# Beyond One-to-One: Rethinking the Referring Image Segmentation (Supplementary Material)

Yutao Hu[1]*, Qixiong Wang[2]*, Wenqi Shao[2], Enze Xie[3],
Zhenguo Li[3], Jungong Han[4], Ping Luo[1,2]†
[1]The University of Hong Kong [2]Shanghai AI Laboratory
[3]Huawei Noah's Ark Lab [4]The University of Sheffield

In the Supplementary Material, we present more experimental statistics and visualization results to further demonstrate the effectiveness of our DMMI network and reflect the value of the proposed Ref-ZOM dataset. To avoid confusion, we utilize Fig. S-XX, Table S-XX to denote the figures and tables in the Supplementary Material, while Fig. M-XX, Table M-XX, and Sec. M-XX are employed to denote the corresponding items in the main paper. If not otherwise specified, the term "G-Ref" in the Supplementary Material refers to the UMD partition of G-Ref, *i.e.,* G-Ref_(U). Meanwhile, all the experiments in the Supplementary Material are conducted with Swin-B as the visual encoder.

## 1. More details about Ref-ZOM

In Fig. S-1(a), we present a pie graph that illustrates the distribution of different types of text inputs in Ref-ZOM. Generally speaking, there are 56972 one-to-one samples in the dataset, occupying 63.16% of the entire dataset. Meanwhile, 23.60% and 13.23% cases are under one-to-many and one-to-zero conditions, leading to a total of 21290 and 11937 image-text pairs. Ref-ZOM is the first referring image segmentation dataset containing one-to-zero, one-to-one and one-to-many samples simultaneously, making it more comprehensive than previous datasets. In Fig. S-1(b), we depict the word cloud of our dataset, in which the size of each word corresponds to the square root of its frequency in the text expressions. Meanwhile, in Table S-1, we compare our Ref-ZOM with three mainstream referring image segmentation datasets.

Additionally, in Fig. S-2, we illustrate more samples in Ref-ZOM. From top to bottom are samples under one-to-many, one-to-one and one-to-zero conditions, respectively, which demonstrates the challenging nature of our Ref-ZOM. For instance, in the first sample of the first row, the text expression refers to two Asian men next to a no smoking sign. The model needs to recognize the specific sign and distinguish the corresponding Asian men from the crowd, which poses a high demand on the model. Similarly, in the second sample of the second row, the model needs to comprehend the word about color and recognize the entity "camera", which calls for high capacity to understand the expression. Furthermore, for the last sample in the penultimate row, the man in the picture is not around a bunch of flowers, which is inconsistent with the text expression and resulting in a one-to-zero case. To address this sample, instead of only concentrating on the subject "man", the network needs to understand the entire sentence comprehensively.

Furthermore, as introduced in the Sec. M-4 in the main paper, we collect one-to-many samples in three different ways. To be more specific, we manually create 9234 image-text pairs and annotate the 23558 objects via the two-player game. Meanwhile, we create 8388 image-text pairs with 10984 annotated objects through the combination of samples from existing datasets [4, 7, 6]. For two sentences describing different targets in one image, we concatenate the two text expressions via the "and" as the conjunction. Notably, we select samples from RefCOCO [4], RefCOCO+ [4] and G-Ref [7, 6] simultaneously. Finally, we create 3668 image-text pairs based on the category information in COCO. In these images, we select the category that has more than two corresponding objects. Then, we generate the sentence through the template like "[xxx] in the picture" where [xxx] indicates the category name, or directly utilize the category name "[xxx]" as the text expression. This strategy contributes to 8563 annotations of one-to-many samples in the dataset.

## 2. Training/test on Ref-ZOM

In this section, we present a detailed comparison of the network trained on Ref-ZOM with state-of-the-art methods. The results are listed in Table S-2. Compared to Table M-2

*Equal contribution.
†Corresponding author.

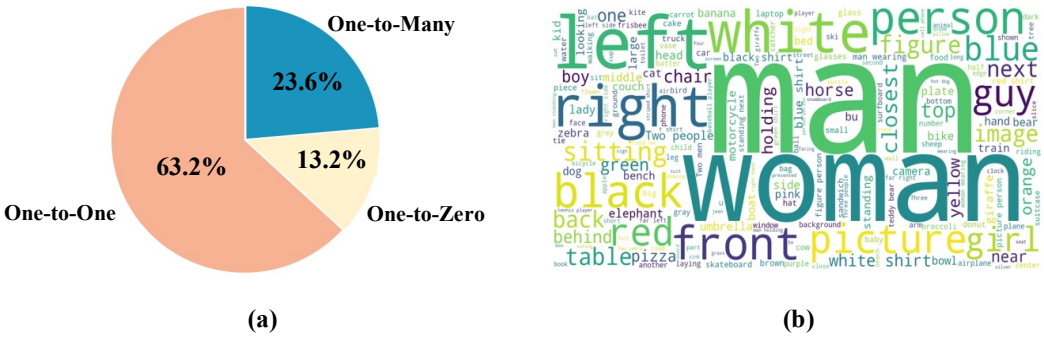


Figure 1: More detailed information of our Ref-ZOM dataset. **(a)** The distribution of samples under different settings in the Ref-ZOM. **(b)** Word clouds of Ref-ZOM dataset, where the size of each word represents its frequency in the text expression.

