

Triple Alignment Strategies for Zero-shot Phrase Grounding under Weak Supervision

Anonymous Authors

1 APPENDIX A

We adopt six benchmark datasets for evaluation:

1) **Flickr30K Entities** contains 224K phrases describing bounding boxes in 31K images, each image includes 5 captions. We also select 1000 images from the test split to evaluate in the same way as MG.

2) **MSCOCO 2014** contains 82783 train images and 40504 validation images. Each image is described by 5 captions. As with MG, the train split of the dataset is used to train our method.

3) **Visual Genome** consists of 77398 training images, 5000 test images, and 5000 validation images. Each image possesses a series of annotations which are in a free-text format.

4) **ReferIt** has 20,000 images and 99,535 segmented regions in the IAPR TC-12 and SALAPR - 12 datasets, respectively. There are approximately 130K entity captions. We use the same 9K training, 1K validation, and 10K test dataset construction strategy as MG.

5) **Flickr-Split-S0** is a zero-shot subset based on Flickr30K. The principal characteristic is that phrases in the test split do not appear in the train split, but the possibility that the captions belong to the same category cannot be avoided. For example, while “man” is in the train split, “woman” can also appear in the test split, so this split is zero-shot for phrases. We construct the dataset according to Case 0 in ZSGNet.

6) **Flickr-Split-S1** is a zero-shot subset based on Flickr30K. In the dataset, the phrases in the test split do not appear in the train split and no phrases in the train split belong to the same category as any text phrase. Flickr30K has several common categories (such as “people” and “animal”) and an “other” category. Case 1 in ZSGNet uses the samples in “other” as the validation and test splits, and the samples in the other categories are used for the train split, so it is zero-shot for phrase categories.

7) **VG-Split** includes VG-Split-S2 and VG-Split-S3. In VG-Split-S2, phrases in the training and test sets are from different synsets, and no test images contain phrase categories in the training synsets. This split corresponds to Case 2 in ZSGNet; In VG-Split-S3, each test image contains, a category belonging to the training synsets, in addition to the category to which the phrase refers. This split corresponds to Case 3 in ZSGNet.

2 APPENDIX B

In this section, we first show the visualization results of several unseen image-object instances, including fine-grained classes, such as classical car and kayak (Figure 1, the 1st, 2nd, 5th, 6th rows); novel concepts, such as celebrity names, anime names (Figure 1, the 3rd, 4th, 7th, 8th rows). Compared with CLIP-based heatmaps, our grounding heatmaps cover the image object more comprehensively and accurately, in agreement with the quantitative results. More results are shown in Figure 2.

We further compare the qualitative results with ground truths. In Figure 3, we visualize some examples of our framework on the

