Shatter and Gather: Learning Referring Image Segmentation with Text Supervision

— Supplementary Material —

Dongwon Kim[1*] Namyup Kim[1*] Cuiling Lan[2] Suha Kwak[1]
[1]POSTECH [2]Microsoft Research Asia
http://cvlab.postech.ac.kr/research/sag

This supplementary material presents experimental results omitted from the main paper due to the space limit. We first summarize the notations in our paper in Table A1. Sec. A analyzes performance according to the number of iterations of the entity discovery module T and thresholding value τ. In Sec. B, we describe experimental details of reproducing open-vocabulary segmentation methods [5, 6]. Sec. C provides the details of the decoder used in reconstruction loss. We then present quantitative results with the DenseCRF [1] in Sec. D. Finally, Sec. E offers more qualitative results of our model on the RefCOCO *val* set. Since our model involves randomness when sampling the entity slots, all of the results in the main paper and supplementary material are obtained by averaging the results from 3 experiments, where the standard deviation across them is always less than 0.06.

A. Impact of hyperparameters T and τ

The number of iterations T: In Table A2, We investigate the impact of the number of iterations in the entity discovery module T. The model consistently achieves high IoU and accuracy when T is greater than 2. Notably, we observe a significant performance degradation when T equals 1 (3.77%p mIoU drop compared to when T is 2). These results highlight the significance of iteratively applying the aggregation and interaction blocks, which enables effective visual entity discovery through the progressive refinement of slots. Furthermore, it is worth noting that even the least-performing model with a single iteration still outperforms the previous work [3, 5, 6].

The threshold value τ: In Table A3, we present the mIoU of predictions from our model with the varying threshold value τ. Our model consistently attains high IoUs when τ ranges between 0.4 and 0.6, which is why we set τ to 0.5 for all experiments in the main paper. These findings indicate that our model is insensitive to the setting of τ.

Symbol	Description
Feature extraction	
$\mathbf{x}^{\mathcal{V}}$	The visual feature
$\mathbf{x}^{\mathcal{T}}$	The textual feature
Entity discovery module	
Φ	The entity discovery module
`Agg`	The aggregation block
`normalize`	ℓ_1 normalization of column vectors
`MLP`	A multi-layer perceptron with layer normalization
$\mathbf{S}^0$	Initial slots
$\mathbf{S}^t$	Slots at t-th iteration
$\hat{\mathbf{S}}^t$	Output slots of the aggregation block at t-th iteration
$\mathbf{S}_i^t$	Slots sampled from the i-th distribution at t-th iteration
`Inter`	The interaction block
`SA`	A self-attention transformer
Modality fusion module	
Ψ	The modality fusion module
`CA`	A cross-attention transformer
Training	
$\langle \cdot, \cdot \rangle$	A inner product between two vectors
`sg`	A stop-gradient operation
f_{dec}	The reconstruction decoder
Inference	
A^{slot}	The patch-wise attention map from the entity discover module
A^{fuse}	The entity-wise attention map from the modality fusion module
Hyperparameters	
N	The number of image patches
K	The total number of slots
K_g	The number of different Gaussian distributions for entity slot
K_s	The number of slots sampled from each distribution for entity slot
D, D_h	The output and hidden dimension
T	The number of iterations in entity discovery module
τ	The threshold for predicting mask

Table A1: Summary of notations used in the main paper together with descriptions.

B. Reproducing MaskCLIP and GroupViT

To evaluate referring image segmentation performance of the open-vocabulary segmentation methods, we consider two open-sourced models that do not require mask supervision during training: MaskCLIP [6] and GroupViT [5].

For MaskCLIP, the target referring query and image are input to the model, with inference made similarly to our approach, *i.e.*, obtaining prediction mask by thresholding the similarity map between all image patches and target query, using a threshold value of 0.5.

T	$\mathcal{A}$@0.3	$\mathcal{A}$@0.5	$\mathcal{A}$@0.7	cIoU	mIoU
8	53.08	23.30	5.98	29.38	33.85
6	55.02	24.99	6.35	**30.40**	**34.76**
4	52.03	22.55	5.85	28.95	33.24
2	48.40	19.92	5.09	27.67	31.77
1	40.36	12.93	2.73	25.11	28.00

Table A2: Performance analysis according to the number of iteration T of the entity discovery module on RefCOCO *val* set. The setting used in the main paper is indicated with a grey-colored row.