Table 1: The comparisons of the Ref-ZOM with three mainstream referring image segmentation datasets.

| | # Images | one-to-one | one-to-many | one-to-zero |
|---|---|---|---|---|
| RefCOCO [4] | 19994 | ✓ | – | – |
| RefCOCO+ [4] | 19992 | ✓ | – | – |
| G-Ref [7] | 26711 | ✓ | – | – |
| Ref-ZOM | 55078 | ✓ | ✓ | ✓ |

in the main paper, Table S-2 additionally reports the performance on one-to-one and one-to-many samples respectively. In other words, Table S-2 is an extended version of Table M-2. We can find DMMI network exhibits outstanding performance. Particularly, for the one-to-many samples, the proposed DMMI surpasses the previous methods by a large margin. This demonstrates that, through the reconstruction of text embedding, the visual feature is encouraged to fully incorporate the semantic clues about entity from the linguistic feature. Therefore, the text-to-image decoder could produce an accurate segmentation map based on the sufficient understanding of expressions, rather than just segmenting the target with the highest response.

Meanwhile, we visualize more segmentation maps generated by different methods in Fig.S-3. It is clear that our DMMI network exhibits the best performance. Taking the second row as an example, our method can precisely localize the three jumping men while the other methods fail to do so. Specifically, VLT [2] and LAVT [8] fail to distinguish the man running on the ground and localize all the persons in the image. Although MCN [5] could localize three jumping persons, its segmentation boundary does not align well with the objects, especially for the middle person, resulting in an inaccurate prediction map. In contrast, our DMMI network not only localizes the three corresponding persons successfully, but also generates accurate boundaries and segments the targets well. Additionally, in the seventh row, a one-to-zero sample, "The bird in front looking left", does not refer to any object in the image. It is noticeable that prior methods still tend to segment the salient object in the image, while our method handles this situation more effectively. This indicates that previous methods tend to overfit on the region with the highest response, even if it is irrelevant to the text expression. Differently, through the dual learning in the training, DMMI network can better comprehend the target entity in the expression and incorporate the semantic clues into the visual features, leading to the more accurate segmentation when facing one-to-zero samples.

## 3. Ablation Study

In this part, we present more statistics about the ablation study on G-Ref_(U) and Ref-ZOM datasets. Compared to the ablation study in the main paper (Table M-3), we additionally report the mIoU and prec@$\{0.5, 0.7, 0.9\}$ to further investigate the effectiveness of each component in DMMI network. The results are listed in Table S-3. As shown, the proposed MBA module and image-to-text decoder bring significant performance gains. Particularly, observing the third and the seventh row in Table S-3, we can find the reconstruction of text embedding contributes a lot to the final results. This suggests that, to support the reconstruction of text embedding in the training, the information of the target entity is fully incorporated into the visual feature, thus benefiting the generation of segmentation maps.

In Fig. S-4, we illustrate the segmentation maps generated by different ablation models. Specifically, we ablate the Multi-scale Bi-direction Attention (MBA) in the third column of Fig. S-4, while we remove the image-to-text decoder in the fourth column. We observe that the entire DMMI delivers the best performance. Specifically, as shown in the fourth column of the third row, the model only localizes two sheep in the image and omits the leftmost one

Table 2: Comparisons with state-of-the-art methods on Ref-ZOM dataset. On the one hand, we report the accuracy on one-to-one, one-to-many and one-to-zero samples separately. On the other hand, in "Overall Targets", we report the combined accuracy of one-to-one and one-to-many samples together.

| Method | One-to-One | | One-to-Many | | Overall Targets | | One-to-Zero |
|---|---|---|---|---|---|---|---|
| | oIoU | mIoU | oIoU | mIoU | oIoU | mIoU | Acc |
| MCN [5] | 52.09 | 53.14 | 58.04 | 57.21 | 55.03 | 54.70 | 75.81 |
| CMPC [3] | 52.46 | 52.89 | 60.23 | 60.27 | 56.19 | 55.72 | 77.01 |
| VLT [2] | 59.07 | 58.96 | 61.42 | 62.79 | 60.21 | 60.43 | 79.26 |
| LAVT [8] | 63.21 | 64.56 | 65.69 | 65.14 | 64.45 | 64.78 | 83.11 |
| **DMMI (Ours)** | **65.43** | **66.83** | **72.20** | **70.44** | **68.77** | **68.21** | **87.02** |

