 2^n distinct combinations of objects in GREC. Unlike classic REC datasets, which are susceptible to models exploiting biases (Cirik et al., 2018) the large pool of possible distinct combinations in the GREC task makes it extremely difficult for models to exploit biases. Furthermore, a top-1 selection strategy (or top-k variant) which is prevalent in classic REC models with the one-to-one assumption (Yan et al., 2023; Deng et al., 2021), is unsuitable when there are varying numbers of targets. A confidence-score threshold-based selection strategy is a viable alternative. However, we demonstrate that the threshold-based approach also leads to a significant drop in performance when current REC models are trained and evaluated on GREC datasets.

To overcome these limitations in classic REC methods, we introduce RECANTFormer: a transformer-based framework for *R*eferring *E*xpression *C*omprehension with V*a*rying *N*umber of *T*argets. To address the challenge of training the model with a varying number of targets, we leverage Hungarian matching loss similar to DETR (Carion et al., 2020), where bipartite matching is calculated between the set of proposed boxes and ground truth boxes. However, differing from DETR, which is a transformer encoder-decoder based approach for object detection, RECANTFormer only employs transformer encoders with

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 21784–21798
November 12-16, 2024 ©2024 Association for Computational Linguistics

simple MLP-based prediction heads. Inspired by the success of using a separate token for grounding (Deng et al., 2021), to allow for multiple potential targets in GREC, we propose a multimodal transformer encoder with multiple learnable localization tokens. Our selection of an decoder-free architecture is driven by the training-inefficient nature of encoder-decoder based DETR-like architectures, as shown in (Chen et al., 2022; Ding et al., 2023). (Also our preliminary experiments in Appendix E support this argument.)

To summarize, our main contribution is introducing RECANTFormer, a transformer-based one-stage framework for GREC. To our knowledge, it is the first model to learn and infer varying numbers of bounding boxes in GREC. Additionally, this is the pioneering work adapting DETR-like Hungarian-matching to an encoder-only architecture for a multimodal task. Our method significantly outperforms state-of-the-art REC methods on three GREC benchmarks and achieves comparable performance on classical REC datasets.

2 Related Work

Classic REC techniques are primarily categorized into two-stage methods, such as (Yu et al., 2018; Hong et al., 2019; Liu et al., 2019), which use a Region Proposal Network to generate candidates, and one-stage methods (Yang et al., 2019, 2020; Huang et al., 2021) that offer a more efficient, end-to-end approach. Recent advances integrate transformers (Vaswani et al., 2017), facilitating multimodal integration, with models like RECANTFormer exemplifying transformer-based one-stage methods trained on task-specific data without visual language pretraining. Unlike models that map expressions to a single region, RECANTFormer can interpret expressions correlating to multiple or no regions. Additionally, although leveraging vision-language pre-training (VLP) has proven beneficial for REC, as demonstrated by models like UNITER (Chen et al., 2020), MDETR (Kamath et al., 2021), and Universal (Yan et al., 2023), RECANTFormer outperforms these VLP-based methods without requiring extensive visual-language data.

Generalized REC: Despite several datasets available (Acharya et al., 2019; He et al., 2023), prior research has not specifically targeted GREC. To the best of our knowledge, our model, RECANTFormer, is the first to focus on this task.

There are several tasks related to GREC with key differences. *Phrase localization* in Flickr30K Entities (Plummer et al., 2015) aims at localizing each noun phrase in a given image with a set of bounding boxes. This task differs from (G)REC in two ways: 1) In both REC and GREC, the entire referring expression must be considered, requiring more sophisticated reasoning over a language query. 2) Evaluation protocols of phrase localization models avoid one expression and many targets scenarios (see appendix for more details). Tasks like *phrase detection* (Plummer et al., 2020) and *open vocabulary object detection* (OVD) also limit language expressions to simple noun phrases. Furthermore, these tasks consider a large, yet finite number of categories, whereas free-form language in REC results in an infinite number of potential categories.

DETR-based Detection There is a body of work that is built on DETR, most of which focuses on further improving DETR for object detection (Liu et al., 2022a,b; Zhang et al., 2023) while a few works (Kamath et al., 2021; Chu and Lee, 2023) have used DETR for multimodal settings. All these works use a full encoder-decoder architecture similar to DETR. Ding et al. (2023) and Chen et al. (2022) have investigated decoder-free DETR, emphasizing the training inefficiency and slow convergence in the encoder-decoder architecture. However, these works focus on language-agnostic object detection, in contrast to the multimodal setting of RECANTFormer.

3 Method

3.1 Model Architecture

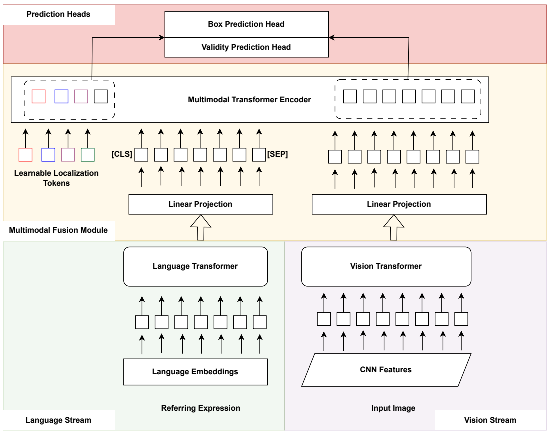


Figure 1: An overview of the proposed RECANTFormer framework consisting of 1) Language Stream, 2) Vision Stream, 3) Multi-modal Fusion module that leverages Learnable Localization Tokens, 4) Prediction Heads.

As illustrated in Figure 1, our method takes in two input streams for vision (in purple) and language (in green), and employs a multi-modal fusion module (in yellow), which includes a multi-modal transformer encoder that serves as the core of the RECANTFormer architecture.

Vision Stream: The vision stream consists of a convolution layer followed by a 6 layer transformer encoder. The transformer encoder in the vision stream extracts embeddings that are capable of capturing spatially long-range correlations in the image. This is particularly crucial for GREC, as resolving most queries (e.g., 'two individuals on the outermost sides') necessitates modeling long-range interactions between different image patches.

Given an image with dimensions $H \times W$, we utilize our backbone, ResNet-50, to generate a lower-resolution activation map of dimensions $C \times H/32 \times W/32$, where C (=2048) is the channel dimension. A 1×1 convolution layer then reduces the channel dimension to C_v (=256). The resulting vector is flattened to obtain $H/32 \times W/32$ tokens, with a hidden dimension of C_v. These token vectors are taken as input by the visual encoder, which outputs a vector of the same dimensions. Considering the 2D nature of the visual features, sine positional encoding is used.