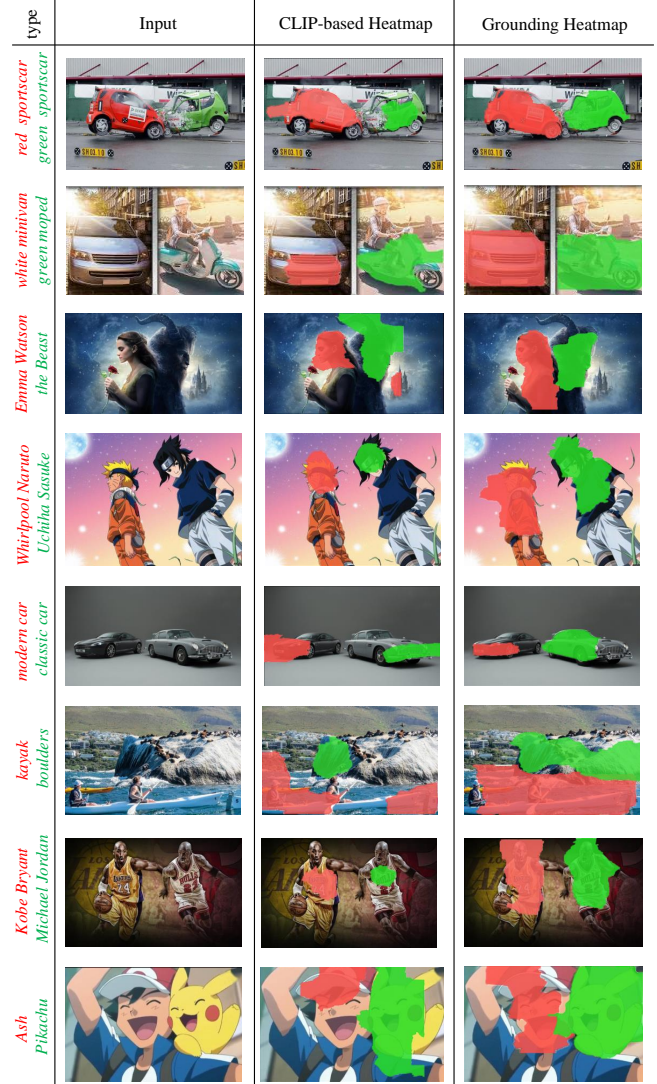


Figure 1: Phrase grounding regions of some unseen object phrases. a) Input image. b) CLIP-based Heatmap. c) Grounding Heatmap.

Flickr30K Entities, which are shown by predicted bounding boxes. We observe that our method performs well on common and uncommon phrases.

3 APPENDIX C

In this section, We give more experimental results comparing our proposed framework with other methods. **Firstly**, our framework has more satisfactory results compared with zero-shot transfer

