AIMS: All-Inclusive Multi-Level Segmentation Supplementary Material

Lu Qi[1] **Jason Kuen**[2] **Weidong Guo**[3*] **Jiuxiang Gu**[2]
Zhe Lin[2] **Bo Du**[4] **Yu Xu**[3] **Ming-Hsuan Yang**[1,5]
[1]UC Merced [2]Adobe Research [3]QQ Brower Lab, Tencent
[4] Wuhan University [5] Google Research

In this supplementary file, we provide more experimental details and empirical results to further demonstrate the benefits of our proposed All-Inclusive Multi-Level segmentation task (AIMS).

- Experimental details on dataset split, sampling strategy and SAM's [1] configuration.
- More ablation studies on sample strategy and dataset usage.
- More visualization results of our AIMS model with various inference modes.
- User study of AIMS and the concurrent work SAM [1]

The code and models to reproduce our experiments will be released.

1 Experimental Details

Dataset Split As outlined in the main paper, all models are trained using five datasets: COCO [2], EntitySeg [3], PascalVOC Part (PPP)[4], PACO[5], and COCO-PSG [6]. Given that the original training and validation splits of these datasets are tailored for single tasks, we collate the images and reorganize them to suit our AIMS task. Initially, we select 1069 and 1000 validation images from PPP [4] (which covers the part and entity levels) and COCO-PSG [6] (which covers the entity and relation levels) respectively. Following this, we eliminate any duplicate images in the unified training set that are present in the validation images, resulting in a refined training set comprised of approximately 236.7K unique images.

Sampling Strategy In Table 2, we present the eight different sample types, labeled as Sample ID 1 through 8, for each iteration, assuming a batch size of 8. Additionally, images for each of these sample types are uniformly selected from the datasets mentioned in the 'Dataset' column.

SAM's configuration To obtain more fine-grained part-level predictions, we follow the hyper-parameters found on SAM's GitHub page: *points_per_side* (2), *pred_iou_thresh* (0.86), *stability_score_thresh* (0.92), *crop_n_layers* (1), *crop_n_points_downscale_factor* (2), *min_mask_region_area* (100).

2 Ablation Studies

Sample Strategy Table 2 presents an ablation study on various sample strategies by examining the impact of each sample type during each iteration. The first row serves as our baseline, adhering to our base framework with a split decoder for varying-level predictions, and it excludes the usage of a prompt mask encoder. The remaining rows illustrate the influence on performance when task prompts are employed during each training iteration.

*Corresponding author.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Sample ID	Mask Prompt	Decoder	Dataset
1	Full Image	Entity	COCO, EntitySeg, PPP, COCO-PSG
2		Part	PPP
3		Relation	COCO-PSG
4	Partial Image	Entity	COCO, EntitySeg, PPP, COCO-PSG, PACO
5		Part	PPP, PACO
6		Relation	COCO-PSG
7	One Entity	Part	PPP, PACO
8	Two Entities	Relation	COCO-PSG

Table 1: The illustration of eight sample types on each training iteration.

Sample ID								PPP (Inference)			COCO-PSG (Inference)		
1	2	3	4	5	6	7	8	AP^P	AP^E	AR^{EP}	AP^R	AP^E	AR^{RE}
○	○	○	○	○	○	○	○	24.5	53.4	69.7	38.9	40.4	50.9
✓	✓	✓	○	○	○	○	○	25.0	53.5	69.8	39.2	40.8	51.3
○	○	○	✓	✓	✓	○	○	22.5	51.4	67.3	37.1	37.6	48.9
✓	✓	✓	✓	✓	✓	○	○	25.7	54.4	70.9	39.6	41.2	51.8
✓	✓	✓	✓	✓	✓	✓	○	26.4	55.9	72.1	39.6	41.3	51.8
✓	✓	✓	✓	✓	✓	○	✓	25.7	54.3	71.0	40.4	42.0	53.0
✓	✓	✓	✓	✓	✓	✓	✓	**26.5**	**56.1**	**72.3**	**40.5**	**42.1**	**53.1**

Table 2: The ablation studies on eight distinct sample types used in each training iteration. The ✓ and ○ symbols are employed to indicate whether a specific sample type is utilized. For a fair comparison, we adjust the learning rate linearly in relation to the batch size. The default learning rate is 1e-8 for a batch size of 8.