Language Stream To encode text, we employ a pre-trained transformer language model, $\text{BERT}_{\text{base}}$ (Devlin et al., 2018) model.

Multi-modal Fusion Module: The objective of this module is to facilitate cross-modality reasoning by word embeddings attending to features of image patches, and vice versa. As shown in Figure 1, the multi-modal fusion module consists of two linear projection layers with one layer from each stream. This is followed by a multi-modal transformer with 6 encoder layers. In addition to the linear projections, we prepend the set of learnable localization tokens to the multi-modal transformer.

Learnable Localization Tokens: Inspired by prior object detection work (Carion et al., 2020) and classic REC (Deng et al., 2021; Ho et al., 2022), we introduce fixed number of learnable tokens (initialized randomly) with a specific focus on object localization. Essentially, each token is designed to correspond to a distinct region in the image. In contrast to REC, where language expressions consistently map to a single region in the image, GREC models require tracking multiple potential targets.

Prediction Heads: RECANTFormers consist of two parallel prediction heads, which take output states of the localization tokens as the input. A bounding box head predicts a fixed number of bounding boxes $N * 4$ with N usually larger than the number of referenced objects. However, only a subset of these coordinates predictions are valid for a given image-text pair. To determine the valid subset of coordinate predictions, a validity prediction head, which predicts the validity of each of N bounding box predictions, is trained in parallel. Both prediction heads are implemented as 3-layer MLPs with ReLU activation.

3.2 RECANTFormer Training Objectives

In our method, similar to the approach used in DETR (Carion et al., 2020), we use Hungarian matching loss with bipartite matching to assign each ground truth bounding box with a unique predicted bounding box from N predictions made by a bounding box head. Predictions with a matched ground truth bounding box are supervised with the corresponding ground truth as the target. A linear combination of L_1 loss and scale invariant Generalized IoU (GIoU) loss (Rezatofighi et al., 2019) is used. The rest of the boxes without a matching ground truth bounding box are labeled as negatives for the validity classification head. Standard cross-entropy loss is used for supervising validity label prediction. In the case of no-target examples, a bounding box of all zeros ($[0, 0, 0, 0]]$) is used as target, while the validity classification head is supervised to predict an invalid label for all the predicted bounding boxes. (More detail on the loss function is provided in Appendix B)

4 Experiments

4.1 Datasets

We conduct our experiments on GREC with three datasets: VQD (Acharya et al., 2019), gRefCOCO (He et al., 2023), and Ref-ZOM (Hu et al., 2023).(See appendix for more details.) We also evaluate RECANTFormer on three standard REC datasets: RefCOCO/+/g.

4.2 Evaluation Metrics

Precision@(F1=1, IoU $\geq$ 0.5) is used to assess the performance in the GREC task as proposed by He et al. (2023). For the VQD dataset, we also report standard **PASCAL VOC AP$^{\textbf{IoU=.5}}$** : from object detection. (Appendix D provides a detailed discussion on evaluation metrics)

5 Results

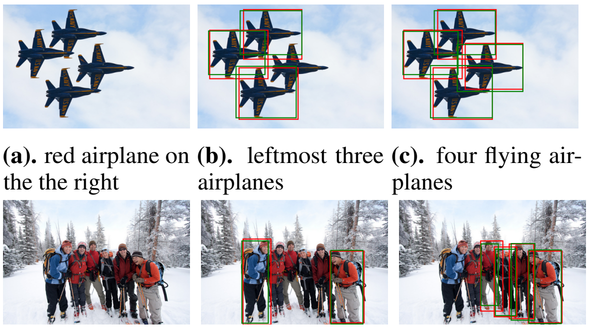
Method	GREC TestA	GREC TestB	REF-ZOM Test	VQD Test
Models Without VL pretraining				
MCN†	32.3	26.8	-	-
VLT†	40.2	30.2	-	-
RESC(L)-MT	20.52	22.47	17.18	45.18
RECANTFormer(5)	**57.82**	**49.49**	56.69	-
RECANTFormer(10)	55.07	48.01	**59.78**	**63.18**
MLLM Zero-shot Evaluation				
KOSMOS-2	22.06	15.96	44.33	21.64
Models With VL pretraining				
MDETR†	50.0	36.5	56.96	-
UNINEXT†	46.4	42.9	-	-

Table 1: Comparison of RECANTFormer performance on 3 datasets with baseline models. For all the compared methods, bounding box predictions are selected using a threshold of 0.7. †: Baselines as reported in (He et al., 2023). Number in parenthesis indicate the # of localization tokens.

Method	$\mathrm{AP}^{IoU=.5}$
DETECT	26.94
Vision+Query	31.03
RECANT(10)	**38.60**

Table 2: Comparison of RECANTFormer results on VQD dataset baselines. Pascal VOC $AP^{IoU=.5}$ is reported.

Generalized REC Table 1 compares the performance of RECANTFormer with 3 types of baselines. First, VLT (Ding et al., 2022), MCN(Luo et al., 2020) and RESC Large(Yang et al., 2020) baselines are strictly trained on the training split of the specific dataset without using any additional VLP data. This is similar to the setting followed by RECANTFormer. It can be seen that our model RECANTFormer outperforms the non-pretrained baselines by a significant margin. Second, we evaluate the GREC datasets on a Multimodal Large Language Model (MLLM), Kosmos-2, in a zero-shot manner. Despite reporting zero-shot accuracy over 50 on classic REC datasets, Kosmos-2 demonstrates poor performance on the GREC datasets. Third, we report results for models subscribing to the *pretrain, then finetune* strategy, which involves pre-training on a large visual-language corpus. For example, MDETR is pre-trained on 1.3M images taking approximately 5300 GPU hours, whereas UNINEXT is trained on 2M images taking 3000 GPU hours. Despite using limited data and compute resources, RECANTFormer outperforms both MDETR and UNINEXT baselines by a significant margin on the gRefCOCO dataset. Furthermore, table 2 reports the standard Pascal VOC $AP^{IoU=.5}$ scores with baselines reported in Acharya et al. (2019).



(a). red airplane on the the right **(b).** leftmost three airplanes **(c).** four flying airplanes

(d). guy sitting **(e).** two individuals on the outermost sides **(f).** the four individuals counted from the right.

