Appendix

A.1 List of Datasets

	semantic	instance	panoptic	grounding	part	training	# images
ADE-150	✓	✓	✓				2000
Pascal VOC	✓						1449
Pascal Context-59	✓						5105
Pascal-Panoptic-Parts	✓	✓	✓		✓	*	10103
COCO	✓	✓	✓			✓	121408
RefCOCO				✓		✓	19994
RefCOCO+				✓		✓	19992
RefCOCOg				✓		✓	26711

Table A1: List of the dataset used. The checkmarks denote whether a dataset has a particular type of annotation and whether the dataset is used in the training process. * Because of a data leak between Pascal-Panoptic-Parts and other Pascal datasets, we use weights trained without Pascal-Panoptic-Parts in those evaluations unless otherwise specified.

We report the statistics of datasets used in training and evaluation in table Table A1. Additionally, we further evaluate our model on 35 object detection datasets and 25 segmentation datasets in Sec. A.3.2. In total, we benchmarked our model on around 70 datasets. These benchmarks show our model can adapt to many different scenarios and retain a reasonable performance in a zero-shot manner.

A.2 Experiment Setup

A.2.1 Implementation Details

For loss functions in Eq. (3), we have $\lambda_{cls} = 2.0, \lambda_{mask} = 5.0, \lambda_{box} = 5.0, \lambda_{ce} = 1.0, \lambda_{dice} = 1.0, \lambda_{L1} = 1.0, \lambda_{giou} = 0.2$. For λ in Eq. (4), we use $\lambda = 0.2$ for seen classes during the training and $\lambda = 0.45$ for novel classes. In close-set evaluation, we set $\lambda = 0.0$ and do not use CLIP. We also do not use CLIP for PAS-21 evaluation (whose classes are mostly covered by COCO) because we find it degrades the performance. We use 800 and 1024-resolution images during the training. For evaluations, we use 1024-resolution images.

A.2.2 Training Process

Stage	Task	Dataset	Batch Size	Max Iter	Step
I	OD&IS	Objects365	64	340741	312346
II	OD&IS	COCO	32	91990	76658
	REC&RIS	RefCOCO/g/+	32		
III	PanoS	COCO	32	150000	100000,135000
	REC&RIS	RefCOCO/g/+	32		
	PartS	Pascal-Panoptic-Parts	32		

Table A2: Training Process. Following UNINEXT [61], We first pretrain our model for object detection on Object365 for 340k iteration (Stage I). Then we fine-tune our model jointly on COCO for object detection, instance segmentation, referring expression comprehension (REC), and referring segmentation (RIS) for 92k iteration (Stage II). We further jointly train our model on Panoptic Segmentation, REC, RIS, and Part Segmentation for 150k iteration (Stage III)

We train all our models on NVIDIA-A100 GPUs with a batch size of 2 per GPU using AdamW [41] optimizer. We use a base learning rate of 0.0001 and a weight decay of 0.01. The learning rate of the backbone is further multiplied by 0.1. Following UNINEXT [61], We first pretrain our model for object detection on Object365 for 340k iteration (Stage I). Then we fine-tune our model jointly on COCO for object detection, instance segmentation, referring expression comprehension (REC), and referring segmentation (RIS) for 91k iteration (Stage II). We further jointly train our model on Panoptic Segmentation, REC, RIS, and Part Segmentation for 150k iteration (Stage III). In Stage I, the learning rate is dropped by a factor of 10 after 312k iterations. In stage II, the learning rate is dropped by a factor of 10 after 77k iterations. In Stage III, the learning rate is dropped by a factor of 10 after 100k and 135k iterations. In all stages, we sample uniformly across datasets when there are multiple datasets. The global batch size is 64 in Stage I and 32 in Stage II and III. Notably, our stage I and II is identical to the setup of UNINEXT. For ablation studies, we train stage III only and reduce the schedule to 90k iterations. The learning rate schedule is also scaled accordingly. The details of training recipe is shown in Table A2.

A.3 Additional Evaluations

A.3.1 Referring Expression Comprehension

Method	Backbone	RefCOCO		RefCOCO+		RefCOCOg	
		oIoU	P@0.5	oIoU	P@0.5	oIoU	P@0.5
MAttNet [63]	RN101	56.5	76.7	46.7	65.3	47.6	66.6
VLT [9]	Dark56	65.7	76.2	55.5	64.2	53.0	61.0
RefTR [43]	RN101	74.3	85.7	66.8	77.6	64.7	82.7
UNINEXT [61]	RN50	77.9	89.7	66.2	79.7	70.0	84.0
UNINEXT [61]	ViT-H	82.2	92.6	72.5	85.2	74.7	88.7
HIPIE	RN50	78.3	90.1	66.2	80.0	69.8	83.6
HIPIE	ViT-H	82.6	93.0	73.0	85.5	75.3	88.9
vs. prev. SOTA		**+0.4**	**+0.4**	**+0.5**	**+0.3**	**+0.6**	**+0.2**

Table A3: Comparison on the referring expression comprehension (REC), and referring image segmentation (RIS) tasks. The evaluation is carried out on the validation sets of RefCOCO, RefCOCO+, and RefCOCOg datasets using Precision@0.5 and overall IoU (oIoU) metrics for REC and RIS, respectively.

In addition to Referring Segmentation reported in Table 6, we further report results on Referring Expression Comprehension (REC). We establish new state-of-the-art performance by an average of $+0.3$ P@0.5 and $+0.5$ oIoU across three datasets.