Dataset					PPP (Inference)			COCO-PSG (Inference)		
COCO	COCO-PSG	PPP	PACO	EntitySeg	AP^P	AP^E	AR^{EP}	AP^R	AP^E	AR^{RE}
✓	○	○	○	○	-	44.4	-	-	41.2	-
○	✓	○	○	○	-	44.5	-	39.6	41.7	51.8
○	○	✓	○	○	24.3	48.6	65.8	-	25.7	-
✓	✓	✓	○	○	25.7	54.4	70.9	39.6	41.2	51.8
✓	✓	✓	✓	○	26.4	55.9	72.1	39.6	41.2	51.8
✓	✓	✓	○	✓	25.7	54.3	71.0	40.4	42.0	53.0
✓	✓	✓	✓	✓	**26.5**	**56.1**	**72.3**	**40.5**	**42.1**	**53.1**

Table 3: The ablation study of performance influence with dataset usage. Similar to Table 2, we adjust the learning rate linearly considering the lack of some task prompts due to the non-provided dataset.

For instance, the second row reveals that the exclusive usage of a full image task prompt brings a marginal performance improvement over the baseline. Since full-image mask prompts are always the same, injecting them into the original image features is identical to not using any mask prompts. Conversely, the third row shows that employing solely partial-image mask prompts can considerably deteriorate performance, which is inconsistent with the inference process that involves full-image mask prompts.

In the fourth row, we observe that using both full and partial mask prompts can further enhance performance by providing more accurate signals to the network. Additionally, the introduction of mask prompts at the entity-to-part and relation-to-entity levels consistently yields performance improvements, as depicted in subsequent rows.

Dataset Usage Table 3 displays an ablation study focusing on the utilization of various datasets. The first three rows present the model performance when trained on a single dataset. This results in degraded performance on cross-dataset evaluation due to inconsistencies in annotations across different datasets. As shown in the fourth row, combining the three datasets for training enhances the overall validation performance. This improvement is due to the consistency maintained between the training and validation splits. Incorporating PACO and Entityseg datasets at part and entity levels further improve the performance for the respective levels.

3 Visualization

The flexibility of our AIMS task. Figure 1 illustrates how our proposed AIMS task provides the flexibility for segmenting anything. This is to address the subjective annotation issues in existing datasets. For instance, in the image (a), the throw pillows and the sofa are predicted as a single entity,

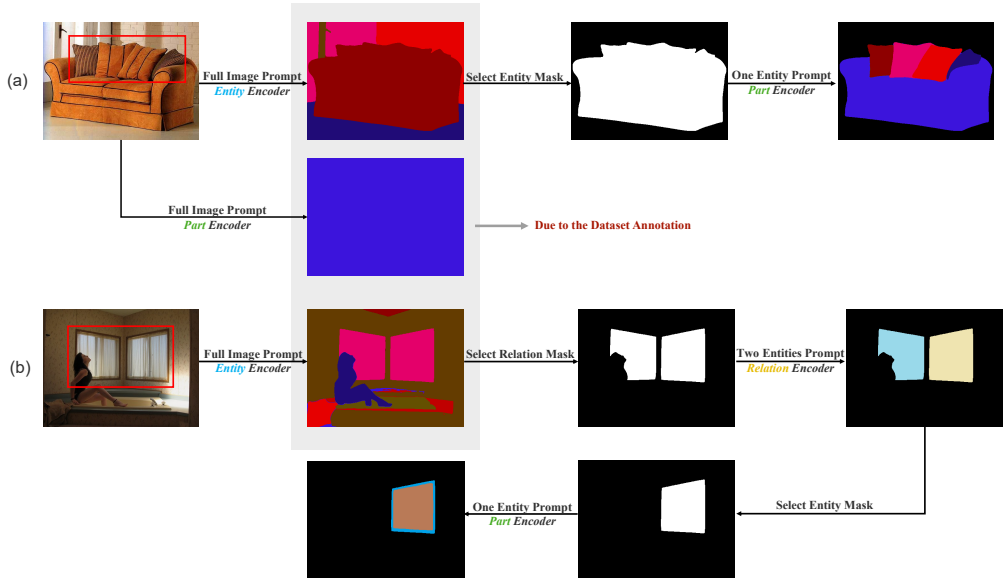

Figure 1: The illustration of the flexibility of AIMS task to tackle the subjective annotation issues in existing datasets.