Table 3: Ablation study of different components in DMMI network on G-Ref and Ref-ZOM datasets. Notably, "Bi-D" indicates the bi-direction operation in MBA module, I2T denotes the "Image-to-Text" decoder and "P@$X$" indicates the percentage of test images with an IoU score higher than the threshold $X$.

| | MBA | | I2T | | G-Ref | | | | | Ref-ZOM | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bi-D | MS | $\mathcal{L}_{sim}$ | $\mathcal{L}_{con}$ | P@0.5 | P@0.7 | P@0.9 | oIoU | mIoU | P@0.5 | P@0.7 | P@0.9 | oIoU | mIoU | Acc |
| 1 | ✓ | ✓ | | | 72.71 | 61.11 | 26.23 | 61.76 | 64.36 | 72.43 | 62.33 | 23.29 | 65.77 | 63.95 | 83.91 |
| 2 | ✓ | ✓ | ✓ | | 74.45 | 63.52 | 27.14 | 62.47 | 65.82 | 74.62 | 63.05 | 24.85 | 67.25 | 66.07 | 85.55 |
| 3 | ✓ | ✓ | | ✓ | 73.67 | 62.19 | 25.98 | 62.13 | 65.15 | 73.41 | 62.66 | 23.68 | 66.36 | 64.58 | 84.73 |
| 4 | | | ✓ | ✓ | 72.79 | 62.32 | 26.59 | 62.05 | 64.37 | 73.94 | 63.07 | 24.13 | 67.13 | 65.42 | 85.09 |
| 5 | ✓ | | ✓ | ✓ | 71.47 | 60.78 | 25.63 | 62.20 | 63.39 | 75.27 | 64.84 | 25.19 | 67.31 | 66.99 | 85.82 |
| 6 | | ✓ | ✓ | ✓ | 74.51 | 63.13 | 26.76 | 62.48 | 65.82 | 75.87 | 65.72 | 26.77 | 67.52 | 67.36 | 86.11 |
| 7 | ✓ | ✓ | ✓ | ✓ | **74.98** | **65.34** | **28.96** | **63.46** | **66.48** | **76.30** | **66.32** | **29.85** | **68.77** | **68.21** | **87.02** |

when the image-to-text decoder is ablated, demonstrating the dual learning plays an important role for DMMI in solving one-to-many cases. In the dual learning, the visual feature is encouraged to fully incorporate semantic clues about the target entity to accurately reconstruct the text embedding, which promotes the understanding of the linguistic feature. Therefore, the text-to-image decoder produces segmentation maps consistent with the text expression. Meanwhile, as illustrated in the third row, although the model without MBA module localizes all the entities correctly, it fails to produce accurate boundaries. Differently, as shown in the last column, the accurate segmentation map is generated when DMMI is equipped with MBA module. This suggests through the feature interaction in local regions with different sizes, DMMI network could better handle the details in the segmentation boundary.

Furthermore, in Fig. S-5, we present the heat map of visual features extracted from ablation models. Here, we extract $Y_2$ from text-to-image decoder for visualization. Specifically, $Y_2$ is first compressed along the channel dimension and then normalized to [0,1]. We observe that, in heat maps extracted from the complete DMMI model, the target entity referred to has the highest response, which is consistent with the segmentation maps in Fig.S-4.

## 4. Zero-shot to Ref-ZOM

In this part, we visualize segmentation maps produced by MCN [5], VLT [2], LAVT [8] and our DMMI when they are trained on G-Ref and then applied to Ref-ZOM without fine-tuning. The results are illustrated in Fig. S-6, where our DMMI produces the best segmentation map. Specifically, as shown in the third row, previous methods tend to be disturbed by other targets in the image and fail to segment the two buses accurately. On the contrary, DMMI can precisely localize the two buses, which reflects the great capacity of our method in handling one-to-many samples. Furthermore, as shown in the sixth row, the model needs to localize the guy looking at tickets, which requires to distinguish the specific person from the others. We can find the compared methods cannot accurately comprehend the text expression and segment the other person in the image. In comparison, our DMMI generates the correct segmentation map, suggesting the capacity of our method in understanding the text expression. Meanwhile, this also reflects that enhancing the understanding of the target entity benefits the one-to-one case as well.

