Supplementary Material of Improving Weakly Supervised Visual Grounding by Contrastive Knowledge Distillation

In this document, we provide further ablation analysis of our knowledge distillation scheme and plot additional visualization of region-phrase matching.

1. The Effectiveness of Distillation on Phrase Localization

We further compare the results between our model using only contrastive loss (NCE) and our full model (NCE+Distill) by breaking down the results presented in our ablation study (Table 4).

1.1. Per-category localization accuracy

Method	people	clothing	body parts	animals	vehicles	instruments	scene	other
NCE	64.12	37.57	9.75	77.49	72.70	54.90	34.05	34.81
NCE+Distill	66.20	40.74	15.77	76.89	75.21	60.13	37.15	35.63

Table 1: Per-category phrase localization accuracy(%) of the NCE model and NCE+Distill model on the Flickr30K Entities dataset. Both models use Res101 CC as backbone. The NCE+Distillation model uses IRV2 OI as Detector$_K$.

Table 1 shows the per-category phrase localization accuracy on Flickr30K Entities dataset. Comparing with the NCE model, the NCE+Distill model has the largest relative improvement on following categories: body parts ($+62\%$), instruments ($+9.5\%$), scene ($+9\%$), and clothing ($+8\%$). These categories mainly contain phrases that are covered by the Open Images object classes.

1.2. Per-phrase localization accuracy on phrases correspond to detector classes

Fig. 1 shows the accuracy of the NCE only and the NCE+Distill model on the most frequent phrase categories in Flickr30K Entities [3] that are also presented in Open Images [2]. The goal of this experiment is to verify if our distillation scheme can help to improve the accuracy of phrase categories by leveraging external knowledge from the object detector. Across all 14 categories, our full model performs on par with NCE for mouth and jeans, and outperforms NCE for 12 categories including people and clothing. We note the category of "mouth" has zero accuracy for both models. This is indeed bounded by the object proposals — only 10% of the "mouth" was covered by the proposals, leading to unsatisfactory performance of both models.

2. Visualization of Region-Phrase Matching

Moving forward, we provide additional visualization of our learned region-phrase matching function, as shown in Fig. 2 (samples from Flickr30K Entities [3]) and Fig. 3 (samples from ReferItGame [1]). On both datasets, our learned matching function can identify meaningful regions associated with the phrases, as shown in Figures 2 and 3.

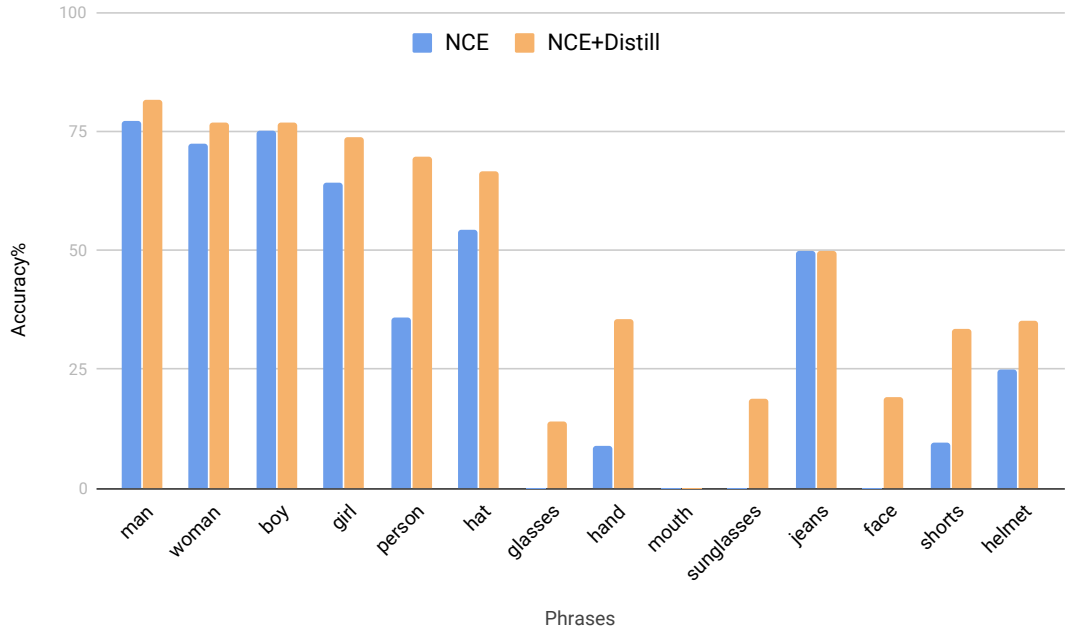

Figure 1: Phrase grounding accuracy(%) for frequent phrases on Flickr30K Entities that are also presented in Open Images detector. We compare the results of two variants of our model (NCE vs. NCE+Distill). Our full model (NCE+Distill) helps to improve those phrase categories that lie in Open Images.