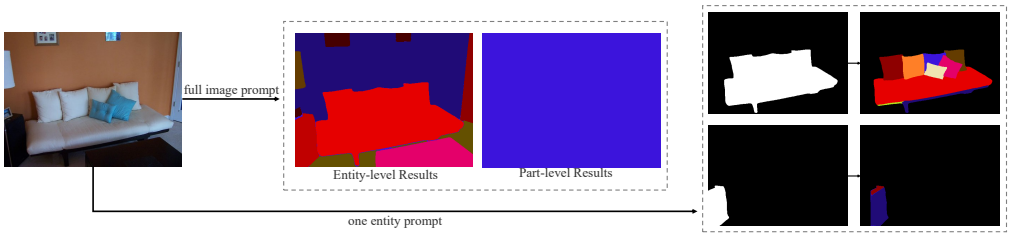

Figure 2: The example of prediction inconsistency between two prompt types on unseen classes.

which aligns with the ground truth annotation. Nonetheless, in certain scenarios, a user might wish to edit the throw pillows independently. With our AIMS method, utilizing a full image prompt for part-level prediction does not yield these separated masks. However, by selecting this entity mask as an entity prompt for part-level prediction, we can successfully differentiate the three throw pillows.

Additionally, in image (b), a user may want to segment the windows into several components. However, original ground truth annotations typically consider the two windows as a single entity. To meet this requirement, we initially utilize a full image prompt and an entity encoder to identify the mask of the whole windows. Following this, we apply this two-entity mask prompt and relation Decoder to split them into two independent window masks. Ultimately, we can select one window for further segmentation into two parts: the window edge and curtains.

Prediction consistency. We investigate the prediction consistency across two inference modes: full and partial image prompts. Figure 2 displays an example where our AIMS model fails to segment anything at the part level with a full image prompt, but it is able to break down an entity-level mask prompt into a more detailed level, similar to the sofa and throw pillows in the image. Given that the pillows have never been labeled in our utilized dataset, we surmise that our AIMS model might yield inconsistent prediction results for unseen classes when different prompts are used. Using entity-level mask prompts could help our model to drill down to the next level.

However, the scenario changes when we turn our attention to the known classes. Figure 3 illustrates that three cars yield similar part-level prediction results, regardless of whether a full image or a single

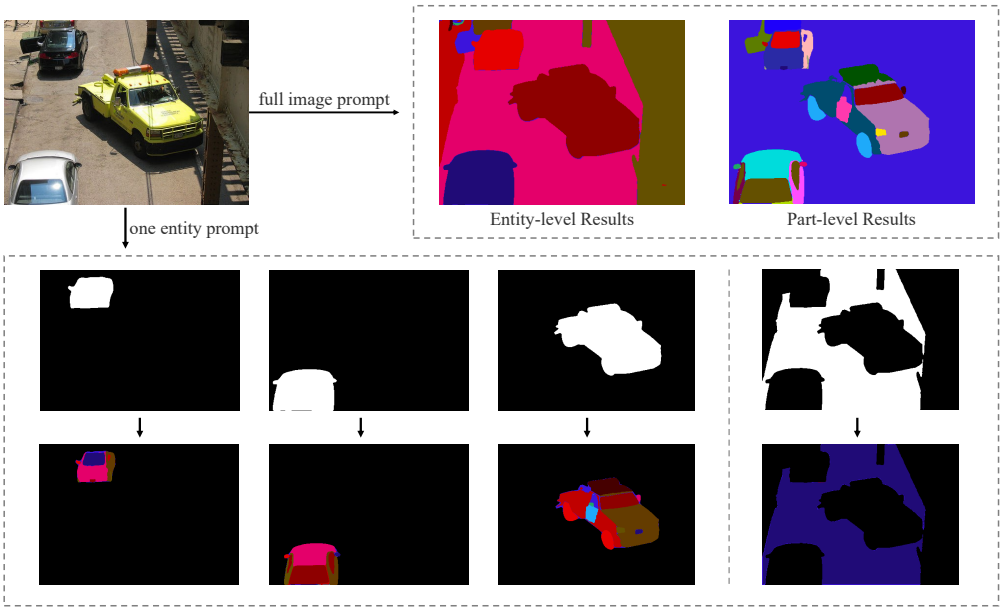

Figure 3: The example of prediction consistency between two prompt types on seen classes.