## 5. Zero-shot to Cityscapes

To further demonstrate the generalization ability of DMMI network and highlight the potential value of Ref-ZOM dataset, we visualize segmentation maps in Fig. S-7

when the network is transferred to Cityscapes [1] dataset without fine-tuning. Since Cityscapes does not contain text expressions, we manually create some expressions describing the targets in the image and feed them to the model as the text input. Here, we select the prior state-of-the-art method, LAVT [8], as a comparison. Meanwhile, the G-Ref and Ref-ZOM are employed as the training set, respectively. Since the image styles of G-Ref and Ref-ZOM are quite different from Cityscapes, it is extremely challenging to directly apply the model to Cityscapes. Moreover, G-Ref only contains one-to-one samples, which also brings difficulties for the model to handle various types of text inputs.

In Fig. S-7, we can easily find that our DMMI network outperforms LAVT substantially. On the one hand, as shown in the eighth row, our method precisely handle the one-to-zero sample even though it is trained on the G-Ref, while LAVT fails in this condition. On the other hand, through the training of our Ref-ZOM dataset, both LAVT and DMMI obtain significant performance improvements. Taking the second row as an example, when trained on G-Ref, both LAVT and DMMI fail to localize the stop sign and two cars simultaneously. In contrast, when utilizing Ref-ZOM as the training set, DMMI generates the best segmentation maps. Meanwhile, based on the Ref-ZOM, although LAVT still segments some incorrect regions, its performance is much better than the G-Ref-trained one. This demonstrates that our Ref-ZOM endows the network with more powerful capacity to solve various types of text inputs and promotes the generalization ability of the model.

## References


[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding.

[2] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16321–16330, 2021.

[3] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring image segmentation via cross-modal progressive comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10488–10497, 2020.

[4] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, 2014.

[5] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10034–10043, 2020.

[6] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11–20, 2016.

[7] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Proceedings of the European Conference on Computer Vision*, pages 792–807. Springer, 2016.

[8] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022.