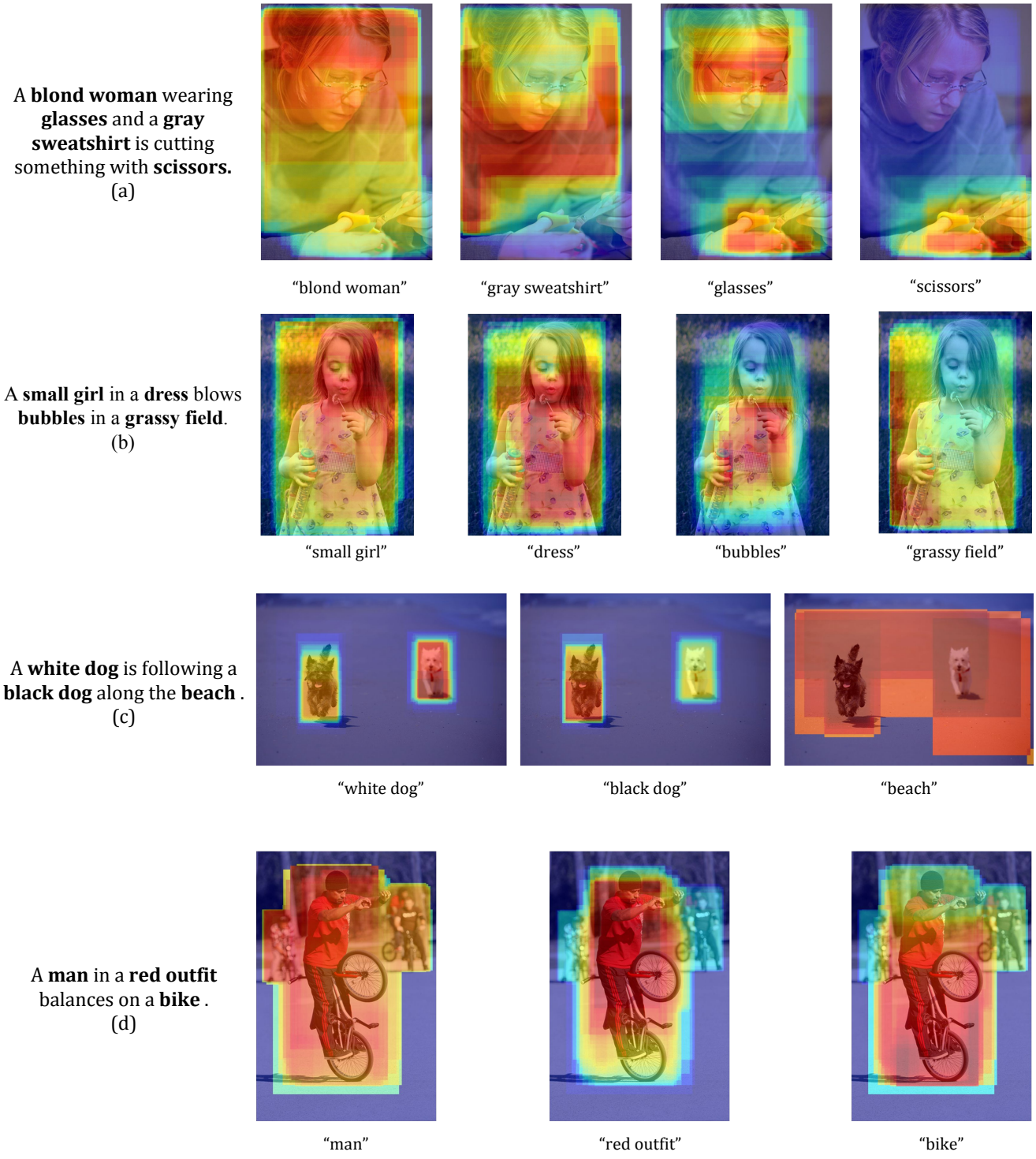

Figure 2: Visualization of region-phrase matching results using our full model (NCE+Distill) on Flickr30k Entities dataset. We present 4 sample images (a—d). For each sample, from left to right: the sentence with parsed phrases, the attention map of region-phrase matching for each phrase. For each pixel, we compute a matching score by averaging scores from all proposals covering the pixel. The red color corresponds to high matching scores.