Figure 2: Example results of our method on the gRefCOCO dataset. If exist, predicted boxes and ground truth boxes are shown in green and red colors respectively

Qualitative Examples Figure 2 shows some qualitative examples of the RECANTFormer model on the gRefCOCO dataset. The model demonstrates its ability to differentiate objects based on color in identifying the absence of a "red airplane" in Figure 2a. Figure 2d presents a no-target sample that demands the model to differentiate objects based on an action noun ("sitting"). Multi-target samples in Figures 2b, 2c, 2e and 2f use counting words ("two", "three", and "four") when referring to objects. Figures 2b, 2e and 2f requires the model to comprehend spatial adjectives ("leftmost", "outermost", "from the right") in referring expressions.

Methods	RefCOCO testA	RefCOCO testB	RefCOCO+ testA	RefCOCO+ testB	RefCOCOg val
CNN Based					
FAOA	74.35	68.50	60.23	49.60	56.12
ReSC-Large	80.45	72.30	68.36	56.81	63.12
Transformer Based					
TransVG	82.67	**78.12**	68.15	55.63	**66.56**
VGTR	82.09	73.31	69.65	55.33	62.88
RECANTFormer(1)	**83.08**	76.51	**70.43**	**58.08**	65.40

Table 3: Comparison of RECANTFormer with state-of-the-art methods on classic datasets; RefCOCO, RefCOCO+, RefCOCOg

Classic REC Results in Table 3 indicates that RECANTFormer achieves superior or comparable performance to state-of-the-art REC models on classic REC tasks. This is despite not resolving the one-to-one assumption, which significantly eases the task for baseline models.

Ablation Study An evaluation on Table 4 shows the impact of localization tokens N on GREC performance. Increasing N from 5 to 10 in gRefCOCO decreased Pr@(F1=1) by 2%. In Ref-ZOM, $N = 10$ slightly improved by 0.09% over $N = 5$, but $N = 20$ declined over 2%. For RefCOCO (classic REC), $N = 1$ vs $N = 5$ differed by over 5%. We hypothesize that this behavior is attributed to the diluted loss signal caused by most the predicted boxes remaining unassigned during Hungarian matching.

N	gRefCOCO Val	Ref-ZOM Test	RefCOCO Val
1	-	-	**81.30**
5	**57.73**	59.69	76.06
10	55.10	**59.78**	-
20	54.27	56.40	-

Table 4: Variation of performance in gRefCOCO, Ref-ZOM, and RefCOCO datasets with the number of localization tokens.

6 Conclusion

This paper presents RECANTFormer, the first framework focused on the challenging task of Generalized Referring Expression Comprehension (GREC). RECANTFormer has a simple, decoder-free transformer-based architecture and demands minimal visual-language training data. Our model effectively utilizes the powerful multimodal fusion capabilities of transformers encoders to outperform GREC benchmarks across 3 datasets. By effectively handling referring expressions with a varying number of target objects, including no-target scenarios, RECANTFormer expands the range of applications for REC.

Limitations

Detecting Hard Negatives Notwithstanding its substantial improvement over baselines, RECANTFormer's performance demonstrates a marked deterioration in the face of challenging negative samples. Further elaboration is provided in Table 5, which presents RECANTFormer's accuracy in processing samples with no targets (N-acc) across various datasets. Upon comparing the results across datasets, it becomes apparent that the N-acc value on the gRefCOCO dataset is significantly lower than that of the other two datasets, attributed to the presence of difficult negative examples.

Supervised Learning We employ a fully supervised setup for training RECANTFormer. Given the considerable annotation cost associated with creating GREC datasets, we consider a fully supervised setup to be a significant constraint for GREC. We believe that a semi-supervised setup (Ouali et al., 2020; **?**), leveraging both unannotated and annotated data, offers a promising direction for future research.

		GREC		REF-ZOM	VQD
Method	Val	TestA	TestB	Test	Test
RECANT(5)	52.70	53.38	54.53	88.24	-
RECANT(10)	52.73	53.07	54.81	88.24	94.16

Table 5: No-target accuracy of the models across datasets.

Ethical Statement

All the datasets used in this study have been previously published. Since the GREC task that we address is a core skill in multimodal information processing, this work has the potential to impact wide range of important applications such as voice controlled autonomous driving, social robots, multimodal dialogue agents, and interactive photo editing. However the capabilities of these models may be used for harmful applications such as surveillance without consensus and illegal information retrieval from images, which must be addressed.

Computational Budget

Compute budget for the entire research is around 4000 GPU hours. This includes, failed experiments, hyper-parameter tuning, ablation studies and training baseline methods. We mainly used NVIDIA A100 GPUs with 80GB of GPU memory for training alongside NVIDIA 2080RTX GPUs with 16GB GPU memory. Our infrastructure facilitate maximum of 4GPUs per job.

Use of AI

AI assistants were not utilized for the research or coding; however, they were employed to enhance the writing in certain paragraphs of the paper.

Acknowledgements

This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk/).This work also used the Cirrus UK National Tier-2 HPC Service at EPCC funded by the University of Edinburgh and EPSRC (EP/P020267/1).

References

Manoj Acharya, Karan Jariwala, and Christopher Kanan. 2019. Vqd: Visual query detection in natural scenes. *arXiv preprint arXiv:1904.02794*.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *ECCV (1)*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer.

Peixian Chen, Mengdan Zhang, Yunhang Shen, Kekai Sheng, Yuting Gao, Xing Sun, Ke Li, and Chunhua Shen. 2022. Efficient decoder-free object detection with transformers. In *European Conference on Computer Vision*, pages 70–86. Springer.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Shih-Yun Chu and Ming-Sui Lee. 2023. Mt-detr: Robust end-to-end multimodal detection with confidence fusion. In *WACV*, pages 5241–5250. IEEE.

Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. 2018. Visual referring expression recognition: What do systems actually learn? *arXiv preprint arXiv:1805.11818*.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. 2021. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. 2022. Vlt: Vision-language transformer and query generation for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tonghe Ding, Kaili Feng, Yanjun Wei, Yu Han, and Tianping Li. 2023. Deot: an end-to-end encoder-only transformer object detector. *Journal of Real-Time Image Processing*, 20(1):1.

Qiaozi Gao, Govind Thattai, Xiaofeng Gao, Suhaila Shakiah, Shreyas Pansare, Vasu Sharma, Gaurav Sukhatme, Hangjie Shi, Bofei Yang, Desheng Zheng, et al. 2023. Alexa arena: A user-centric interactive platform for embodied ai. *arXiv preprint arXiv:2303.01586*.

Shuting He, Henghui Ding, Chang Liu, and Xudong Jiang. 2023. GREC: Generalized referring expression comprehension. *arXiv preprint arXiv:2308.16182*.