τ	0.3	0.4	0.5	0.6	0.7
mIoU	32.58	34.39	**34.76**	33.57	30.72

Table A3: Performance analysis according to thresholding hyperparameter τ on RefCOCO *val* set.

For GroupViT, the target referring query and image are input to the model along with an additional dummy query (*e.g.*, "`A photo of a nothing`"). This dummy query functions similarly to the background class in semantic segmentation methods. During inference, the model assigns image segments to the query with higher similarity, and segments assigned to the target query are considered the final prediction. We have also tried a similar inference protocol with our model for a fair comparison, but it failed to produce meaningful performance.

For the fine-tuning, we noticed that the loss does not decline with the original training configurations. Therefore, we set the batch size and learning rate for both models to 32 and 1e-5, respectively, matching the hyperparameters used for our model.

C. Decoder f_{dec} for the reconstruction loss

For the reconstruction decoder, we follow the model architecture of the spatial broadcast decoder [4] and its application to slot attention [2]. Specifically, each entity slot is broadcasted to the N sequence, which is the same length as the visual feature. Next, the broadcasted slots are augmented with sinusoidal positional encoding. These augmented slots are then fed into a four-layer multi-layer perceptron (MLP) with ReLU activation functions. The output dimension of the MLP is $D + 1$, with the additional dimension being used to compute weighting values via softmax. Finally, we compute the reconstruction of the visual feature by taking the weighted sum of all slots in each sequence position, using the calculated weighting values.

D. Additional qantitative results

Following the previous work that utilizes post-processing to enhance segmentation quality, we present the performance of our model with DenseCRF [1] in Table A4. The results demonstrate that employing DenseCRF yields performance improvements across all benchmarks. Notably, the performance of our model without DenseCRF surpasses that of the previous method, TSEG [3], even when TSEG uses DenseCRF. In the main paper, we only report the performance of our model without DenseCRF, as its computation is time-consuming and the resulting benefits are relatively marginal.

Methods	RefCOCO			RefCOCO+			Gref	PC
	val	*testA*	*testB*	*val*	*testA*	*testB*	*val*	*val*
GroupViT	10.82	11.11	11.29	11.14	10.78	11.84	12.77	9.41
MaskCLIP	19.45	18.69	21.37	19.97	18.93	21.48	21.11	23.80
TSEG	25.44	-	-	22.01	-	-	22.05	28.77
TSEG†	25.95	-	-	22.62	-	-	23.41	30.12
Ours	34.76	34.58	35.01	28.48	28.60	27.98	28.87	33.45
Ours†	**35.75**	**35.52**	**36.03**	**29.30**	**29.41**	**28.67**	**30.02**	**35.67**

Table A4: Comparison with other methods, including the post-processing by DenseCRF. [1]. The results are reported in mIoU (%). PC and † denote the PhraseCut dataset and the models post-processed by DenseCRF, respectively.

E. Additional qualitative results

In Fig. A1 and Fig. A2, qualitative results of our model on the RefCOCO *val* set are presented. The results demonstrate that our model, trained solely with image-text pair supervision, can successfully discover visual entities and integrate them into segmentation masks corresponding to free-form text queries. For instance, our model predicts accurate masks for referring expressions about non-human objects (row 2-3 in Fig. A1, A2), occluded objects (row 3 in Fig. A1, A2), and partially appeared objects (rows 5 in Fig. A1, A2). Furthermore, the results of the top-3 discovered entities reveal that the entity discovery module effectively identifies visual entities, and the modality fusion module accurately infers their relevance to the text query.

In Fig. A3, we present additional qualitative results of our framework, featuring three types of slots: entity (Ours), random, and query slots. For the entity slots, those sampled from the same distribution are represented with identical boundary colors. The results indicate that using random slots leads to noisy entity discovery due to insufficient semantic specificity. In contrast, utilizing query slots generates not adequately fine-grained entities, as they are bound to particular semantic categories. Our proposed entity slot effectively addresses these shortcomings. It produces accurate visual entities by maintaining an awareness of semantic properties and facilitating fine-grained entity discovery. Specifically, we can observe that the entity slots sampled from the same Gaussian distribution share the same semantic properties (*e.g.*, head, chair, and color) while capturing individual entities without redundancy.

Figure A1: Qualitative results of our framework on RefCOCO *val* set. We present the discovered entities from A^{slot} and their relevance scores from A^{fuse}. Top-3 entities in terms of relevance to query expression are presented.