A.3.2 Object Detection and Segmentation in the Wild

To further examine the open-vocabulary capability of our model, we evaluate it on the Segmentation in the Wild (SeginW) [69] consisting of 25 diverse segmentation datasets and Object Detection in the Wild (OdinW) [32] Benchmark consisting of 35 diverse detection datasets. Since OdinW benchmark contains Pascal VOC and some of the classes in SeginW benchmark are covered by Pascal-Panoptic-Parts, we use a version of our model that is not trained on Pascal-Panoptic-Parts for both benchmarks for a fair comparison.

We report the results in Table A6 and Table A7. Notably, our method establishes a new state-of-the-art of SeginW benchmark by an average of $+8.9$ mAP across 25 datasets. We achieve comparable performance under similar settings. In particular, our ResNet-50 baseline outperforms GLIP-T by $+3.1$ mAP. We note that other methods such as GroundingDINO [39] achieve better absolute performance by introducing more grounding data, which can be critical in datasets whose classes are not common objects. (For example, the classes of Boggle Boards are letters, the classes of UnoCards are numbers, and the classes of websiteScreenshots are UI elements).

A.4 Other Ablation Studies

We provide further ablations on a few design choices in this section.

Text Encoder. We experiment with replacing the BERT text encoder in UNINEXT with a pre-trained CLIP encoder. Additionally, following practices of ODISE [59], we prompt each label to a sentence "a photo of <label>". For RIS and REC tasks, the language expression remains unchanged. We report the result in Table A4. We find that while CLIP and BERT achieve similar performance on panoptic

	COCO		RefCOCO
	PQ	AP^{Mask}	oIoU
CLIP	**51.5**	44.3	48.7
BERT	51.3	**44.4**	**77.3**

Table A4: Ablation Studies on the choice of Text Encoder. We find that while CLIP and BERT achieve similar performance on panoptic and instance segmentation, BERT performs significantly better on Referring Instance Segmentation (+28.6 oIoU).

	COCO		RefCOCO
	PQ	AP^{Mask}	oIoU
w/o OTA	50.9	43.6	76.3
w/ OTA	**51.3**	**44.4**	**77.3**

Table A5: Ablation Studies on the SimOTA matching process. Introducing SimOTA leads to performance improvement in all evaluation metrics.

and instance segmentation, BERT performs significantly better on referring instance segmentation (+28.6 oIoU). We hypothesize that this may be caused by the lack of explicit language-focused training which can help achieve a better understanding of referring expression.

SimOTA.Following UNINEXT [61] we adopted simOTA in the matching process for "thing" classes during the training. We experiment with removing simOTA matching and use standard one-to-one matching instead. We report the result in Table A5. We find that simOTA improves the performance on both panoptic segmentation and referring instance segmentation.

A.5 Limitations

We've showcased experimental evidence supporting our method across a diverse set of tasks, including open vocabulary panoptic and semantic segmentation, instance and referring segmentation, and object detection. However, it will be crucial for future work to test our methodology on video-related tasks, such as object tracking and segmentation, to draw comparisons with other universal models like UNINEXT [61]. Furthermore, it's worth considering additional pretraining of our vision encoder on newer, more complex datasets that encompass a vast amount of masks and information. For instance, SA-1B [27], which includes over 1 billion masks, would serve as an ideal training ground. Lastly, it would be advantageous to measure the change in performance when training on supplementary hierarchical datasets. Such an approach will allow us to demonstrate more varied object part segmentations, thereby expanding the capabilities and versatility of our model.

A.6 Broader Impact

Our research introduces a potent approach to hierarchical and universal open vocabulary image segmentation, aiming to address the ever-increasing demand for more data and advanced model architectures. As the demand increases, practical methodologies such as universal segmentation are projected to play a vital role in constructing and training foundational models. Our model, HIPIE, shows promise for fostering progress in a multitude of fields where hierarchical data are critical, including self-driving cars, manufacturing, and medicine. However, it's imperative to acknowledge potential limitations. Given that our model is trained on human annotations and feedback, it can inadvertently replicate any errors or biases present in the datasets. The architecture's complexity is further enhanced when multiple models are integrated, which, in turn, compromises the explainability of the final predictions. Therefore, as with the introduction of any novel technology, it's crucial to implement safety protocols to mitigate misuse. This includes mechanisms for ensuring the accuracy of inputs and establishing procedures to comprehend the criteria the model employs for predictions. By doing so, we can improve the model's reliability and mitigate potential issues.

A.7 Qualitative Results

A.7.1 More Visualizations

We provide more visualizations of panoptic segmentation, part segmentation and referring segmentation in Figs. A1 to A3.

A.7.2 Combining with SAM

We integrate our model with the mask outputs generated by the ViT-H Image encoder from Segment Anything (SAM) [27]. The encoder is trained on SA-1B which encompasses a broad spectrum of objects and masks within each image, enabling us to enhance our segmentation output by utilizing the high-quality masks from the SAM encoder to generate finer, more detailed masks.

To elaborate, in the context of panoptic segmentation, we implement a voting scheme between our pixel-wise annotations and the masks from Segment Anything (SAM), enriching these masks with our labels. For objects where our model demonstrates a strong understanding of hierarchy, such as "person" or "bird", we substitute the SAM masks with ours. This approach enables us to optimize hierarchical outcomes in the face of highly complex images.