Chih-Hui Ho, Srikar Appalaraju, Bhavan Jasani, R Manmatha, and Nuno Vasconcelos. 2022. Yoro-lightweight end to end visual grounding. In *European Conference on Computer Vision*, pages 3–23. Springer.

Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. 2019. Learning to compose and reason with language tree structures for visual grounding. *IEEE transactions on pattern analysis and machine intelligence*, 44(2):684–696.

Yutao Hu, Qixiong Wang, Wenqi Shao, Enze Xie, Zhenguo Li, Jungong Han, and Ping Luo. 2023. Beyond one-to-one: Rethinking the referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4067–4077.

Binbin Huang, Dongze Lian, Weixin Luo, and Shenghua Gao. 2021. Look before you leap: Learning landmark features for one-stage visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16888–16897.

Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. 2021. Talk-to-edit: Fine-grained facial editing via dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13799–13808.

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

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017a. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017b. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Chang Liu, Henghui Ding, and Xudong Jiang. 2023. GRES: Generalized referring expression segmentation. In *CVPR*.

Chang Liu, Xudong Jiang, and Henghui Ding. 2022a. Instance-specific feature propagation for referring segmentation. *IEEE Transactions on Multimedia*.

Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. 2022b. Dab-detr: Dynamic anchor boxes are better queries for detr. In *ICLR*. OpenReview.net.

Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. 2019. Improving referring expression grounding with cross-modal attention-guided erasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1950–1959.

Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. 2020. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 10034–10043.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee.

Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. 2016. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 792–807. Springer.

Yassine Ouali, Céline Hudelot, and Myriam Tami. 2020. An overview of deep semi-supervised learning. *arXiv preprint arXiv:2006.05278*.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2022. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Bryan A Plummer, Kevin J Shih, Yichen Li, Ke Xu, Svetlana Lazebnik, Stan Sclaroff, and Kate Saenko. 2020. Revisiting image-language networks for open-ended phrase detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(4):2155–2167.

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

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666.

Shikhar Sharma, Dendi Suhubdy, Vincent Michalski, Samira Ebrahimi Kahou, and Yoshua Bengio. 2018. Chatpainter: Improving text to image generation using dialogue. *arXiv preprint arXiv:1802.08216*.

Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. 2016. End-to-end people detection in crowded scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2325–2333.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. 2023. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15325–15336.

Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. 2020. Improving one-stage visual grounding by recursive sub-query construction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 387–404. Springer.

Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. 2019. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4683–4693.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1307–1315.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. 2023. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *ICLR*. OpenReview.net.

A Baselines

A.1 No Additional Data Setting

Modified RESC(Large) : RESC(Yang et al., 2020) is a CNN based one stage approach proposed for classic REC. RESC build on FAOA (Yang et al., 2019), which is developed with the main idea of fusing VOLVO3 features with text query embedding. RESC improves FAOA perfomance on classic REC with recursive subquery construction. Original RESC model assume with one expression one region. Therefore, during training process model select the best matching anchor prediction (specifically; using Softmax over all the anchor predictions) and calculate regression losses against the single target bbox. To facilitate many targets scenario, replace this loss calculation with YoloV5 loss function with IoU based loss calculation. Loss calculation w.r.t recursive sub-query construction is kept unchanged.

A.2 Zero-shot Setting

We use (Peng et al., 2023) as a zero-shot evaluation baseline. Kosmos-2 is a grounded Multimodal Large Language Model (MLLM) trained on 115M text spans over 90M images. Model record a zero-shot accuracy scores over 60 on RefCOCOg splits in classic REC task.

A.3 Pretrain-finetune Setting

We report the results of MDETR and UNINEXT from (He et al., 2023). In the case of MDETR, fine-tuning process follows the same pre-trained checkpoint and procedure as classic REC tasks. Initially, the training dataset is preprocessed using Spacy to identify the roots of the referring expressions, and then the model is fine-tuned on the pre-processed data. In UNINEXT (Yan et al., 2023), the stage-1 pre-train checkpoint is fine-tuned to avoid data leakage.

B Training Objectives

RECANTFormer follow Hungarian matching based calculation similar to DETR(Carion et al., 2020) object detection. This section presents details of loss calculation for completeness.

In RECANTFormer, the bounding box head always predicts N$_{loc}$ boxes for a given sample. Additionally, the validity head predicts N$_{loc}$ predictions in parallel with a validity label for the corresponding to each predicted box. The aim is to evaluate these predictions considering the varying number of ground truth bounding box targets. The loss calculation involves two steps: 1) Matching the predictions to the ground truth targets using the Hungarian Algorithm based on similarity. 2) Calculating the losses of the validity labels and predicted bounding boxes based on the assigned ground truth boxes from step 1, if any.

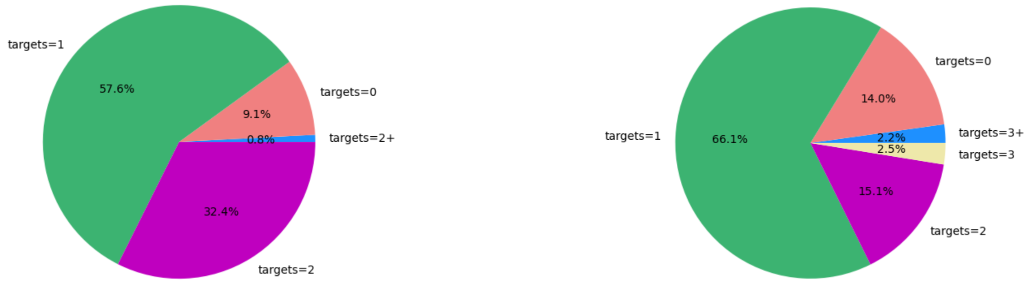
Step-1: Target-Prediction Matching We represent predicted bounding boxes as $\hat{Y}_{bbox} \in N_{loc} \times 4$ and the predicted validity scores as $\hat{Y}_{val} \in N_{loc} \times 2$. Similarly, we denote the set of ground truth bounding boxes asY_{bbox}. Assuming N$_{loc}$ larger than the number of ground truth bounding boxes, we pad ground truth bounding boxes so that $Y_{bbox} \in N_{loc} \times 4$. When generating target validity labels Y_{val}, we assign an invalid label to the positions with padded boxes, while marking the remaining (actual) ground truth positions as valid. To find a bipartite matching between the sets; $Y = (Y_{bbox}, Y_{val})$ and $\hat{Y} = (\hat{Y}_{bbox}, \hat{Y}_{val})$; we search for a permutation of N_{loc} elements $\sigma \in \mathfrak{S}_{N_{loc}}$ with the lowest cost:

$$\hat{\sigma} = \arg\min_{\sigma \in \mathfrak{S}_{N_{loc}}} \sum_{i}^{N_{loc}} \mathcal{L}_{match}(Y_i, \hat{Y}_{\sigma(i)}) \quad (1)$$

where $\mathcal{L}_{match}(Y_i, \hat{Y}_{\sigma(i)})$ is the pair-wise matching cost between ground truth Y_i with the prediction at index $\sigma(i)$. Following prior work (Carion et al., 2020; Stewart et al., 2016) we use Hungarian Algorithm for calculating optimal assignment. Matching cost calculation take both validity labels and the bounding box similarity between ground truth and predictions into account. After the first step, each ground truth bounding box at index i is matched with the prediction at index $\hat{\sigma}(i)$.