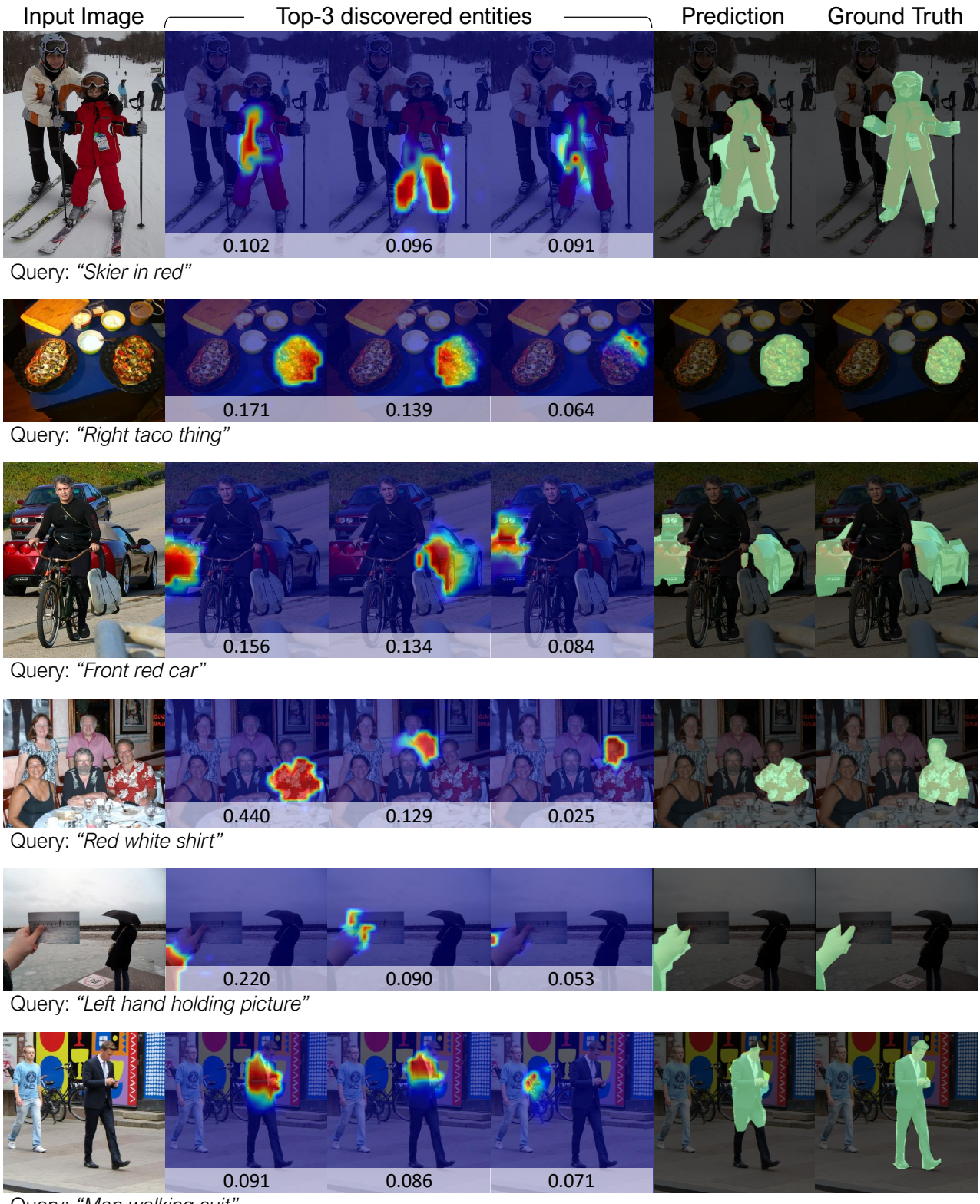

Figure A2: Qualitative results of our framework on RefCOCO *val* set. We present the discovered entities from A^{slot} and their relevance scores from A^{fuse}. Top-3 entities in terms of relevance to query expression are presented.

