Limitations and Broader Impacts. The training of PolyFormer requires accurate bounding box and polygon annotations. How to reduce such dependence and utilize weakly-supervised data for region-level image understanding needs further exploration. For the data and model, we need to further understand the broader impacts including but not limited to fairness, social bias and potential misuse.

A. Additional Dataset Details

We evaluate PolyFormer on four benchmark image datasets, RefCOCO [10], RefCOCO+ [10], RefCOCOg [5, 6], and ReferIt [2]. All images of RefCOCO, RefCOCO+, and RefCOCOg are from the MS COCO dataset [4] and annotated with referring expressions. We further evaluate PolyFormer models for the Referring Video Object Segmentation (R-VOS) task on Ref-DAVIS17 [3].

RefCOCO/RefCOCO+: These two datasets are collected using a two-player game [10]. RefCOCO has 142,209 annotated expressions for 50,000 objects in 19,994 images, and RefCOCO+ consists of 141,564 expressions for 49,856 objects in 19,992 images. These two datasets are splitted into training, validation, test A and test B sets, where test A contains images of multiple people and test B contains images of multiple instances of all other objects. Compared to RefCOCO, location words are banned from the referring expressions in RefCOCO+, which makes it more challenging.

RefCOCOg: This dataset is collected on Amazon Mechanical Turk, where workers are asked to write natural language referring expressions for objects. RefCOCOg consists of 85,474 referring expressions for 54,822 objects in 26,711 images. RefCOCOg has longer, more complex expressions (8.4 words on average), while the expressions in RefCOCO and RefCOCO+ are more succinct (3.5 words on average), which makes RefCOCOg particularly challenging. We use the UMD partition [6] for RefCOCOg as it provides both validation and testing sets and there is no overlapping between training and validation images.

ReferIt: ReferIt contains 130,364 referring expressions for 99,296 objects in 19,997 images collected from the SAIAPR-12 dataset [1]. We use the cleaned Berkeley split of the dataset, which consists of 58,838, 6,333, and 65,193 referring expressions in train, validation, and test sets, respectively. Compared to RefCOCO, RefCOCO+ and RefCOCOg, ReferIt contains more stuff segmentation masks, *e.g.*, sky, ground.

Ref-DAVIS17: Ref-DAVIS17 contains 90 videos from the DAVIS17 [7] dataset, where language descriptions are provided for specific objects in each video. It contains 1,544 referring expressions for 205 objects. The dataset is split into a training set and a validation set, containing 60 and 30 videos respectively. For each referred object, each of the two annotators provides the descriptions of the first-frame and the full-video. For the Ref-DAVIS17 dataset, we use the standard evaluation metrics: Region Jaccard ($\mathcal{J}$), Boundary F measure ($\mathcal{F}$), and their average value ($\mathcal{J}\&\mathcal{F}$).

$B_H \times B_W$	RefCOCO	RefCOCO+	RefCOCOg
32×32	75.07	70.15	68.49
64×64	**75.96**	**70.65**	**69.36**
128×128	74.99	70.01	68.69

Table A. Ablation study on the size of 2D coordinate codebook.

B. Additional Implementation Details

The dimension of image feature C_v is 1024 for PolyFormer-B and 1536 for PolyFormer-L. The dimensions of language feature C_l and coordinate embedding C_e are 768. We use a linear layer to project the language and image features into the same dimension of 768. We adopt 12 attention heads in the self-attention and cross-attention layers, and GELU activations in the transformer encoder and decoder layers. For L_{cls}, we set the label smoothing factor to 0.1.

C. Additional Experiment Results

To obtain the accurate coordinate embedding, we build a 2D coordinate codebook, $\mathcal{D} \in \mathbb{R}^{B_H \times B_W \times C_e}$, where B_H and B_W are the numbers of bins along the height and width dimensions, respectively. We train PolyFormer-B models with different number of bins $B_H \times B_W$ and the results are summarized in Table A. We observe that using coordinate book with 64×64 bins achieves the best result, which is adopted by default in all the other experiments.