entity prompt is used. This indicates that our AIMS model can maintain prediction consistency for known classes.

The three-level predictions Figure 4 displays the three-level prediction outcomes of our AIMS model. Despite the potential subjectivity in defining the three levels, our AIMS model shows strong promises in fulfilling various user needs and intentions in image editing. For instance, the red region in the image is not detected at the entity-level prediction results, as it is merged into the tray. However, by injecting an entity-level mask prompt for the tray region into the part encoder, we see that the separate red region can be obtained, as shown in the second column of the 'part-level results'. Additionally, the relation-level masks of any two entities that share some semantic relationship can be predicted, and all relation-level prediction results can be represented in a scene graph.

More visualization results. In the Fig. 7, we show more visualization results at entity and part level in the wild (Open-Image [7] dataset), manifesting the generalization ability of our proposed AIMS model at the entity and part level. Considering the entity level is the principal part of image editing, we also show some cases on Laion400M [8] to show the effectiveness of our AIMS model.

4 User Study

We conducted a user study in which there were 480 individuals who were identified as Adobe Photoshop users who regularly used the software for image manipulation/editing. In this user study, we randomly selected 40 images in the wild and provided the users with a visualization of the three-level prediction results of our AIMS model, as shown in Figure 4. For each image, we asked each user about his/her degree of satisfaction with treating semantically meaningful and -coherent segments for three levels, with respect to their relevance to and suitability for image manipulation/editing applications. The satisfactory scores are aggregated from all users for the individual images. We find that the average score of each image is large than 7.8 on the condition that the maximum score is 10. Most of the selected images' scores are larger than 6.0, This confirms that the users are highly satisfied with our model's prediction results in the context of image manipulation/editing. To better present the user study's findings, we summarize the data of Table 4. The image IDs with the minimum, median, and maximum user scores are 20, 19, and 35. We show these images and their corresponding entity annotations in Fig. 6.