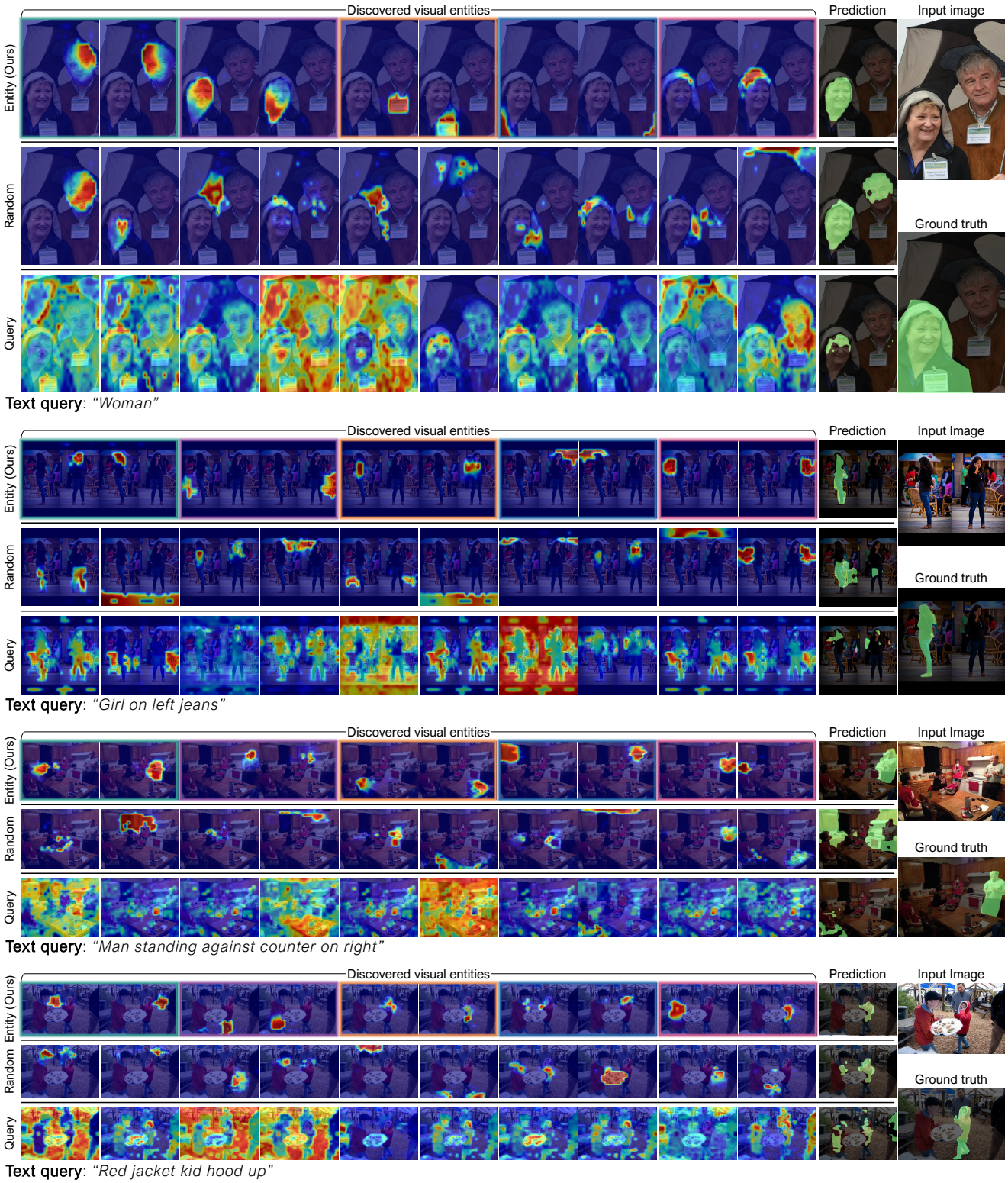

Figure A3: Qualitative results of our framework with entity slot (Ours), random slot, and query slot on RefCOCO *val* set. For each slot type, we present the 10 discovered entities from A^{slot} and final predictions. In the case of entity slots, the color of the boundaries indicates the Gaussian distribution that the slot sampled.

References

[1] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2011. 1, 2

[2] Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, and Francesco Locatello. Bridging the gap to real-world object-centric learning. In *Proc. International Conference on Learning Representations (ICLR)*, 2023. 2

[3] Robin Strudel, Ivan Laptev, and Cordelia Schmid. Weakly-supervised segmentation of referring expressions. *arXiv preprint arXiv:2205.04725*, 2022. 1, 2

[4] Nick Watters, Loic Matthey, Chris P. Burgess, and Alexander Lerchner. Spatial broadcast decoder: A simple architecture for disentangled representations in VAEs. In *ICLR Workshop on Learning from Limited Labeled Data*, 2019. 2

[5] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[6] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *Proc. European Conference on Computer Vision (ECCV)*, 2022. 1