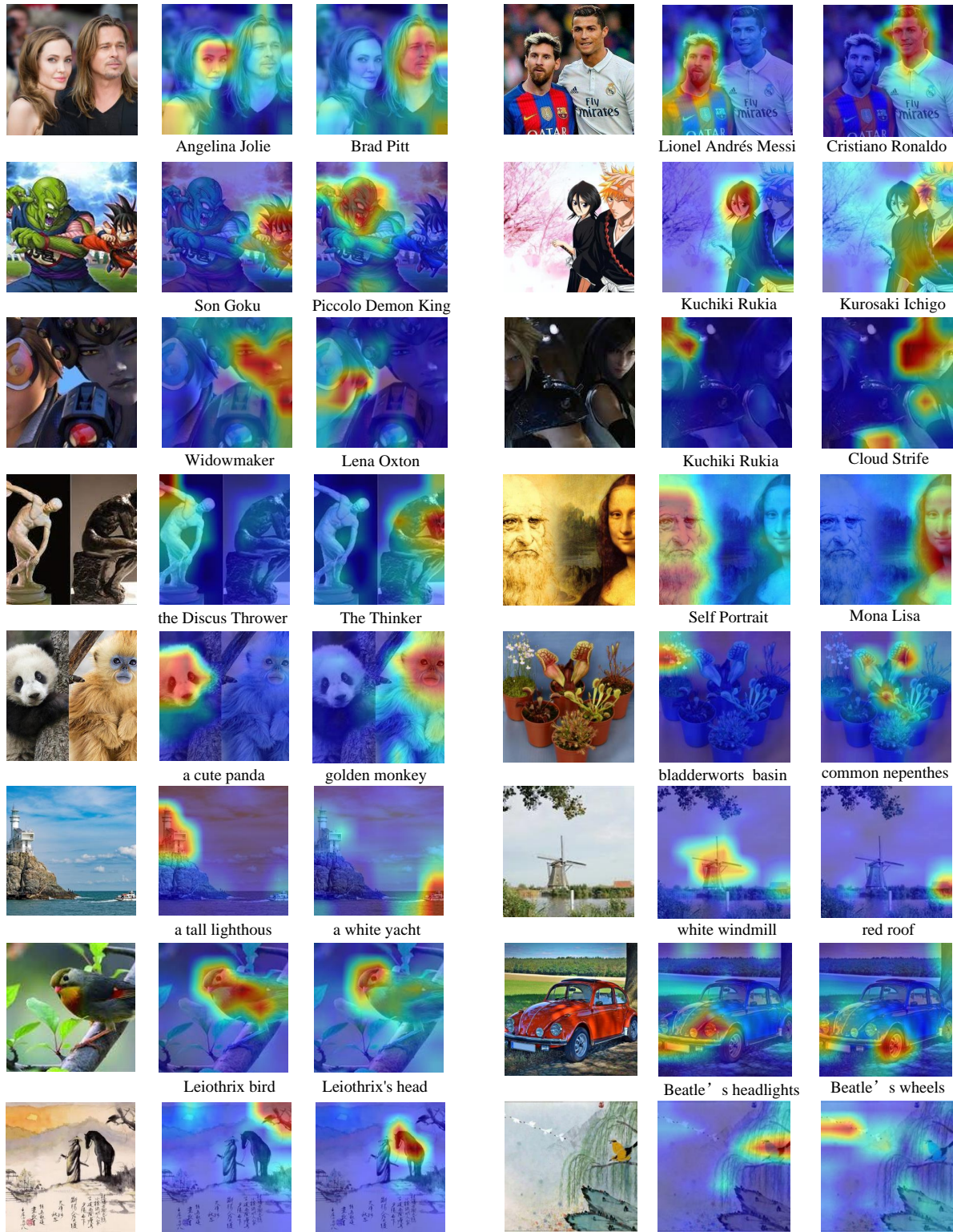


Figure 2: Heatmaps of unseen object categories from our method. The categories from top to bottom are: CElebrity names(CE), ANime names(AN), GAME character names(GA), ARTwork names(AR), RARE plant & animal phrases(RA), SMall object phrases(SM), EXclusive category phrases(EX) and SEntence-level phrases(SE).

| Method | Inference Time | Overall | People | Animals | Vehicles | Instruments | Bodyparts | Clothing | Scene | Other |
|--------------|----------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| MaskCLIP | 61 ms | 34.26 | 37.46 | 40.93 | 52.25 | 36.42 | 9.56 | 29.36 | 48.4 | 25.87 |
| AdaptingCLIP | 1360 ms | 29.47 | 29.23 | 40.15 | 45.00 | 24.69 | 13.19 | <u>27.23</u> | 41.86 | 24.92 |
| GAE | 137 ms | 25.56 | 26.76 | 39.72 | 38.12 | <u>36.76</u> | 9.14 | 19.56 | 33.72 | 22.22 |
| CH | 188 ms | <u>43.75</u> | <u>56.33</u> | 62.31 | 58.60 | 39.39 | <u>11.03</u> | 24.61 | 52.78 | <u>32.26</u> |
| Ours w/ GAE | 114 ms | 36.35 | 43.58 | 48.22 | 52.72 | 14.69 | 9.09 | 24.85 | <u>55.94</u> | 26.44 |
| Ours w/ CH | <u>114 ms</u> | 45.46 | 56.44 | <u>59.95</u> | <u>57.68</u> | 26.94 | 7.16 | 25.53 | 70.04 | 32.53 |

Table 1: Category-wise bounding box accuracy on Flickr30K Entities. In bold black: best results; Underline: suboptimal results.

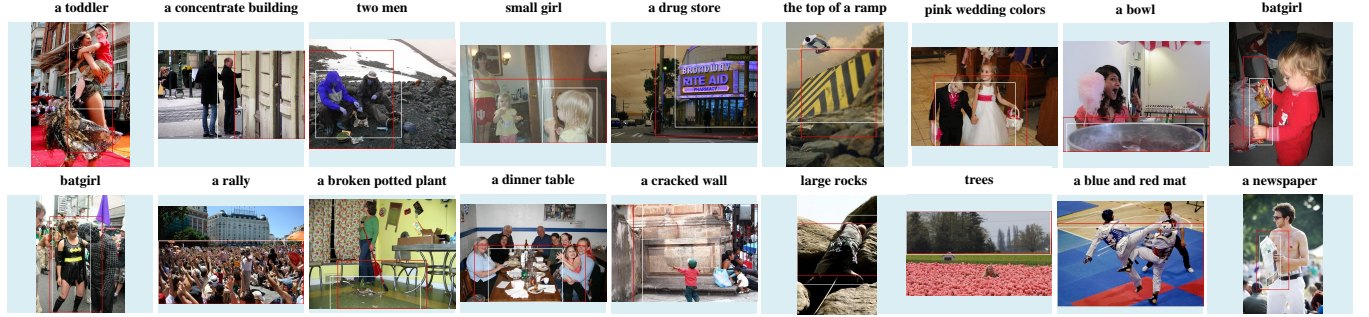


Figure 3: Qualitative results from our method. We visualize ground-truth bounding boxes in white and our predicted boxes in red.

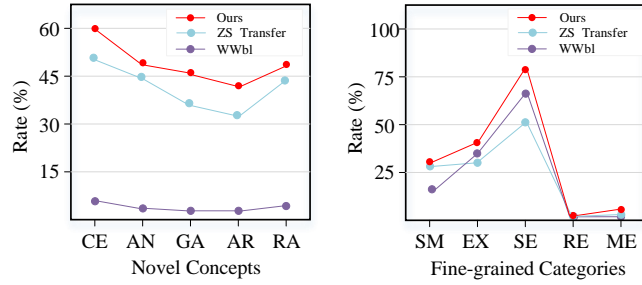


Figure 4: Comparison of the bounding box accuracy in different models with unseen object categories.