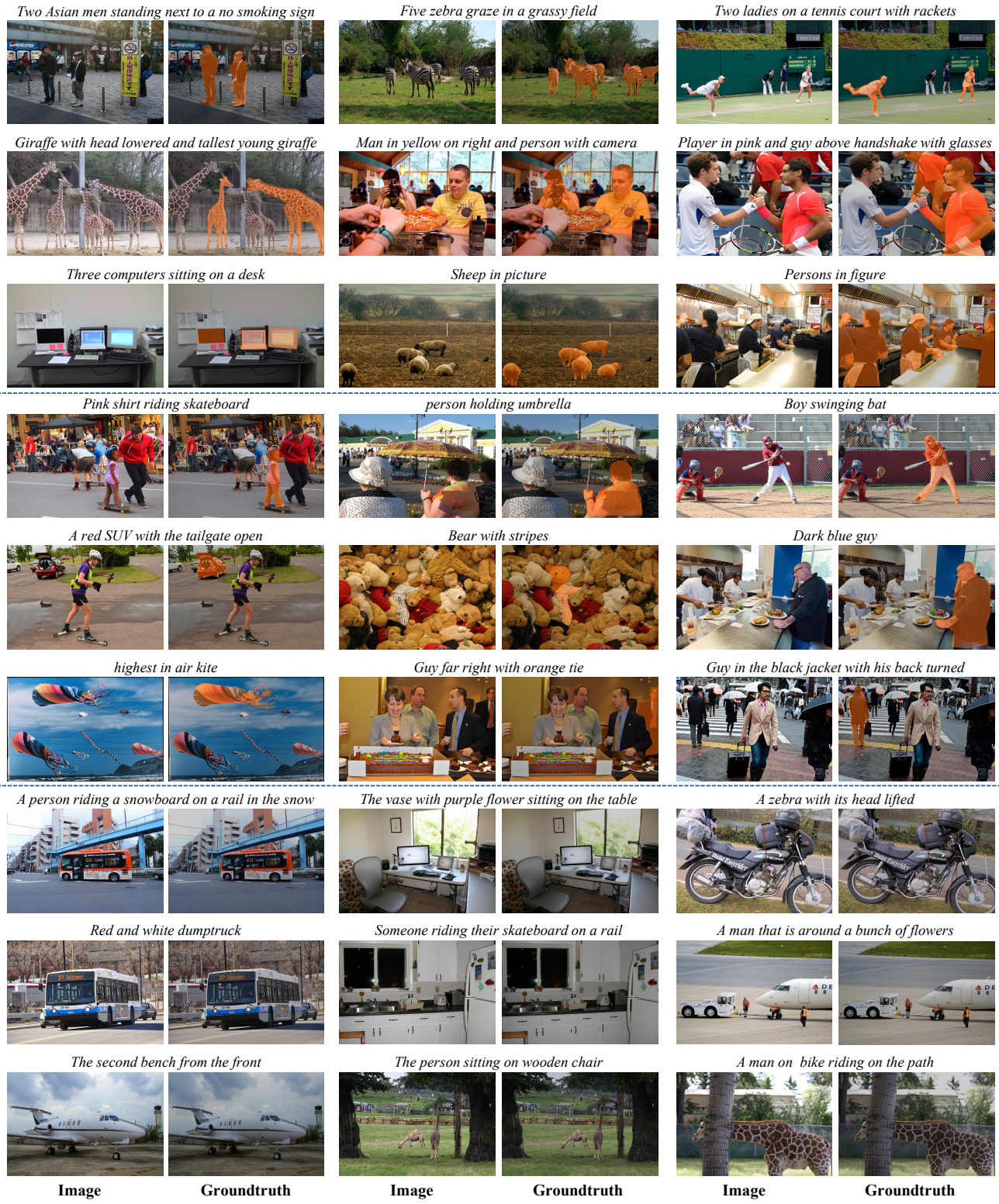


Figure 2: Illustrations of representative samples selected from our newly proposed Ref-ZOM dataset. From top to down, the first three rows are one-to many samples, while the second three rows and the last three rows present the cases under one-to-one and one-to-zero conditions, respectively.


Figure 3: Illustrations of segmentation maps generated by different methods. Here, all the models are trained on our Ref-ZOM and then evaluated on the test set of it. From top to bottom, the first four rows illustrate the segmentation results of one-to-many samples. The fifth and sixth rows present the segmentation maps of one-to-one cases, while the last two rows are one-to-zero samples.