Based on our observations from the figures, it's evident that Grounding DINO generates instance segmentation bounding boxes and subsequently uses SAM for the application of the segmentation masks. While this method proves effective for most datasets, SA-1B is a highly complex set featuring a vast array of whole objects, parts and subparts. Our qualitative findings suggest that the a single granularity instance segmentation model may fail to fully capture all objects/parts within an image or may incorrectly identify them. This consequently leads to SAM receiving sub-optimal bounding boxes for segmentation, resulting in fewer and less accurate masks (see third columns in Figs. A4 to A8). In contrast, our methodology (see last columns in Figs. A4 to A8) integrates the SAM encoder masks with our annotations and hierarchical masks wherever feasible. This results in a significantly more fine-grained and accurate output, proving superior in handling complex datasets such as SA-1B.

A.7.3 Combining with Stable Diffusion

As an interesting experiment, we combined our model with image generation model Stable-Diffusion[49] in Fig. A9. Given a source expression and target prompt, we first use HIPIE's segmentation capability to find the corresponding masks, which are then used for image inpainting. Notably, our model can uniquely achieve fine-grained control over object parts by providing part segmentation masks.

	HIPIE (H)	X-Decoder(L)[69]
Mean	**41.2**	32.3
Median	**45.1**	22.3
Airplane-Parts	**14.0**	13.1
Bottles	**45.1**	42.1
Brain-Tumor	1.9	**2.2**
Chicken	**46.5**	8.6
Cows	**50.1**	44.9
Electric-Shaver	**76.1**	7.5
Elephants	**68.6**	66.0
Fruits	61.1	**79.2**
Garbage	31.2	**33.0**
Ginger-Garlic	**24.3**	11.6
Hand	**94.2**	75.9
Hand-Metal	**64.0**	42.1
House-Parts	6.8	**7.0**
HouseHold-Items	**53.4**	53.0
Nutterfly-Squireel	**79.7**	68.4
Phones	7.0	15.6
Poles	6.7	20.1
Puppies	**64.6**	59.0
Rail	2.2	**2.3**
Salmon-Fillet	**41.8**	19.0
Strawberry	**81.5**	67.1
Tablets	8.8	**22.5**
Toolkits	**17.9**	9.9
Trash	**31.2**	22.3
Watermelon	**50.6**	13.8

Table A6: Segmentation Result on SeginW benchmark across 25 datasets. We report mAP. We outperform X-Decoder by a large margin (+8.9)

	HIPIE		GLIP-T [32]	MDETR[24]
	ViT-H	R50	Swin-T	EffNet-B5
Pretraining Data	O365,COCO,RefCOCO		O365	GOLDG,RefCOCO
Mean	**17.9**	14.5	11.4	10.7
Median	**5.5**	3.9	1.6	3.0
AerialMaritimeDrone_large	**10.9**	5.2	8.3	0.6
AerialMaritimeDrone_tiled	16.6	9.6	**17.1**	5.4
AmericanSignLanguageLetters	2.8	**2.9**	0.1	0.3
Aquarium	**18.3**	8.6	16.0	1.7
BCCD	**8.0**	6.0	1.7	6.7
boggleBoards	**0.1**	0.0	0.0	0.0
brackishUnderwater	**2.7**	0.9	1.7	0.7
ChessPieces	**5.5**	3.8	0.0	3.0
CottontailRabbits	**75.7**	69.5	57.0	66.5
dice_mediumColor	0.3	**0.5**	0.5	0.0
DroneControl	**1.6**	0.7	0.1	3.8
EgoHands_generic	**6.6**	5.8	1.1	5.9
EgoHands_specific	0.5	0.2	0.1	**3.5**
HardHatWorkers	1.8	1.4	**2.7**	0.4
MaskWearing	**1.1**	0.8	0.6	0.4
MountainDewCommercial	8.5	**37.7**	15.3	3.0
NorthAmericaMushrooms	**42.7**	27.4	5.9	39.8
openPoetryVision	0.0	0.0	0.0	0.0
OxfordPets_by-breed	7.2	**7.8**	0.3	0.0
OxfordPets_by-species	**2.7**	2.5	1.6	0.7
Packages	56.2	**68.1**	58.3	63.6
Pascal VOC	**66.0**	58.6	51.2	5.6
Pistols	**66.8**	36.4	31.6	15.9
PKLot	**2.6**	1.1	0.0	0.0
plantdoc	3.6	**3.7**	1.6	0.5
Pothole	2.9	3.9	1.6	**12.7**
Raccoon	49.7	33.4	6.2	**50.6**
selfdrivingCar_fixedLarge_export_	7.3	5.3	**7.4**	2.8
ShellfishOpenImages	**49.6**	27.5	15.9	8.1
ThermalCheetah	0.3	0.5	0.2	**4.5**
thermalDogsAndPeople	**53.3**	24.5	38.7	42.8
UnoCards	0.0	0.0	0.0	0.0
Vehicles-OpenImages	53.5	53.9	**55.0**	13.4
websiteScreenshots	0.4	0.3	0.3	**0.7**
WildfireSmoke	0.3	0.0	0.0	12.5

Table A7: Object Detection Result in OdinW benchmark. We report mAP. We achieve comparable performance under similar settings. In particular, our ResNet-50 baseline outperforms GLIP-T by $+3.1$.

Figure A1: More visualizations showcasing panoptic segmentation, part segmentation, subpart segmentation, and referring segmentation results on RefCOCO. It is recommended to view the results in color and zoom in for better detail.