Step-2: Loss Calculation In the second step, predicted bounding boxes with a matching is evaluated against corresponding ground truth bounding box assigned in step-1. In addition, validity prediction loss is calculated between N_{loc} predictions and generated ground truth labels (including padded position). Hungarian loss can be denoted as:

$$\mathcal{L}_{Hungarian}(Y, \hat{Y}) = \sum_{i=1}^{N_{loc}} -\log \hat{p}_{\hat{\sigma}(i)}(Y_{val}(i)) + \sum_{i=1, Y_{val}(i)=valid}^{N_{loc}} \mathcal{L}_{bbox}(Y_{bbox}(i), \hat{Y}_{bbox}(\sigma(i))) \quad (2)$$



(a). Distribution of 209344 gRefCOCO training examples

(b). Distribution of 68429 Ref-ZOM training examples

Figure 3: Distribution of data in train splits of gRefCOCO and Ref-ZOM w.r.t number of ground truth targets

Specifically, $\mathcal{L}_{bbox}$ is a linear combination of L1 loss and GIoU loss. Removing the inputs for simplicity:

$$\mathcal{L}_{bbox} = \lambda_{L1}\mathcal{L}_{L1} + \lambda_{GIoU}\mathcal{L}_{GIoU} \tag{3}$$

Note that in eq. (2) $\mathcal{L}_{bbox}$ is calculated only if there is a valid ground-truth bounding box.

B.1 RECANTFormer+: Extension for GRES

When extending RECANTFormer loss calculation, matching step remain unchanged. Therefore same matching indexes are used. In addition to the two components in eq. (2), joint training for GREC and GRES includes $\mathcal{L}_{segm}$ where:

$$\sum_{i=1, Y_{val}(i)=valid}^{N_{loc}} \mathcal{L}_{segm}(Y_{segm}(i), \hat{Y}_{segm}(\sigma(i)) \tag{4}$$

$\mathcal{L}_{segm}$ is a linear combination of Focal loss and DICE loss:

$$\mathcal{L}_{segm} = \lambda_{Focal}\mathcal{L}_{Focal} + \lambda_{DICE}\mathcal{L}_{DICE} \tag{5}$$

C Datasets

We conduct our experiments on GREC using 3 datasets: VQD (Acharya et al., 2019), gRefCOCO(Liu et al., 2023) and Ref-ZOM(Hu et al., 2023). Referring expressions across all datasets are in English. In this section we provide statistics of these 3 datasets.

In addition to the gRefCOCO and Ref-ZOM datasets, we also evaluated RECANTFormer on three mainstream referring expression comprehension datasets: RefCOCO, RefCOCO+ (Yu et al., 2016) and RefCOCOg (Nagaraja et al., 2016). For RefCOCO and RefCOCO+, we used the UNC partition, while for RefCOCOg, we used the Google partition.

C.1 Dataset Links

We used gRefCOCO and Ref-ZOM datasets for our experiments in GREC and GRES tasks. Both the datasets are publicly available and downloaded links are provided at following git repositories:

- gRefCOCO: https://github.com/henghuiding/gRefCOCO
- VQD: https://github.com/manoja328/VQD_dataset
- Ref-ZOM: https://github.com/toggle1995/RIS-DMMI

C.2 Dataset Statistics

Table 6 provide number of expressions (image-text pairs) in different splits in gRefCOCO and Ref-ZOM datasets. In appendix A.3 breaks-down training split of each dataset by the number of ground truth targets for given image-expression pair. Note that 99% of gRefCOCO training examples have zero target, single target or two-targets. In the case of Ref-ZOM, just over 95% samples have two or less ground truths, while 2.5% of training examples having three ground-truth targets.

D Discussion on Evaluation Metrics

In this section, we discuss the selection of Pr@(F1=1, IoU $\geq$ 0.5) as the evaluation metrics against several other alternatives.

In traditional REC research, where expressions always correspond to a single object instance and thus a single bounding box, top-1 accuracy is commonly used as a metric. The predicted bounding

	GREC				REF-ZOM	VQD		
Sample Category	Train	Val	TestA	TestB	Train	Test	Train	Test
Total Samples	209344	16870	18712	14933	68249	21770	431363	190174
Zero-target	19140	6966	4189	4242	9610	2327	161494	80025
Multi-target	69580	5905	8835	5744	13601	7387	55148	20048

Table 6: Number of image-text pairs in gRefCOCO and Ref-ZOM dataset splits.

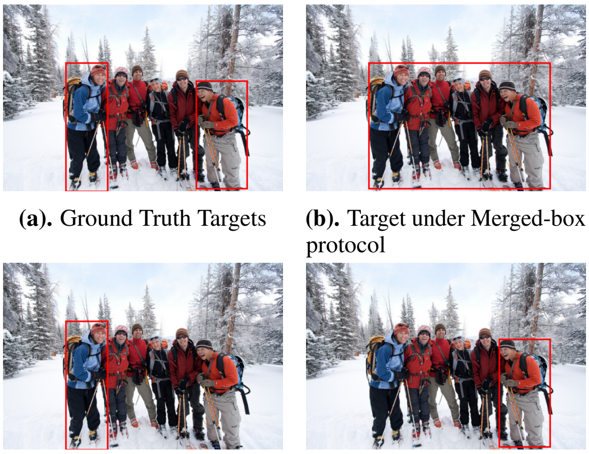
box with the highest confidence is compared to the ground-truth bounding box, and the prediction is deemed correct if the Intersection over Union (IoU) between the two bounding boxes exceeds a specified threshold (typically 0.5). However, this approach is not applicable when the number of ground-truth bounding boxes is unknown in advance and can vary, including cases where there are zero, one, or multiple target bounding boxes.