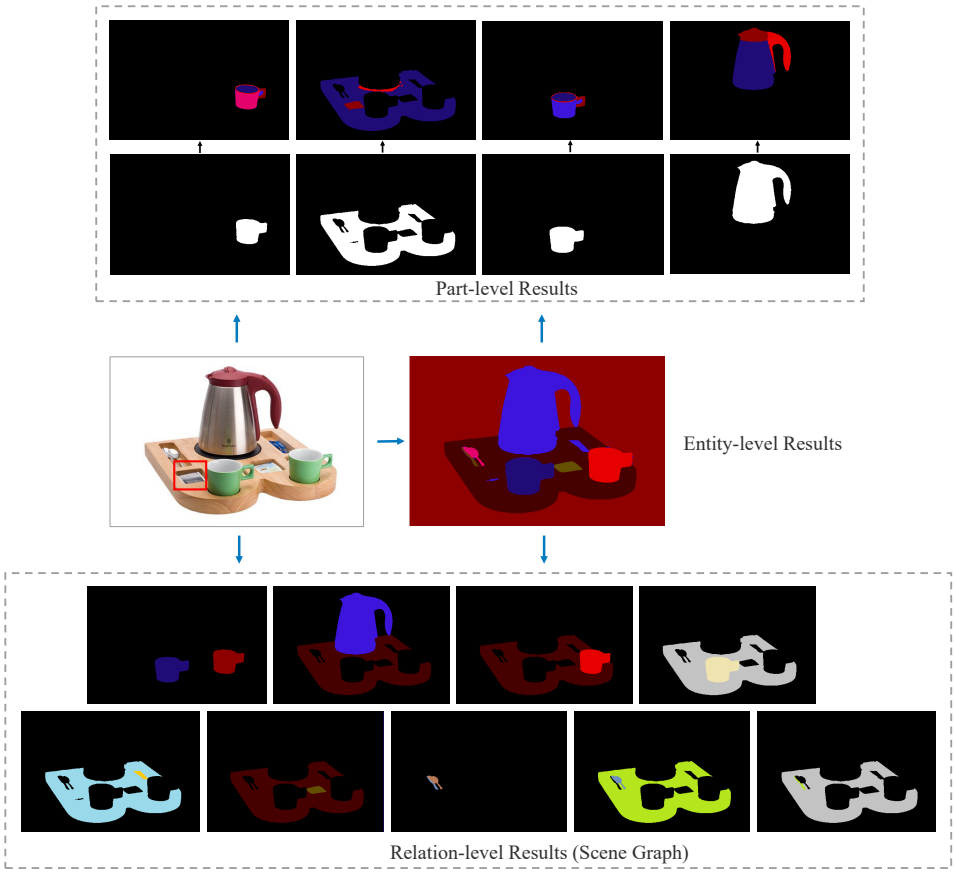

Figure 4: The illustration of how the AIMS model can be used to obtain multi-intention segmentation results, including a scene graph for physical-touch relations.

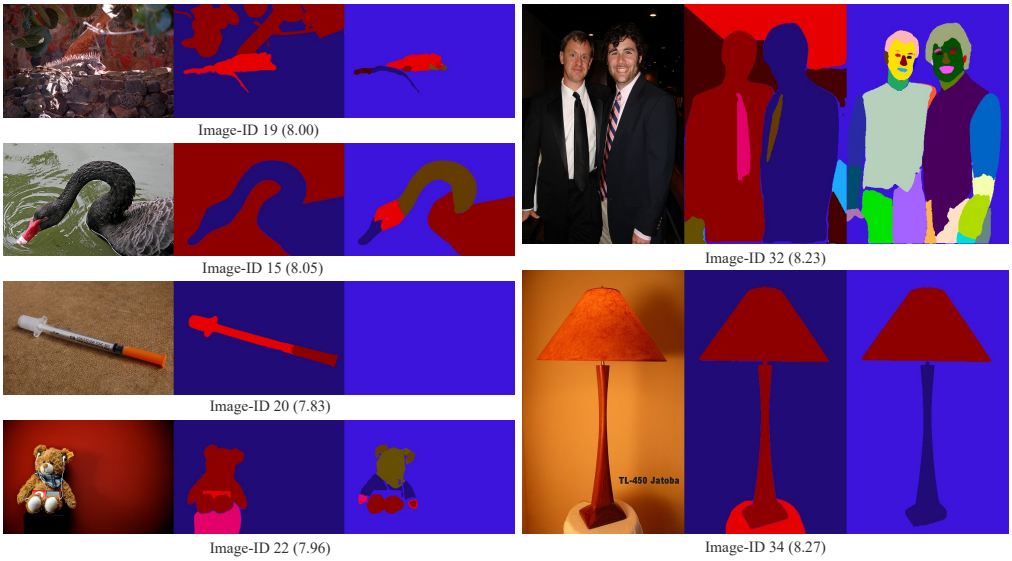

Figure 5: The selected images used in our user study. For brevity, we only show the entity- and part-level predictions the user most care.

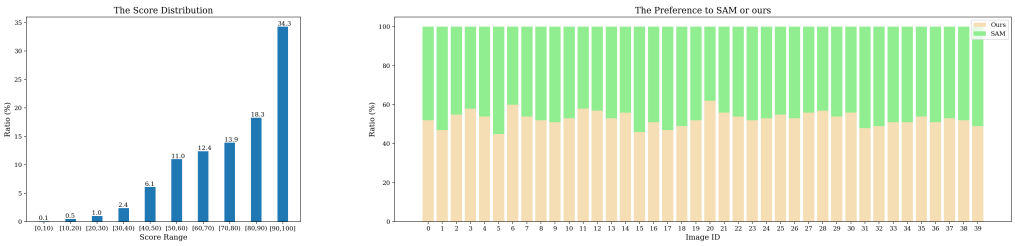

Figure 6: The statistical visualizations of the data from our user and survey studies. **(a)** The histogram here represents the distribution of the users based on their given scores in the three-level predictions. **(b)** Each horizontal bar here indicates the proportions of votes given by the survey participants to SAM [1] (green) and *ours* (orange) on each of the 40 images.

