An Efficient and Effective Transformer Decoder-Based Framework for Multi-Task Visual Grounding Supplementary Material

Wei Chen[1], Long Chen[2], and Yu Wu[1]*

[1]Wuhan University [2]The Hong Kong University of Science and Technology
{weichencs,wuyucs}@whu.edu.cn, longchen@ust.hk

1 More Analysis of Elimination Strategy

Compared to the static elimination strategy, our dynamic elimination strategy offers an improved approach to address the issue of incorrect elimination during the training process. The static strategy eliminates a fixed number of visual tokens, which can lead to premature elimination and negatively impact the model's performance. In contrast, our dynamic strategy gradually increases the number of eliminated tokens as the loss converges, as shown in Figure 1. This prevents incorrect elimination and ensures a more stable training process. This is why our strategy outperforms the static elimination strategy in the results of previous ablation studies.

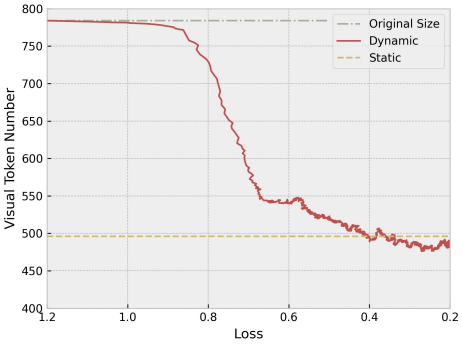

Fig. 1: Visualization of eliminating visual tokens process.

2 More Analysis of Mask Head

We observed that when using an MLP to project visual tokens into a spatial segmentation mask, the resulting mask lacks spatial relationships between pixels, as depicted in Figure 2. In order to address this issue, we propose the incorporation of a 1-channel convolution layer. This layer helps establish connections between neighboring pixels, thus enhancing the spatial coherence of the resulting mask. The ablation study results are shown in Table 1.

Table 1: Ablation study results of convolution layer in our mask head. "w/o Conv" denotes without using convolution layer.

Type	RES		REC	
	val	test	val	test
w/o Conv	67.71	68.18	78.95	79.00
with Conv	**69.15**	**70.01**	**79.60**	**80.24**

* Corresponding author

3 Additional Qualitative Results

As shown in Figure 3, we provide further visualization examples that illustrate the process of our method for eliminating visual tokens as well as compare the predicted results of our approach with those of LAVT [7].

4 Generalization on GRES.

We further validate our approach using the GRES [5] dataset, which includes text corresponding to either multiple objects or none. In this new scenario, we also achieve SOTA, demonstrating the superiority and robustness of our method.

Table 2: Performance comparison on GRES [5]

Methods	val		testA		testB	
	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU
LAVT [42]	57.64	58.40	65.32	65.90	55.04	55.83
ReLA [5]	62.42	**63.60**	69.26	70.03	59.88	61.02
Ours	**64.04**	62.75	**71.65**	**70.93**	**62.77**	**62.79**

5 Implementation Details.

We use the AdamW optimizer [6] and the weight decay is 1e-4. The initial learning rate is 5e-6 for the language backbone, 1e-5 for the visual backbone, and 2.5e-5 for the rest of the model. BERT-base [1] is utilized as the linguistic backbone for extracting linguistic features, while ViT-base [2] serves as the visual backbone. We employ the adaptation introduced by ViTDet [3] to adjust the visual backbone for higher-resolution images (*i.e.*, 448×448 in our method), and it is pre-trained on MS-COCO [4], excluding overlapping images from the val/test sets. The number of Transformer Decoder layers is 3, the hidden dimension is 768, and the feed-forward network dimension is 1024. We train the model for 150 epochs with a batch size of 80 using RTX 4090s. The patch size P is 16, the threshold α is 0.015, k in adaptive spatial attention is 1, the convolutional kernel is 5*5 in our mask head, and λ_{det} and λ_{seg} are 0.1 and 1.

References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2020)
3. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: ECCV (2022)
4. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)