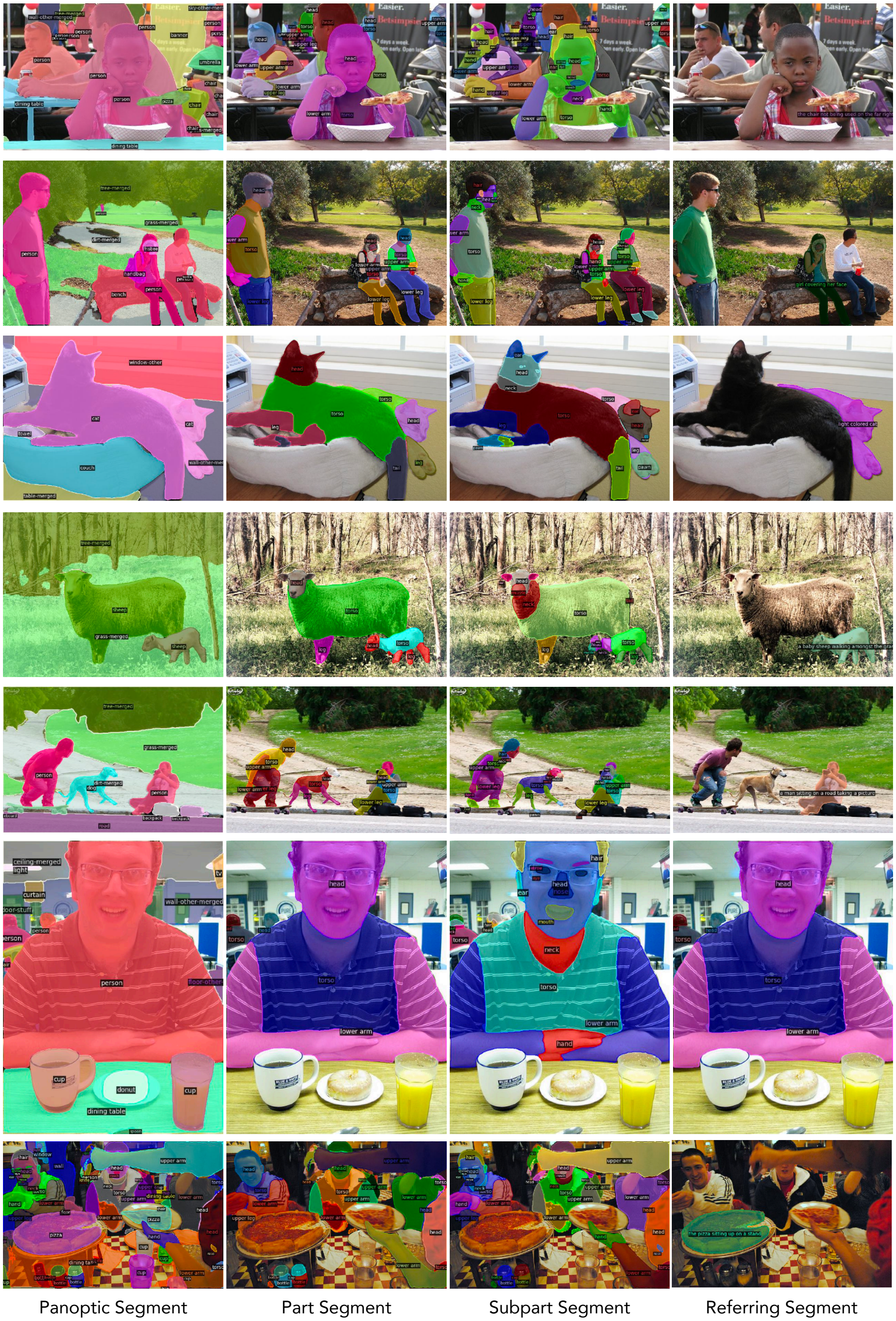

Figure A2: More visualizations showcasing panoptic segmentation, part segmentation, subpart segmentation, and referring segmentation results on RefCOCO. It is recommended to view the results in color and zoom in for better detail.