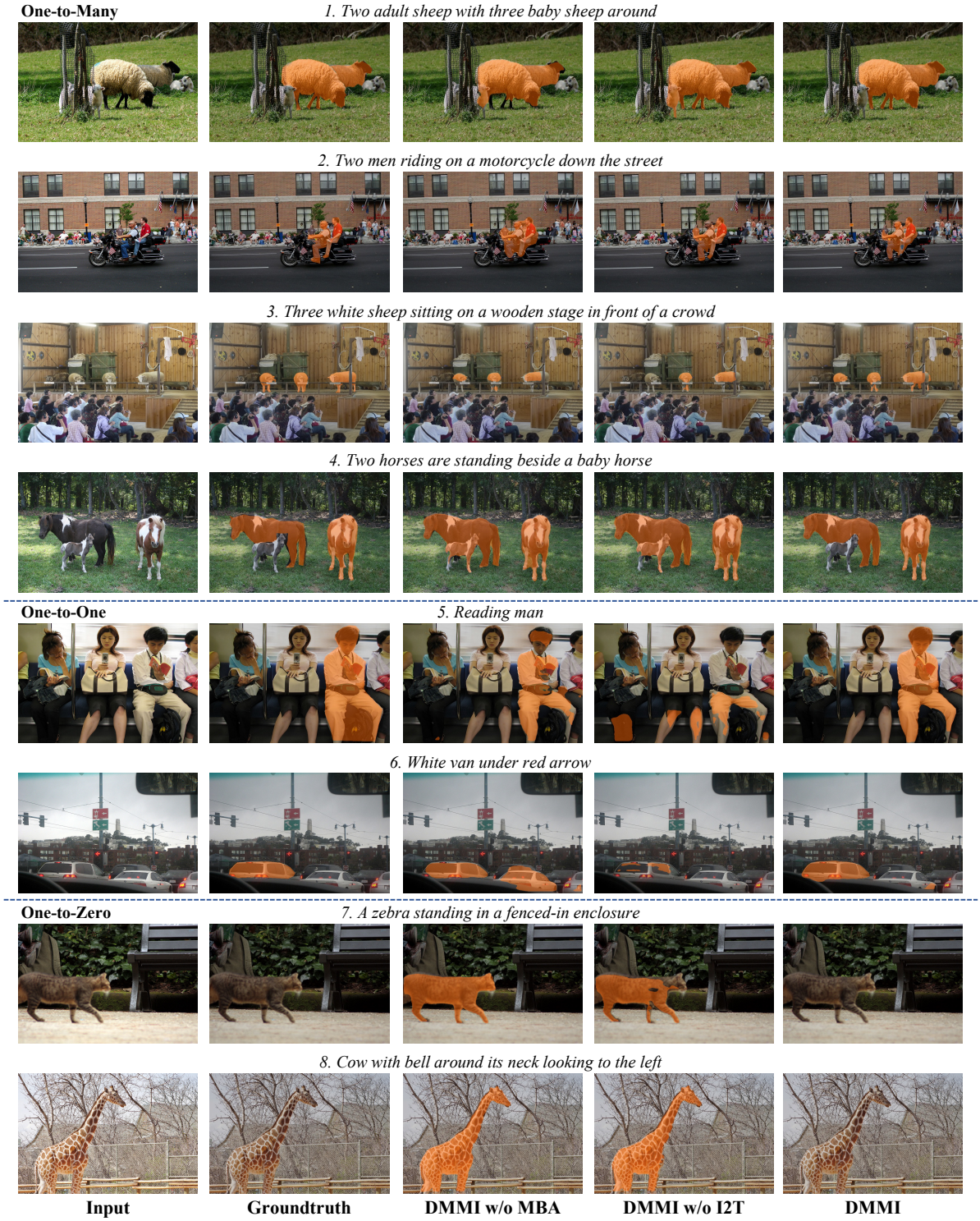


Figure 4: Illustrations of segmentation maps produced by different ablation models. Notably, "I2T" denotes the image-to-text decoder. Here, all the models are trained on our Ref-ZOM and then evaluated on the test set of it. From top to bottom, the first four rows illustrate the segmentation results of one-to-many samples. The fifth and sixth rows present the segmentation maps of one-to-one cases. While the last two rows are one-to-zero samples.