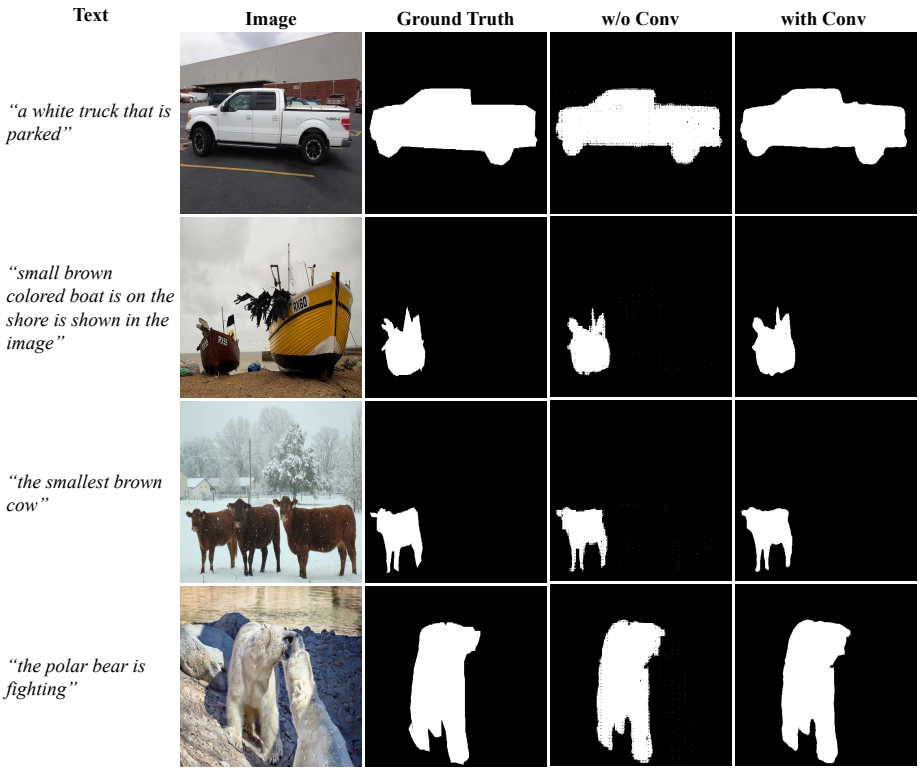

Fig. 2: Comparison of our predicted results between "w/o Conv" and "with Conv". "w/o Conv" means without using convolution layer in our mask head.

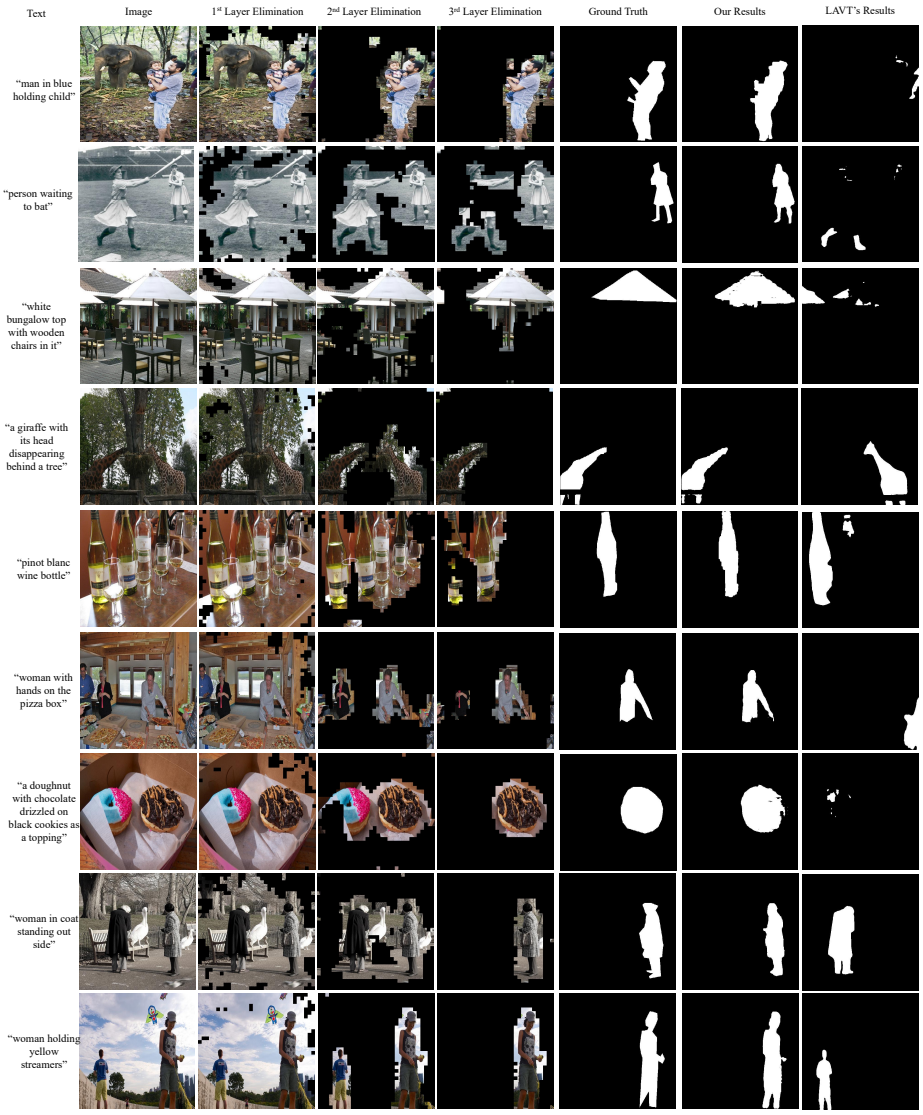

Fig. 3: Additional visualization of our results.

5. Liu, C., Ding, H., Jiang, X.: Gres: Generalized referring expression segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23592–23601 (2023)
6. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2018)
7. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. In: CVPR (2022)