SoTA method [2] and the WWbl method [1] in Figure 4. **Secondly**, we also show a comparison of the methods' performance over all categories of Flickr30K Entities and evaluate their inference speeds in Table 1. Our framework guarantees great inference speeds with competitive accuracy.

4 APPENDIX D

In this section, We further ablate each component of our framework to determine their performance impact on their respective schemes. **Firstly**, we analyze the effect of α , β , l_n on CLIP-based heatmap generation. In fact, the number level of α and β has a great impact on the quality of the CLIP-based heatmap. As is shown in Figure 5, the points surrounded by red circles represent the optimal solutions in our settings. We also compare some qualitative results about our designs in Figure 6. The initial method represents the last layer's

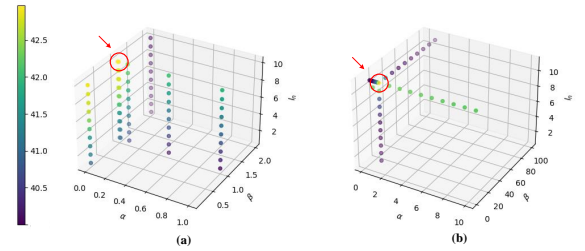


Figure 5: Comparison with CLIP-based heatmaps' qualities in different values of α , β , l_n . Visualization of the quadratic relationship among (a) α , β , the first n transformer layers and bbox accuracy; (b) α , β , the 11-th transformer layer and bbox accuracy.

image embedding from ViT-B/32. After aligning the region with the text, the CLIP-based heatmap becomes phrased relevant.

REFERENCES

- [1] Tal Shaharabany, Yoad Tewel, and Lior Wolf. 2022. What is Where by Looking: Weakly-Supervised Open-World Phrase-Grounding without Text Inputs. *NeurIPS* (2022), 28222–28237.
- [2] Chong Zhou, Chen Change Loy, and Bo Dai. 2022. Extract free dense labels from clip. In *European Conference on Computer Vision*. Springer, 696–712.

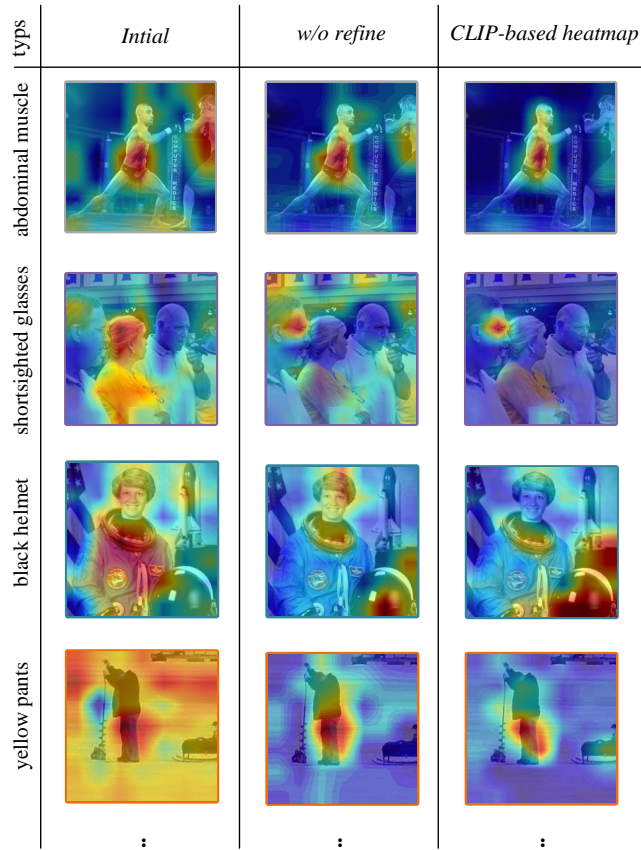


Figure 6: Comparison between the proposed and post-ablation methods. The complete CLIP-based heatmap successfully discovers phrase-related entities such as pants and glasses, while the others often fail to capture individual entities of specific descriptions.