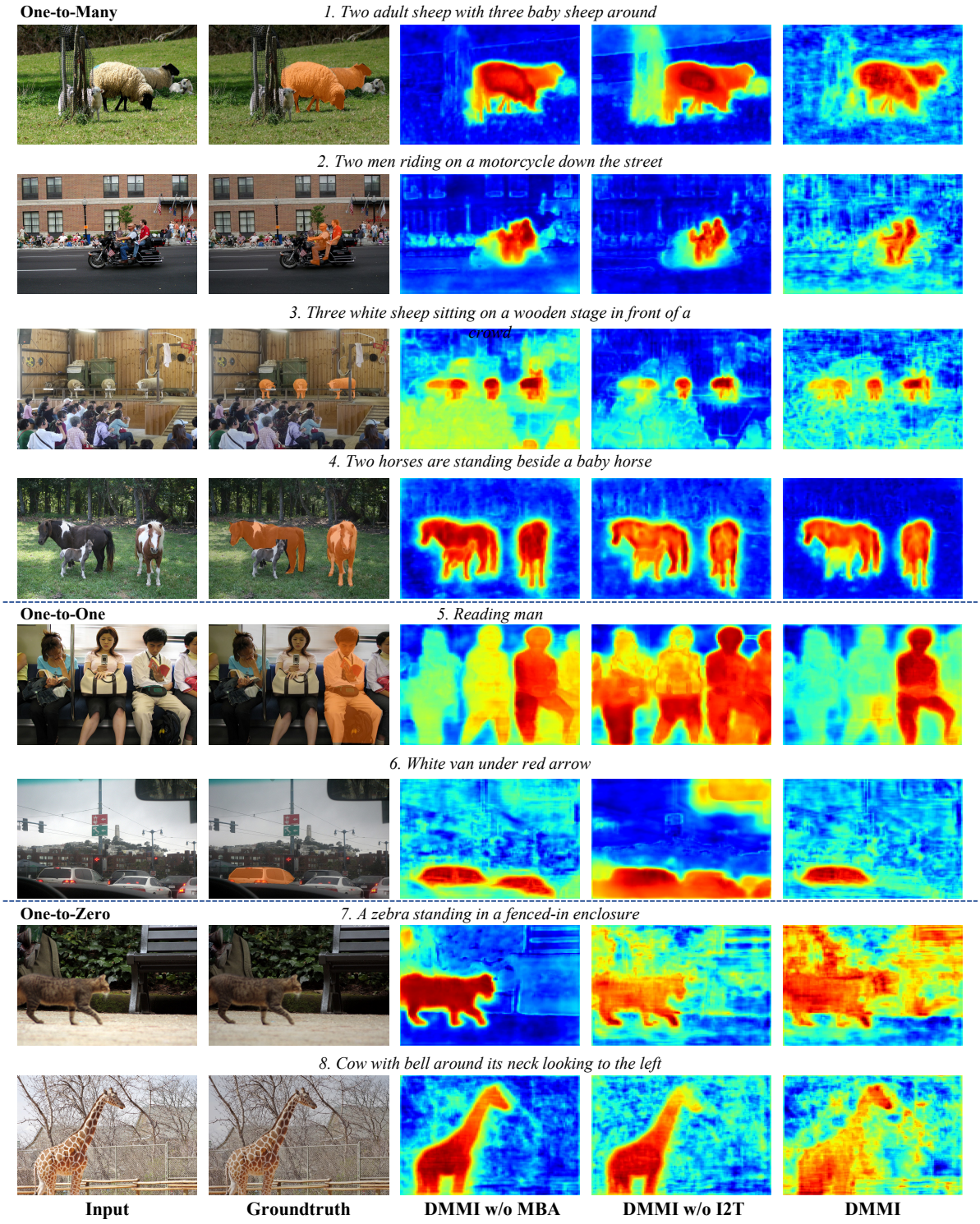


Figure 5: Illustrations of heat maps produced by different ablation models. Notably, "I2T" denotes the image-to-text decoder. Meanwhile, all the models are trained on our Ref-ZOM and then evaluated on the test set of it. Here, we extract the feature maps $Y_2$ from the text-to-image decoder, which are compressed along the channel dimension and then normalized to [0,1] for visualization. From top to bottom, the first four rows illustrate the heat maps of one-to-many samples. The fifth and sixth rows present the heat maps of one-to-one cases. While the last two rows are one-to-zero samples.

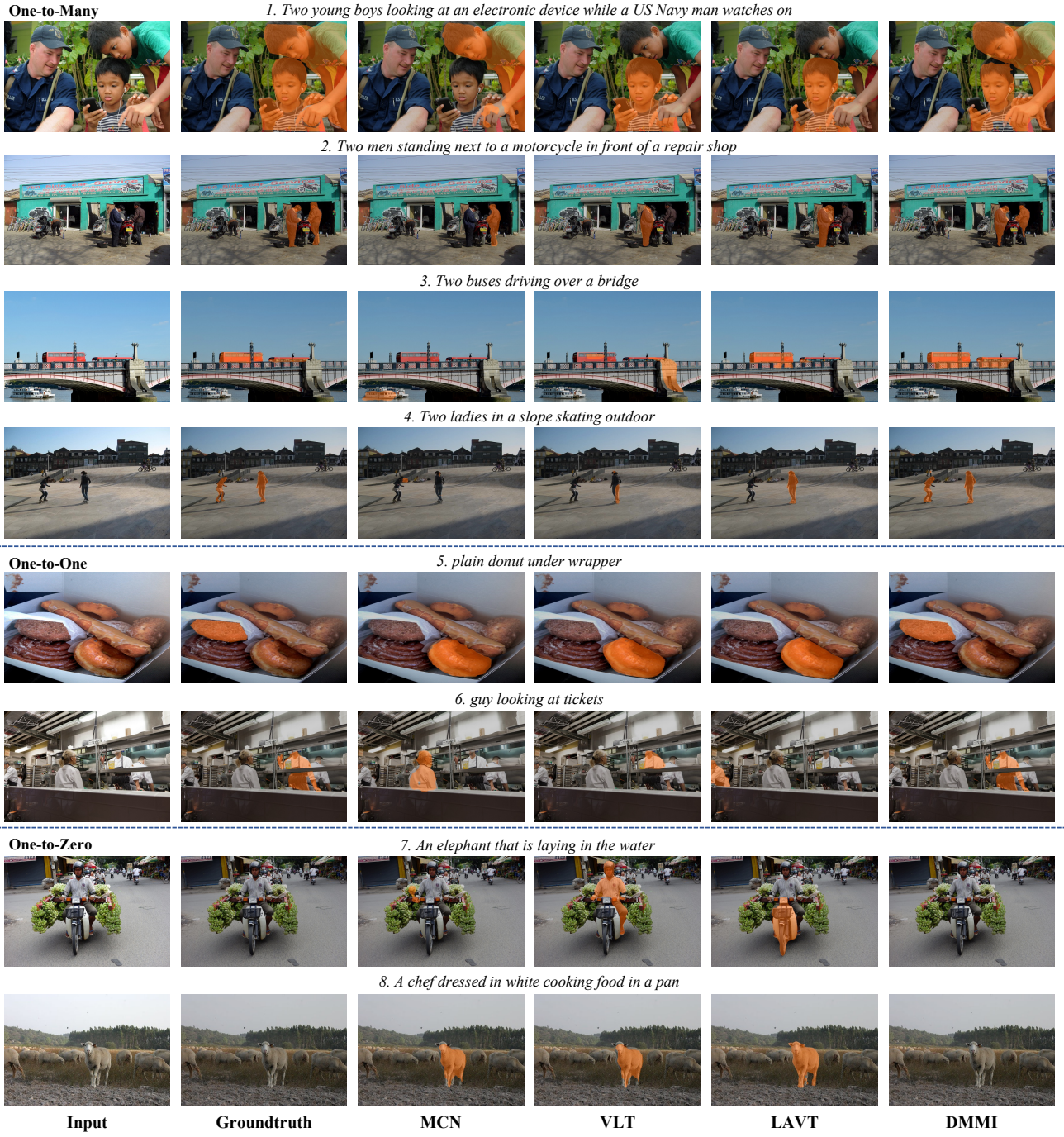


Figure 6: Illustrations of segmentation maps generated by different methods when they are trained on G-Ref and then directly transferred to Ref-ZOM without fine-tuning. From top to bottom, the first four rows visualize the segmentation results of one-to-many samples. The fifth and sixth rows illustrate the segmentation maps on one-to-one cases, while the last two rows present the segmentation results on one-to-zero case.