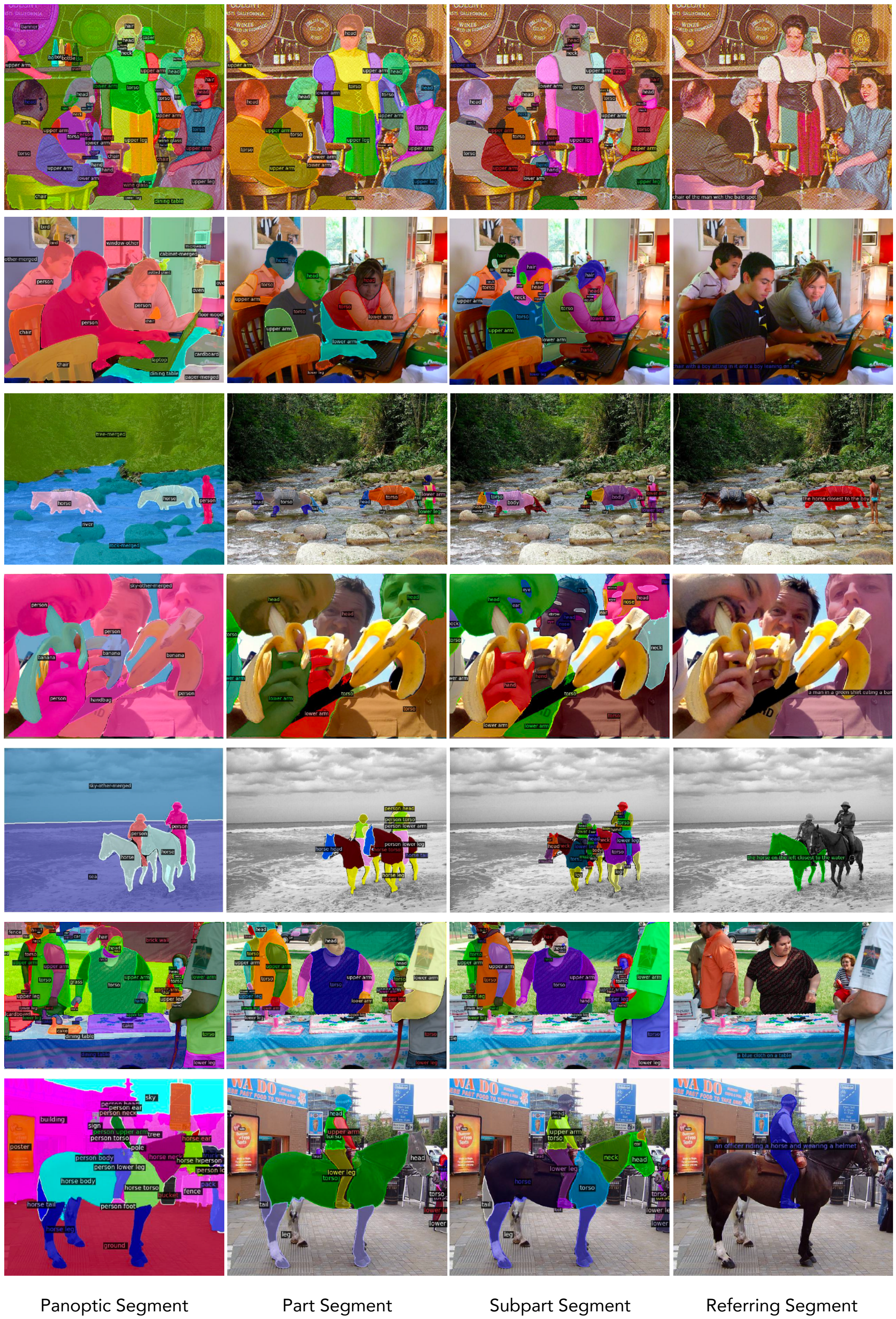

Figure A3: More visualizations showcasing panoptic segmentation, part segmentation, subpart segmentation, and referring segmentation results on RefCOCO. It is recommended to view the results in color and zoom in for better detail.

Figure A4: Results of merging HIPIE with SAM for hierarchical segmentation. By integrating the part masks from our model and conducting a vote among SAM's panoptic masks, we generate finely detailed mask outputs. Our method demonstrates fewer misclassifications and overlooked masks across the SA-1B dataset compared to the Grounding DINO + SAM approach. Furthermore, our technique excels in differentiating between intra-class objects and identifying distinct object parts.