Figure 7: More visualization results on Open-Image Dataset [7] in the 'wild'.

In Fig. 6, we provide statistical information about our user study. The sub-figure (a) is a summary of Table 4, indicating that over 90% of users rated our prediction results higher than 6.0. Sub-figure (b) contrasts user preferences between our results and those from the SAM model. Despite being trained with fewer data, our model garners an appreciation comparable to that of SAM, as can be observed in the sub-figure (b).

Image ID	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	AVG
1	0.00	0.00	1.88	3.13	8.13	11.25	11.25	16.88	9.38	38.13	8.06
2	0.63	0.00	3.13	3.75	5.63	15.00	13.75	13.13	10.00	35.00	7.84
3	1.25	0.63	1.88	1.88	10.63	10.63	15.00	12.50	8.13	37.50	7.86
4	0.63	0.63	0.63	3.13	7.50	14.38	13.13	14.38	8.75	36.88	7.95
5	0.00	0.63	2.50	1.25	7.50	11.88	13.13	18.13	7.50	37.50	8.02
6	0.63	0.63	1.25	1.25	6.88	15.00	14.38	20.00	7.50	32.50	7.88
7	0.00	0.00	0.63	3.13	6.25	13.75	12.50	20.63	10.63	32.50	8.01
8	0.00	0.00	1.25	1.88	8.13	11.88	15.63	13.75	9.38	38.13	8.08
9	0.00	0.63	1.25	3.75	8.13	10.63	15.63	18.13	8.13	33.75	7.90
10	0.00	0.63	3.13	2.50	5.00	16.88	15.00	14.38	6.88	35.63	7.85
11	0.00	0.00	0.00	1.25	5.63	17.50	10.00	19.38	13.75	32.50	8.12
12	0.63	0.63	0.63	2.50	5.00	10.63	16.88	22.50	11.87	31.88	8.00
13	0.00	0.00	0.00	2.50	5.00	14.38	15.00	15.63	12.50	35.00	8.14
14	0.00	0.63	1.25	3.75	6.25	13.75	13.75	19.38	12.50	28.75	7.85
15	0.00	0.00	0.63	1.88	8.13	13.75	11.25	20.63	9.38	34.38	8.05
16	0.63	0.63	1.25	2.50	7.50	11.88	13.75	18.13	11.25	32.50	7.91
17	0.00	0.00	0.63	1.88	4.38	13.75	18.13	16.25	11.88	33.13	8.09
18	0.00	0.00	0.00	3.13	5.63	16.88	15.00	13.75	11.25	34.38	8.02
19	0.00	0.63	0.63	1.88	6.25	15.00	16.88	14.38	8.75	35.63	8.00
20	0.00	0.00	0.63	3.13	6.25	10.63	12.50	20.63	10.63	32.50	7.83
21	0.00	1.88	0.63	0.63	5.00	11.88	15.63	22.50	8.13	33.75	8.05
22	0.00	0.63	1.25	3.13	8.75	10.63	13.75	16.25	11.88	33.75	7.96
23	0.00	0.63	0.63	1.88	5.63	13.75	13.75	18.75	9.38	35.63	8.08
24	0.00	0.63	0.63	1.88	4.38	11.88	15.63	20.00	11.88	33.13	8.11
25	0.00	1.88	0.63	1.88	7.50	11.88	11.88	19.38	11.25	33.75	8.00
26	0.00	1.25	0.63	2.50	5.00	11.25	16.88	18.75	9.38	34.38	8.03
27	0.00	0.00	1.25	1.25	7.50	12.50	13.13	20.63	9.38	34.38	8.06
28	0.00	0.00	0.63	4.38	3.75	12.50	13.13	20.00	13.13	32.50	8.08
29	0.00	0.00	0.63	3.13	5.00	8.13	16.88	18.75	13.75	33.75	8.18
30	0.00	1.25	1.25	5.00	5.63	11.25	11.25	18.13	11.88	34.38	7.96
31	0.00	0.63	1.88	0.63	5.00	12.50	12.50	20.00	12.50	34.38	8.13
32	0.00	0.63	1.88	0.00	4.38	11.25	15.00	18.75	9.38	38.75	8.23
33	0.00	0.63	0.63	2.50	5.00	13.13	9.38	23.75	9.38	35.63	8.13
34	0.00	1.25	0.63	2.50	5.00	11.88	6.88	18.75	15.63	37.50	8.25
35	0.00	0.63	0.63	1.25	4.38	10.00	12.50	21.88	13.13	35.63	8.27
36	0.00	0.63	0.63	3.13	5.00	10.63	13.13	18.13	14.38	34.38	8.14
37	0.00	0.00	0.63	2.50	3.75	10.00	18.13	18.75	16.25	30.00	8.14
38	0.63	0.00	0.63	2.50	10.00	10.63	13.13	23.13	10.00	29.38	7.84
39	0.00	1.25	1.88	1.88	4.38	10.63	18.13	17.50	13.75	30.63	8.04
40	0.00	0.00	0.63	2.50	6.88	11.88	11.25	16.88	13.75	36.25	8.08