D. More Visualization Results

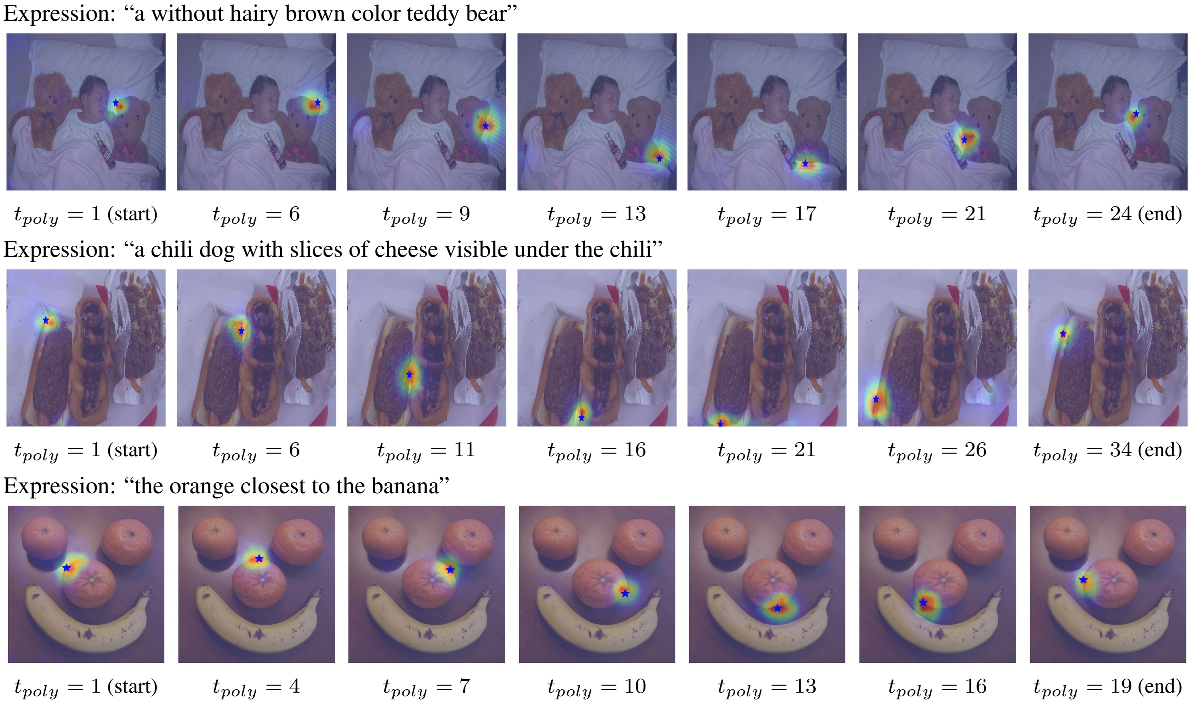
D.1. Cross-attention Map

More cross-attention map visualization is shown in Fig. A. We observe that the cross-attention map concentrates on the object referred by the sentence, and moves around the object boundary during the polygon generation process.

D.2. Prediction Visualization

Fig. B shows more examples on the synthetic images generated by Stable Diffusion [8]. Fig. C shows more examples on the RefCOCOg test set. It can be seen that

PolyFormer is able to segment the referred object in challenging scenarios, *e.g.*, instances with occlusion and complex shapes, instances that are partially displayed or require complex language understanding. In addition, PolyFormer demonstrates good generalization ability on synthetic images and text descriptions that have never been seen during training. In contrast, the state-of-the-arts LAVT [9] and SeqTR [11] fail to generate satisfactory results.



Expression: "a without hairy brown color teddy bear"

$t_{poly} = 1$ (start) $t_{poly} = 6$ $t_{poly} = 9$ $t_{poly} = 13$ $t_{poly} = 17$ $t_{poly} = 21$ $t_{poly} = 24$ (end)