While the zero-target case is not taken into account, efforts have been made in phrase grounding research to address scenarios where multiple ground truth bounding boxes exist. The primary evaluation metric proposed for assessing grounded detection datasets, such as Flickr30K entities(Plummer et al., 2015), is Recall@k. However, Recall@k is not adequately defined for cases involving multiple boxes. To overcome this limitation, prior work on phrase grounding has introduced two distinct protocols: the Merge-box protocol (Deng et al., 2021; Liu et al., 2019; Yang et al., 2019) and the Any-box protocol (Li et al., 2019) (as referred to in (Kamath et al., 2021)).

Merge-box protocol In the Merge-box protocol, all the ground truth bounding boxes corresponding to a given phrase are merged to create the smallest enclosing bounding box. This resulting bounding box is then considered the target for evaluation.

Any-box protocol In the Any-box protocol, a model prediction is deemed correct if it has an Intersection over Union (IoU) higher than the specified threshold (0.5) with any of the ground truth bounding boxes.

As evidenced by fig. 4, both evaluation approaches suffer from significant drawbacks. The merged-box protocol, for instance, sacrifices fine-grained details to an extent that undermines semantic correctness in GREC. This is demonstrated in fig. 4b, where the resulting bounding box encompasses all individuals instead of solely capturing those on the outermost sides. Meanwhile, the any-box protocol fails to assess whether all instances referred to in the expressions are correctly identified. As illustrated in figs. 4c and 4d, identifying any of the individuals would suffice under this protocol, which is problematic given that the expression explicitly references "two individuals." By contrast, Pr@(F1=1, IoU=0.5) represents a more stringent measure that demands fine-grained predictions while still preserving semantic correctness in terms of the number of identified regions, as compared to the aforementioned protocols.



(a). Ground Truth Targets

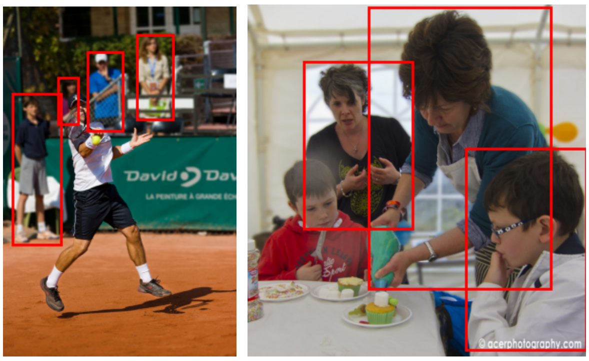
(b). Target under Merged-box protocol

(c). Sufficient target under Any-box protocol

(d). Sufficient target under Any-box protocol

Figure 4: Merged-box and any-box evaluation scenarios for 'two individuals on the outermost sides'

Furthermore, object detection research relies on metrics such as Average Precision (AP), which involve a trade-off between recall and precision. However, in the context of GREC, it is possible for a model to achieve high recall and precision scores while lacking a proper understanding of the expression's semantics. For example, as depicted in fig. 5, a model that selects every person in the image would attain perfect recall and high precision. Nevertheless, it is important to note that these high precision/recall scores conceal the model's failure to comprehend the underlying expression.



(a). four person in the background appearing fuzzy

(b). everyone except the kid in red

Figure 5: 'Two examples where model can achieve perfect recall and high precision by selecting every person, while failing to understand the expression '

E Preliminary Experiments

In our initial experiments, we trained the MDETR architecture on the Ref-ZOM dataset, which is encoder-decoder based, without using pretrained weights. We found that the non-pretrained MDETR model did not perform well on the GREC task after 25 epochs of training (approximately 30 hours) on 4 NVIDIA A100 GPUs, yielding a precision score of only 10.69. In contrast, the RECANTFormer model achieved a Pr@(F1=1) score of 55.74 after 25 epochs in approximately 12 hours. Due to the inefficient use of compute resources, we discontinued experiments with non-pretrained versions. Our intuition suggests that training an encoder-decoder model with cross-attention requires more resources (data and compute) compared to an encoder-only approach. In general, we believe that encoder-only DETR-based models show promise for further investigation, especially in low-resource settings.

F Implementation Details

Our model is trained using the AdamW optimizer. The multimodal fusion module has an initial learning rate of 1e-4, while the vision and language streams have learning rates of 1e-5 and a weight decay of 1e-4. We initialize the backbone and vision encoder using weights from a DETR model encoder (Carion et al., 2020), which was trained on COCO images excluding those in the test/val splits of respective datasets. The language stream is initialized with the BERT$_{\text{base}}$ model (Devlin et al., 2018). We use Xavier initialization for the weights in the multimodal fusion module. Data augmentation follows prior work (Deng et al., 2021), but we exclude random horizontal flipping due to semantic ambiguity. Additionally, random cropping is not used when training on Generalized REC datasets (gRefCOCO. Ref-ZOM and VQD). Images are scaled so that the longest side is 800 pixels, and the language stream uses a maximum of 40 language tokens. We train the model for 90, 90, 40 epochs on gRefCOCO, Ref-ZOM and VQD experiments respectively. For all the classic REC tasks, we train the model for 180 epochs. The learning rate decreases by a factor of 10 after 60 epochs in all experiments.