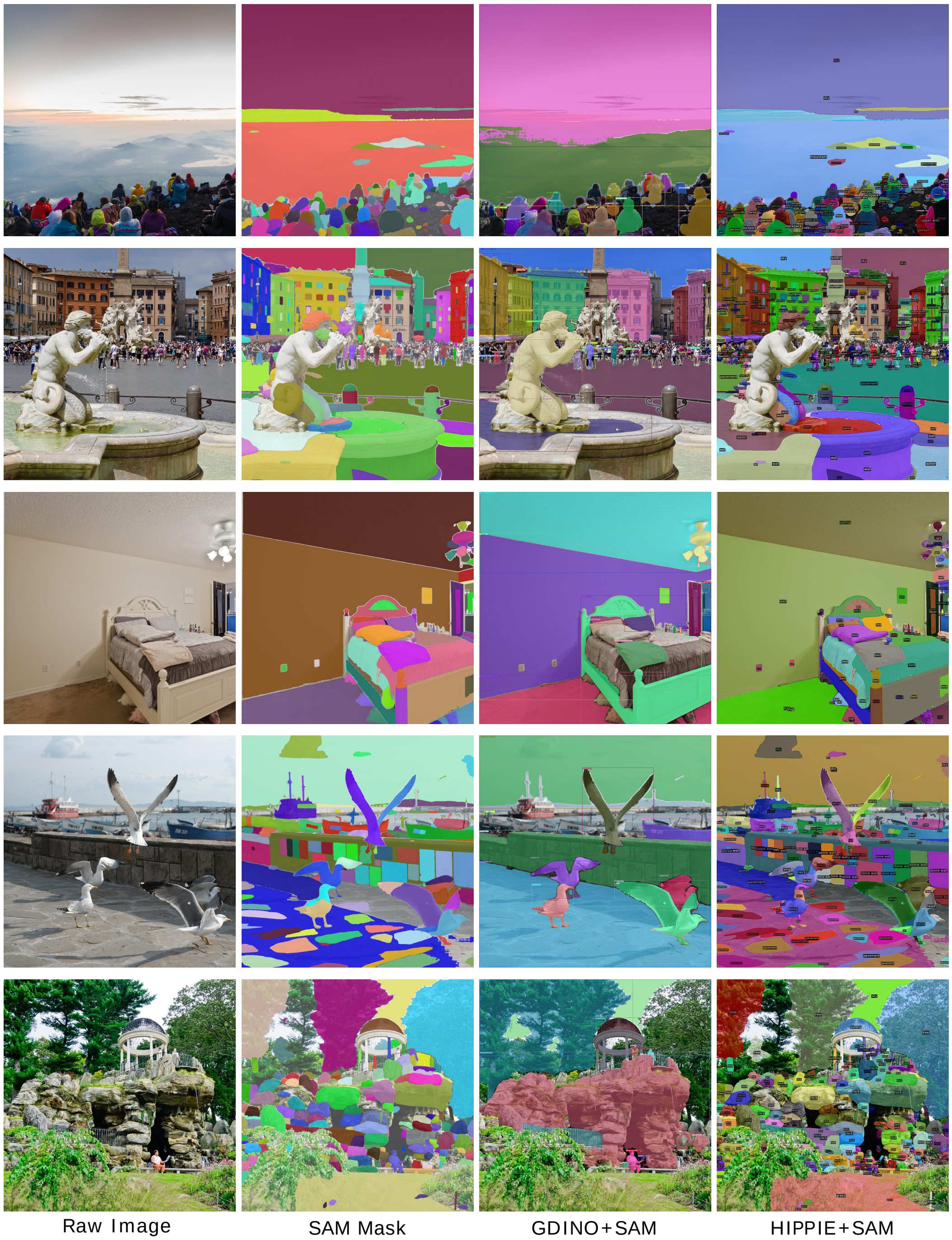

Figure A5: Additional results of merging HIPIE with SAM for hierarchical segmentation. By integrating the part masks from our model and conducting a vote among SAM's panoptic masks, we generate finely detailed mask outputs. Our method demonstrates fewer misclassifications and overlooked masks across the SA-1B dataset compared to the Grounding DINO + SAM approach. Furthermore, our technique excels in differentiating between intra-class objects and identifying distinct object parts.

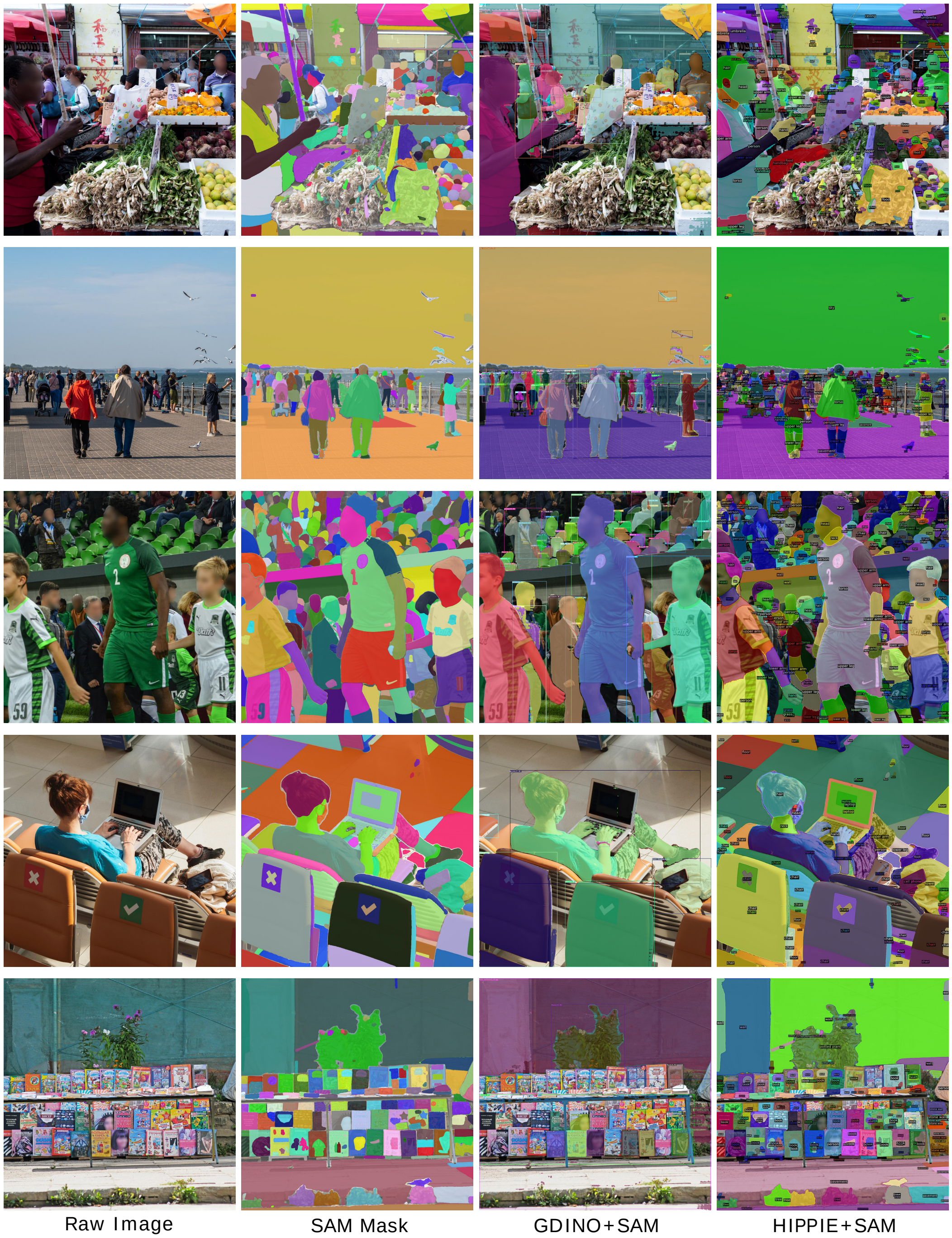

Figure A6: Additional results of merging HIPIE with SAM for hierarchical segmentation. By integrating the part masks from our model and conducting a vote among SAM's panoptic masks, we generate finely detailed mask outputs. Our method demonstrates fewer misclassifications and overlooked masks across the SA-1B dataset compared to the Grounding DINO + SAM approach. Furthermore, our technique excels in differentiating between intra-class objects and identifying distinct object parts.