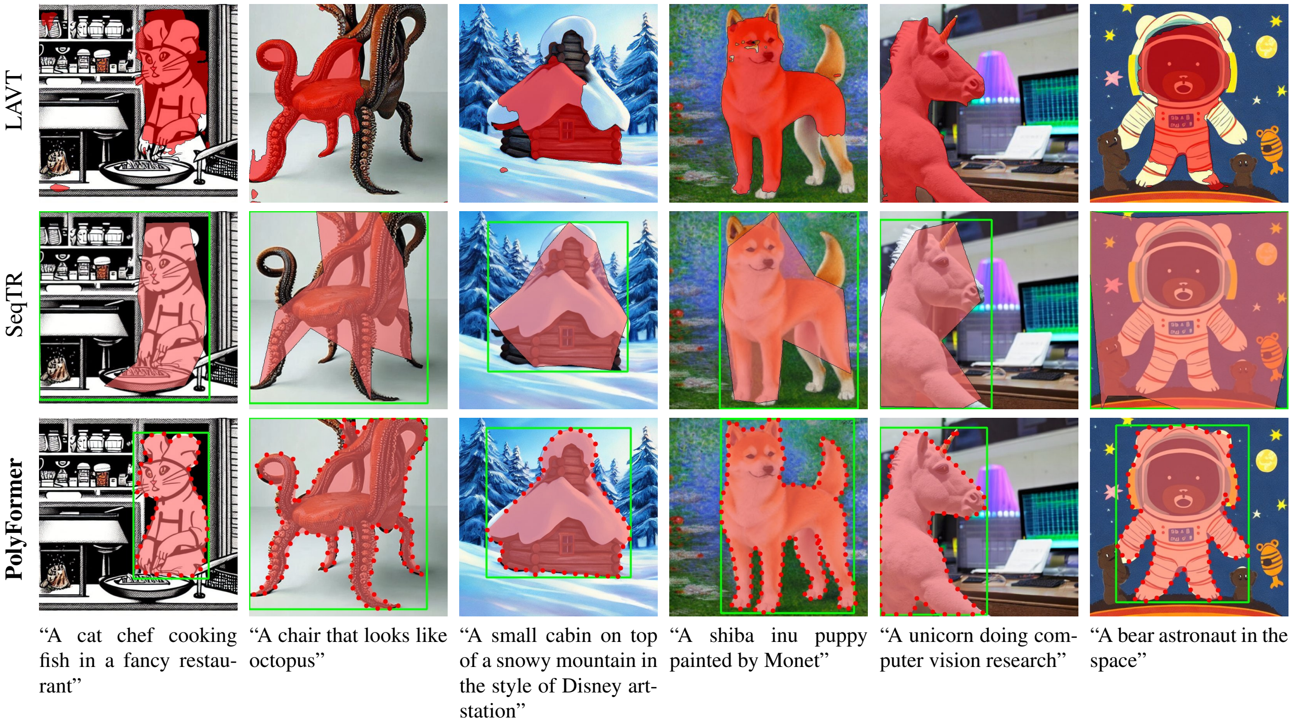
Expression: "a chili dog with slices of cheese visible under the chili"

$t_{poly} = 1$ (start) $t_{poly} = 6$ $t_{poly} = 11$ $t_{poly} = 16$ $t_{poly} = 21$ $t_{poly} = 26$ $t_{poly} = 34$ (end)

Expression: "the orange closest to the banana"