G Ablation

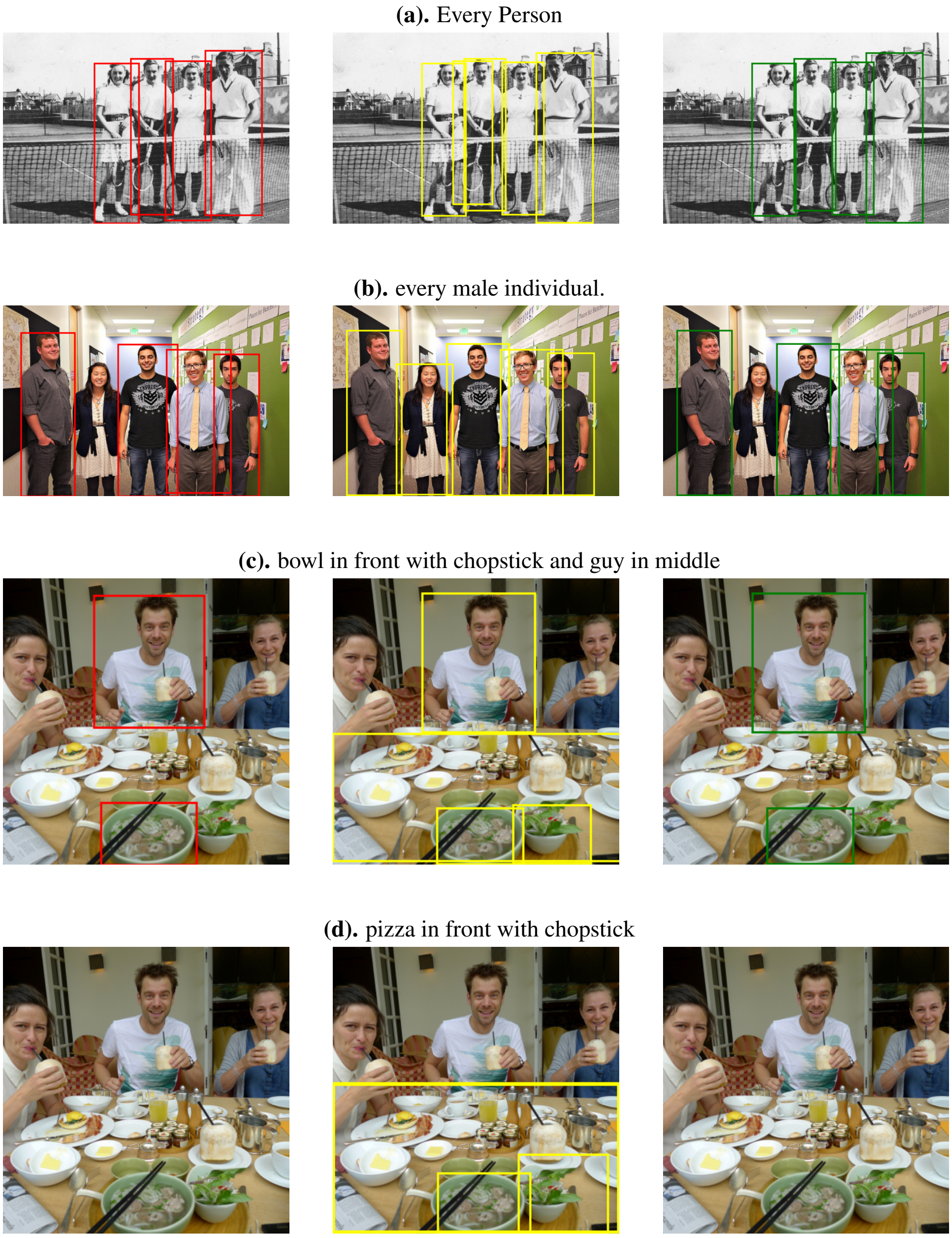
We use gRefCOCO validation set to ablate our choice of bounding box loss components and report results in Table 7.

L1	GIoU	Pr@(F1=1)
✗	✓	56.53
✓	✗	57.44
✓	✓	**57.73**

Table 7: Ablation results of RECANTFormer(5) on gRefCOCO validation set with different bounding box loss components.

H More on Localization Tokens

fig. 6 provides more examples of the RECANTFormer model predictions. In addition to the final set of predicted bounding boxes (shown in column 3 with green bounding boxes), the second column illustrates the predicted regions by the bounding box head without applying validity filtering. In fig. 6a, where there are four persons present, five *probable bounding boxes* predicted by box head includes two highly overlapping regions (around the second person from the left). The validity head correctly selects the best set of bounding boxes, predicting a single bounding box covering each person. In fig. 6b where *'every male individual.'* are referred, box head predicts a box around each of the person including the woman. Validity prediction head filter outs the female individual and correctly select the four target objects for the given expression. Similar behaviour, where bounding box prediction head select set of probable regions for the validity head to filter-out regions irrelevant to the expression, can be also seen in figs. 6c and 6d.



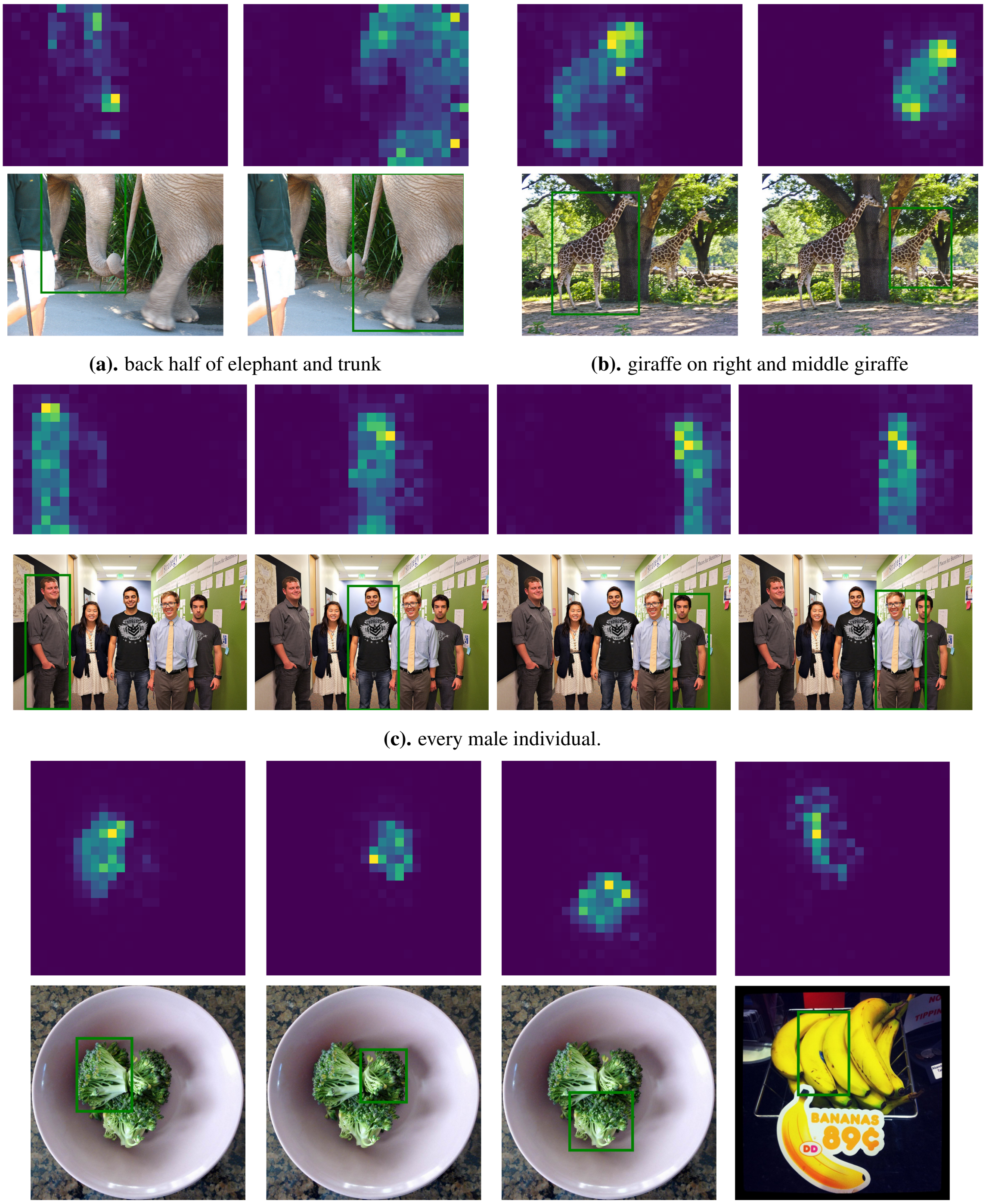
(a). Every Person

(b). every male individual.