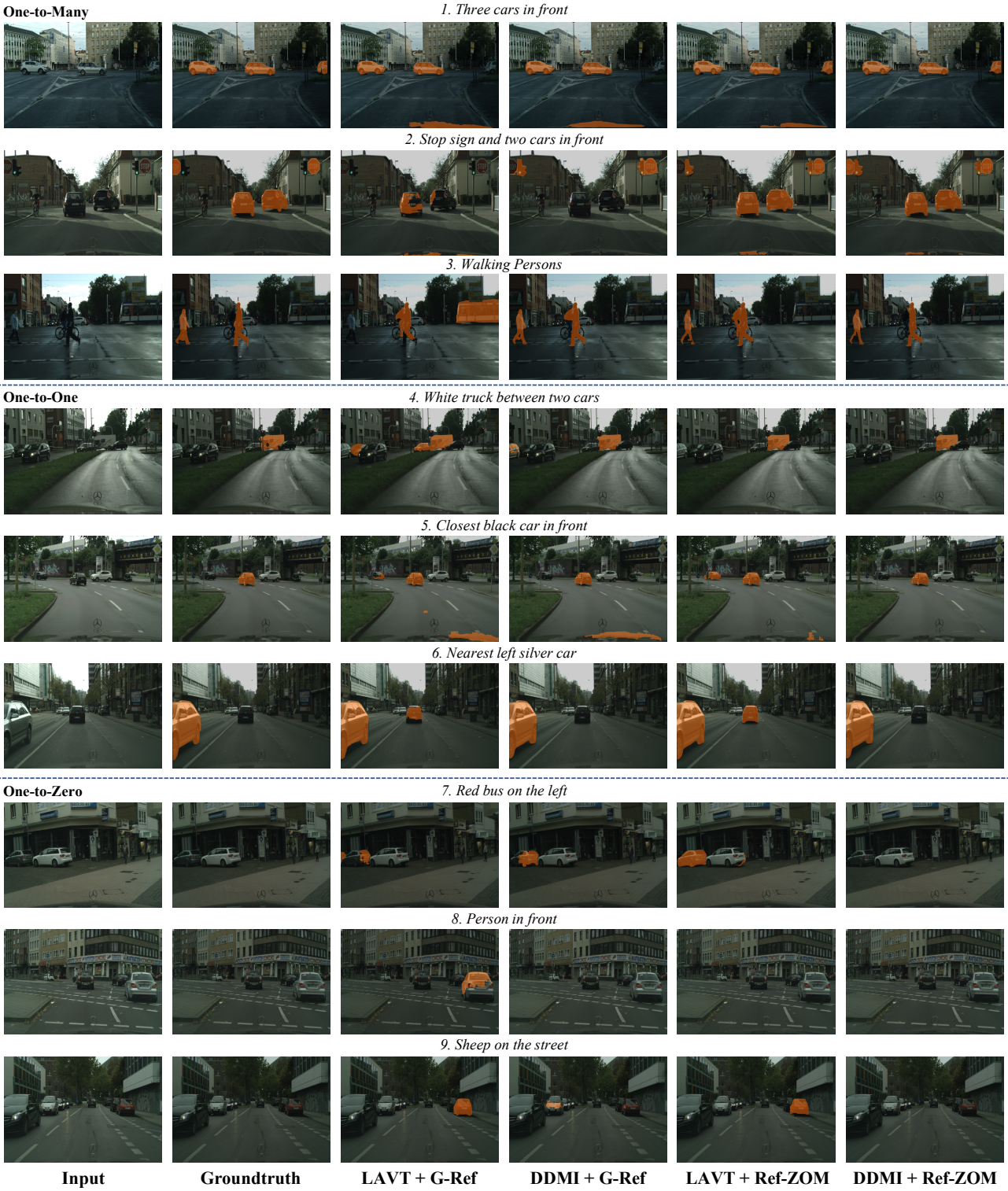


Figure 7: Illustrations of segmentation maps produced by LAVT and our DMMI when they are trained on different datasets and then directly transferred to Cityscapes without fine-tuning. Specifically, in the third and fifth columns, we present the segmentation maps of LAVT when it is trained on G-Ref and Ref-ZOM, respectively. In the fourth and sixth columns, we depict the segmentation maps of our DMMI when it is trained on G-Ref and Ref-ZOM. From top to bottom, these cases are under one-to-many, one-to-one and one-to-zero settings, respectively.