Table 4: Data from the user study on the three-level prediction results of our AIMS model. P_x indicates the percentage of users who give "x" as the score to represent their degrees of satisfaction. "x" ranges from 1 to 10, and 10 represents the highest degree of satisfaction. "AVG" indicates the the score averaged across all users for each image.

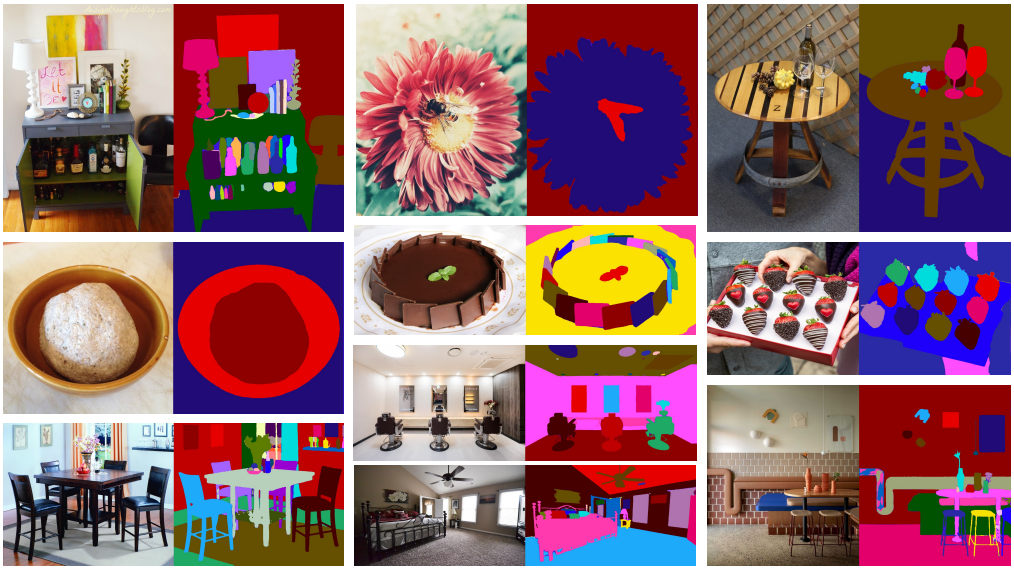

Figure 8: More visualization results on Laion400M [8] in the 'wild'.

References

[1] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *arXiv*, 2023.

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[3] Lu Qi, Jason Kuen, Weidong Guo, Tiancheng Shen, Jiuxiang Gu, Wenbo Li, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. Fine-grained entity segmentation. 2022.

[4] Daan de Geus, Panagiotis Meletis, Chenyang Lu, Xiaoxiao Wen, and Gijs Dubbelman. Part-aware panoptic segmentation. In *CVPR*, 2021.

[5] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *arXiv*, 2023.

[6] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation. In *ECCV*, 2022.

[7] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *IJCV*, 2020.

[8] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Workshop*, 2021.