(c). bowl in front with chopstick and guy in middle

(d). pizza in front with chopstick

Figure 6: GREC examples with regions detected before and after validity filtering. First column shows ground truth bounding boxes in red. Yellow boxes in second column shows all the bounding boxes from box head without applying validity filtering. Last column with green bounding boxes shows final prediction of the model after filtering



(a). back half of elephant and trunk

(b). giraffe on right and middle giraffe

(c). every male individual.

(d). every single broccoli floret.

(e). banana second to left

Figure 7: Attention weights of output state of valid localization tokens to output states of tokens representing visual features in multimodal tranformer encoder. RECANTFormer checkpoint only trained on GREC task is used. Under each visualization of weights, is the bounding box predicted by the particular localization token.

I Attention Weights of Localization Tokens

We hypothesize that the output state of the localization tokens within the multimodal transformer encoder contains crucial information necessary for the generation of a segmentation mask that extends beyond predicting bounding box coordinates. To validate this intuition, we visualize attention weights of the output state of valid localization tokens in relation to the output states of the tokens that represent visual features, as depicted in fig. 7. It is worth noting that these visualizations utilized the checkpoint from RECANTFormer, which was solely trained on the GREC task prior to any joint fine-tuning. Each attention weight is accompanied by an image featuring the corresponding bounding box predicted by that particular localization token. These visualizations validate that the weights of localization tokens contain pertinent information beyond forecasting of box coordinates.

J RECANTFormer+ for GRES

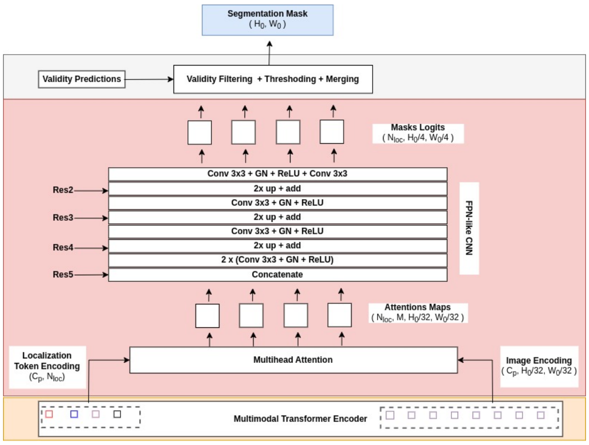

Segmentation Mask
(H_0, W_0)
Validity Predictions
Validity Filtering + Threshoding + Merging
Masks Logits
(N_{loc}, $H_0/4$, $W_0/4$)
Conv 3x3 + GN + ReLU + Conv 3x3
Res2
2x up + add
Conv 3x3 + GN + ReLU
Res3
2x up + add
Conv 3x3 + GN + ReLU
Res4
2x up + add
2 x (Conv 3x3 + GN + ReLU)
Res5
Concatenate
FPN-like CNN
Attentions Maps
(N_{loc}, M, $H_0/32$, $W_0/32$)
Localization Token Encoding
(C_p, N_{loc})
Multihead Attention
Image Encoding
(C_p, $H_0/32$, $W_0/32$)
Multimodal Transformer Encoder

Figure 8: Implementation of segmentation head extending ReCANTFormer for Generalized Referring Expression Segmentation

Mask Prediction Head The mask prediction head extends the RECANTFormer model to generate a segmentation mask per image which is illustrated in Figure 8. Here our key idea is that the self-attention mechanism in the multi-modal transformer, specifically the attention between localization tokens and visual tokens, captures the required information to generate a segmentation mask. This module receives two inputs from the multi-modal transformer encoder: 1) the output states of localization tokens, and 2) the output states of visual tokens. The attention mechanism between localization tokens and visual tokens includes multi-head attention, which in turn generates a set of M heat maps. The FPN approach (Lin et al., 2017a) is used for upsampling. The segmentation mask generates N number of masks. Masks obtained using this segmentation head are finally filtered using the validity classification head. Then selected masks are combined to generate a single segmentation mask. Our design is motivated by the extension of the DETR (Carion et al., 2020) object detector for (panoptic) segmentation. However, DETR being an encoder-decoder architecture, uses multi-head attention between decoder output and the encoded image to generate heatmaps. Linear combination of focal loss (Lin et al., 2017b) and dice loss (Milletari et al., 2016) is used to train the model.

Results on GRES The performance of RECANTFormer+ on GRES task on gRefCOCO is presented in table 8. When models with comparable backbones are considered, RECANTFormer+ outperforms MattNet (Yu et al., 2018), VLT (Ding et al., 2022), and ReLA (Liu et al., 2023) models with respect to gIoU, cIoU and N-acc metrics by significant margins.

Dataset	Visual	Text	val				testA				testB			
	Encoder	Encoder	cIoU	gIoU	N-acc	T-acc	cIoU	gIoU	N-acc	T-acc	cIoU	gIoU	N-acc	T-acc
MattNet	R-101	LSTM	47.51	48.24	41.15	96.13	58.66	59.30	44.04	**97.56**	45.33	46.14	41.32	95.32
VLT	D-53	bi-GRU	52.51	52.00	47.17	95.72	62.19	63.20	48.74	95.86	50.52	50.88	47.82	94.66
ReLA	R-50	BERT	42.04	39.10	29.70	**98.23**	47.42	44.95	35.09	96.56	38.76	36.01	23.39	**97.86**
RECANTFormer(5)+	R-50	BERT	**56.08**	**59.95**	**52.83**	95.94	**62.88**	**64.65**	**53.66**	96.88	**51.64**	**56.54**	**55.96**	93.40

Table 8: Comparison of GRES Results on gRefCOCO dataset. cIoU: Cumulative IoU. gIoU: Generalized IoU N-acc: No-target accuracy. T-acc: Target accuracy