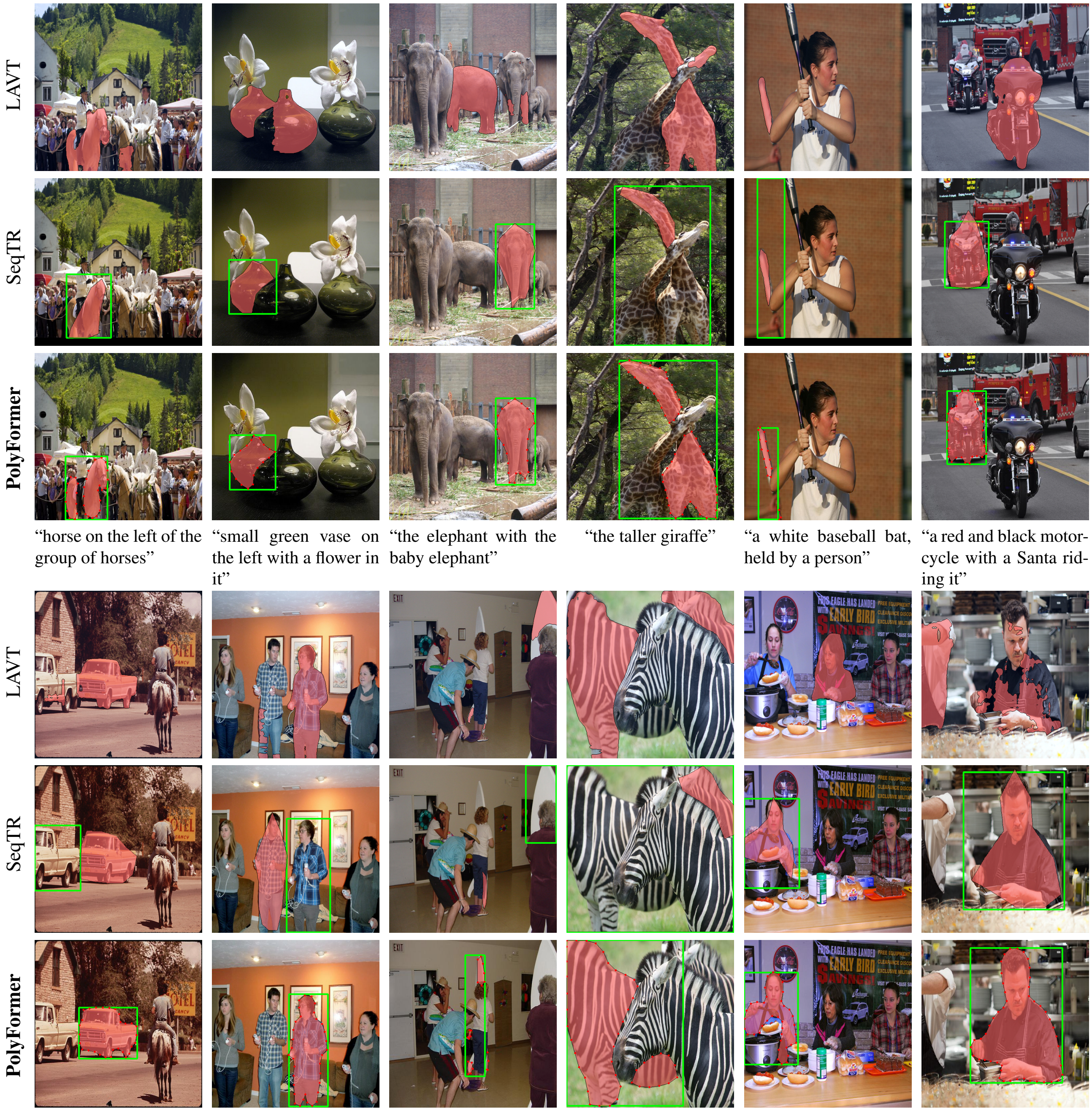
$t_{poly} = 1$ (start) $t_{poly} = 4$ $t_{poly} = 7$ $t_{poly} = 10$ $t_{poly} = 13$ $t_{poly} = 16$ $t_{poly} = 19$ (end)

Figure A. Decoder's cross-attention map when predicting the polygons. $\star$ indicates the vertex prediction at time step t_{poly}.



"A cat chef cooking fish in a fancy restaurant" "A chair that looks like octopus" "A small cabin on top of a snowy mountain in the style of Disney artstation" "A shiba inu puppy painted by Monet" "A unicorn doing computer vision research" "A bear astronaut in the space"

Figure B. The result comparison of LAVT [9], SeqTR [11] and PolyFormer on synthetic images generated by Stable Diffusion [8].



“horse on the left of the group of horses”

“small green vase on the left with a flower in it”

“the elephant with the baby elephant”

“the taller giraffe”

“a white baseball bat, held by a person”

“a red and black motorcycle with a Santa riding it”

“old yellow and white truck parked behind other truck”

“boy with blue plaid shirt and glasses”

“the surfboard the woman in a white shirt and blue capris is holding”

“a zebra with its head not visible but much of its body able to be seen”

“ a girl was cooking the food and serving”

“a man wearing a black shirt and a black and white striped apron stirring something in a metal container”

Figure C. The result comparison of LAVT [9], SeqTR [11] and PolyFormer on RefCOCOg test set. PolyFormer simultaneously predicts the bounding box and polygon vertices that forms the segmentation mask. LAVT is for referring image segmentation only. For SeqTR, we generate the bounding boxes and segmentation masks from the task-specific models as they perform better than the multi-task model.

References

[1] Hugo Jair Escalante, Carlos A. Hernández, Jesus A. Gonzalez, A. López-López, Manuel Montes, Eduardo F. Morales, L. Enrique Sucar, Luis Villaseñor, and Michael Grubinger. The segmented and annotated IAPR TC-12 benchmark. *CVIU*, 114(4):419–428, 2010. 1

[2] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014. 1

[3] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *ACCV*, pages 123–141, 2018. 1

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, pages 740–755, 2014. 1

[5] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016. 1

[6] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *ECCV*, pages 792–807, 2016. 1

[7] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 1

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 3

[9] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. LAVT: Language-Aware Vision Transformer for Referring Image Segmentation. In *CVPR*, pages 18155–18165, 2022. 2, 3, 4

[10] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85, 2016. 1

[11] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. SeqTR: A Simple Yet Universal Network for Visual Grounding. In *ECCV*, pages 598–615, 2022. 2, 3, 4