Raw Image SAM Mask GDINO+SAM HIPPIE+SAM

Figure A7: Additional results of merging HIPIE with SAM for hierarchical segmentation. By integrating the part masks from our model and conducting a vote among SAM's panoptic masks, we generate finely detailed mask outputs. Our method demonstrates fewer misclassifications and overlooked masks across the SA-1B dataset compared to the Grounding DINO + SAM approach. Furthermore, our technique excels in differentiating between intra-class objects and identifying distinct object parts.

Figure A8: Additional results of merging HIPIE with SAM for hierarchical segmentation. By integrating the part masks from our model and conducting a vote among SAM's panoptic masks, we generate finely detailed mask outputs. Our method demonstrates fewer misclassifications and overlooked masks across the SA-1B dataset compared to the Grounding DINO + SAM approach. Furthermore, our technique excels in differentiating between intra-class objects and identifying distinct object parts.

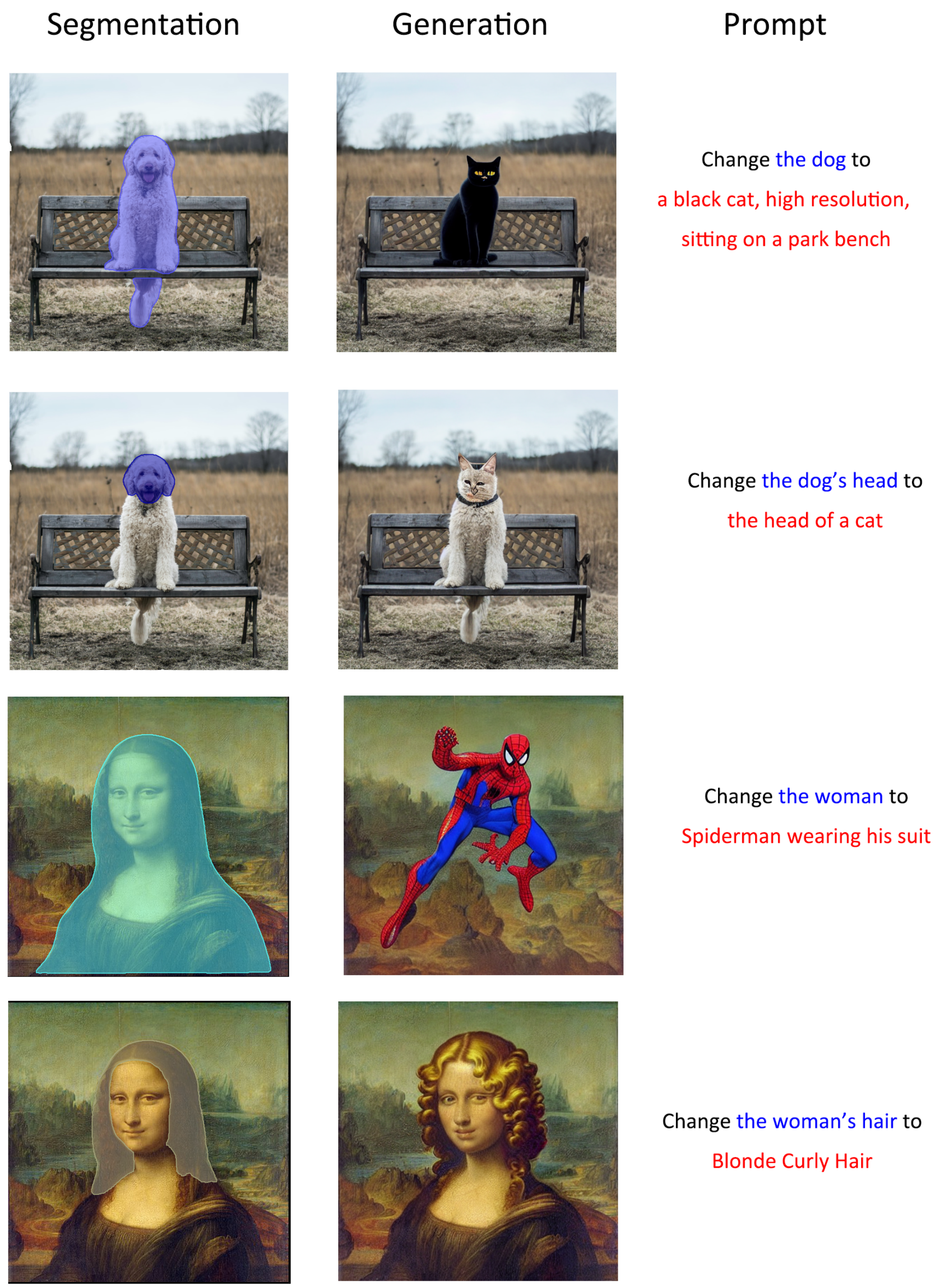

Figure A9: Results of combining HIPIE with Stable Diffusion for Image inpainting. We leverage our segmentation model to generate masks for the redrawing process. Our model can uniquely achieve fine-grained control by providing part segmentation masks.