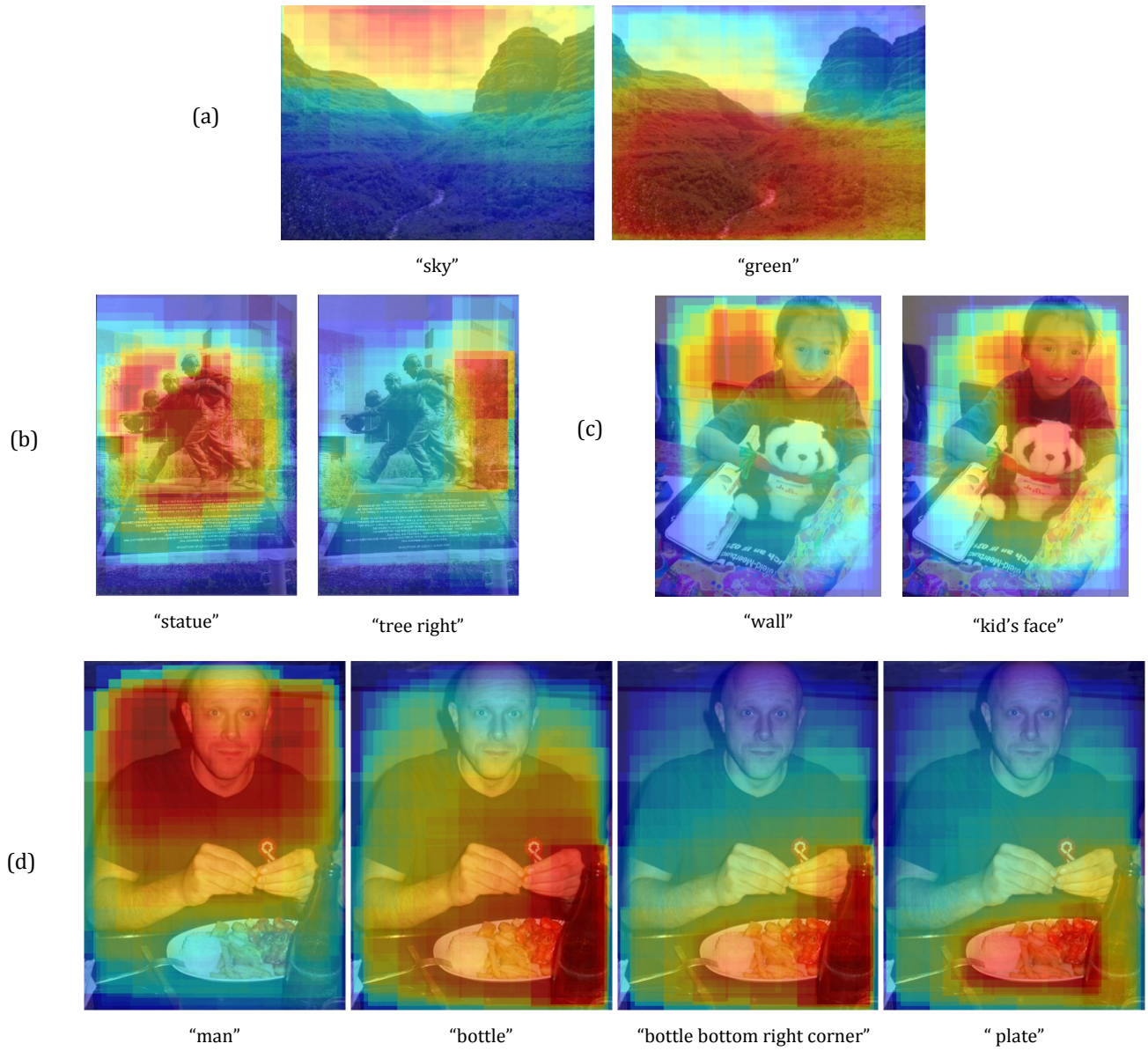

Figure 3: Visualization of region-phrase matching results using our full model (NCE+Distill) on ReferItGame dataset. We present 4 sample images (a—d). For each sample, we visualize the attention map of region-phrase matching for each phrase. Similarly, we aggregate matching scores for each pixel from all nearby proposals. The red color corresponds to high matching scores.

References

[1] 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

[2] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. OpenImages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github. com/openimages*, 2(3):18, 2017. 1

[3] 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 *ICCV*, 2015. 1