References

[1] E. H. Adelson. On seeing stuff: the perception of materials by humans and machines. In *Human vision and electronic imaging VI*, volume 4299, pages 1–12. SPIE, 2001.

[2] M. Bucher, T.-H. Vu, M. Cord, and P. Pérez. Zero-shot semantic segmentation. In *NeurIPS*, 2019.

[3] Z. Cai and N. Vasconcelos. Cascade r-cnn: high quality object detection and instance segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1483–1498, 2019.

[4] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar. Masked-attention mask transformer for universal image segmentation. 2022.

[5] D. de Geus, P. Meletis, C. Lu, X. Wen, and G. Dubbelman. Part-aware panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5485–5494, 2021.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] N. Dhanachandra, K. Manglem, and Y. J. Chanu. Image segmentation using k-means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science*, 54:764–771, 2015.

[8] H. Ding, C. Liu, S. Wang, and X. 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.

[9] H. Ding, C. Liu, S. Wang, and X. Jiang. Vlt: Vision-language transformer and query generation for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[10] Z. Ding, J. Wang, and Z. Tu. Open-vocabulary panoptic segmentation with maskclip. *arXiv preprint arXiv:2208.08984*, 2022.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

[12] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

[13] G. Feng, Z. Hu, L. Zhang, and H. Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15506–15515, 2021.

[14] D. A. Forsyth and J. Ponce. *Computer vision: a modern approach*. prentice hall professional technical reference, 2002.

[15] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[16] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin. Scaling open-vocabulary image segmentation with image-level labels. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 540–557. Springer, 2022.

[17] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[18] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[20] R. Hu, M. Rohrbach, and T. Darrell. Segmentation from natural language expressions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[21] T. Hui, S. Liu, S. Huang, G. Li, S. Yu, F. Zhang, and J. Han. Linguistic structure guided context modeling for referring image segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 59–75. Springer, 2020.

[22] S. K. Jagadeesh, R. Schuster, and D. Stricker. Multi-task fusion for efficient panoptic-part segmentation. *arXiv preprint arXiv:2212.07671*, 2022.

[23] Y. Jing, T. Kong, W. Wang, L. Wang, L. Li, and T. Tan. Locate then segment: A strong pipeline for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9858–9867, 2021.

[24] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.

[25] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019.

[26] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[27] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[28] H. W. Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[29] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022.

[30] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13619–13627, 2022.

[31] F. Li, H. Zhang, H. xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation, 2022.

[32] L. H. Li*, P. Zhang*, H. Zhang*, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, K.-W. Chang, and J. Gao. Grounded language-image pre-training. In *CVPR*, 2022.

[33] M. Li and L. Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in neural information processing systems*, 34:19652–19664, 2021.

[34] Y. Li, H. Mao, R. Girshick, and K. He. Exploring plain vision transformer backbones for object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 280–296. Springer, 2022.

[35] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. *arXiv preprint arXiv:2210.04150*, 2022.

[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[37] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[38] J. Liu, H. Ding, Z. Cai, Y. Zhang, R. K. Satzoda, V. Mahadevan, and R. Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. *arXiv preprint arXiv:2302.07387*, 2023.

[39] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[40] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[41] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

[42] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014.

[43] L. Muchen and S. Leonid. Referring transformer: A one-step approach to multi-task visual grounding. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[44] V. K. Nagaraja, V. I. Morariu, and L. S. Davis. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 792–807. Springer, 2016.

[45] R. Nock and F. Nielsen. Statistical region merging. *IEEE Transactions on pattern analysis and machine intelligence*, 26(11):1452–1458, 2004.

[46] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[47] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, and J. Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022.

[48] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

[49] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[50] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[51] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019.

[52] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017.

[53] R. Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022.

[54] Z. Wang, Y. Lu, Q. Li, X. Tao, Y. Guo, M. Gong, and T. Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022.

[55] J. Wu, X. Li, X. Li, H. Ding, Y. Tong, and D. Tao. Towards robust referring image segmentation. *arXiv preprint arXiv:2209.09554*, 2022.

[56] Y. Xian, S. Choudhury, Y. He, B. Schiele, and Z. Akata. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8256–8265, 2019.

[57] J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022.

[58] J. Xu, J. Hou, Y. Zhang, R. Feng, Y. Wang, Y. Qiao, and W. Xie. Learning open-vocabulary semantic segmentation models from natural language supervision. *arXiv preprint arXiv:2301.09121*, 2023.

[59] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. *arXiv preprint arXiv:2303.04803*, 2023.

[60] M. Xu, Z. Zhang, F. Wei, Y. Lin, Y. Cao, H. Hu, and X. Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 736–753. Springer, 2022.

[61] B. Yan, Y. Jiang, J. Wu, D. Wang, P. Luo, Z. Yuan, and H. Lu. Universal instance perception as object discovery and retrieval. *arXiv preprint arXiv:2303.06674*, 2023.

[62] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. 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.

[63] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, 2018.

[64] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016.

[65] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[66] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu. Unleashing text-to-image diffusion models for visual perception. *arXiv preprint arXiv:2303.02153*, 2023.

[67] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019.

[68] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[69] X. Zou, Z.-Y. Dou, J. Yang, Z. Gan, L. Li, C. Li, X. Dai, H. Behl, J. Wang, L. Yuan, et al. Generalized decoding for pixel, image, and language. *arXiv preprint arXiv:2212.11270*, 2022.

[70] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Gao, and Y. J. Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023.