Supplementary Material

A. Complementary definitions and details

Ground-truth heatmap generation. Note that we generate the ground-truth heatmap from ground-truth bounding boxes with the same method as CenterNet [29]. In particular, we apply a Gaussian kernel on the center of bounding boxes, where the kernel size is calculated according to the box size.

Auxiliary task. During training, our model will predict both faces and bodies, while in the evaluation, we only report the metrics regarding the faces.

Center sampling. Recall that in Equation (6) we define the average pooling positional encoding, now we introduce center sampling

$$\mathrm{cs}(s_i)(u, v) = \phi((\bar{u}, \bar{v})), \tag{10}$$

where it first average the coordinate of an $s_i \times s_i$ window and then encode it. When using center sampling, we replace it with $\mathrm{avgpool}(\boldsymbol{p}, s_i)(u, v)$ with it in (6). Since it discards the scale information, we adopt average pooling in our method.

Evaluation details. For the predicted bounding boxes, we apply a Non-Maximum-Suppression (NMS) IoU no more than 0.7, and retain the top-1000 predicted boxes based on confidence score. When evaluating the outpainting pipeline, we first apply the same NMS and top-1000 filter on the result of each image, then we aggregate all the remaining boxes and apply the NMS and filtering again. For the heatmap, we average the heatmaps over all images.

B. Further discussion on the outpainting baseline

Tab. 7 reports the performance of the outpainting pipeline with varying numbers of samples. Increasing the number of samples improves metrics for outside faces, but degrades CE and AP on truncated faces, revealing a trade-off inherent to this approach. A further limitation is that the pipeline is not end-to-end trainable, making each component a potential bottleneck (Figure 7). Moreover, even with strong generative models, accessing the ideal conditional distribution remains an open challenge.

C. Potential negative Societal impacts

We also note the potential for more troubling applications (dual use). Successfully detecting objects like humans faces beyond what is directly observable could serve opposing ends. Instead of directing the camera to avoid that area, extreme amodal face detection could be used to pursue unseen-but-inferred objects. The existence of such applications does not negate the ethical case for extreme amodal face detection, though, which is based on its safety, privacy, and accessibility-enhancing potential.

Top-μ^2	AP↑	AP_t↑	AP_o↑	AP_{o+}↑	AP_{o-}↑	MAE↓	MAE_t↓	MAE_o↓	MAE_{o+}↓	MAE_{o-}↓	mIoU↑	Recall↑	CE↓	SE↓
15	21.37	62.51	0.80	1.40	0.20	23.99	2.33	37.08	5.64	48.53	18.08	25.51	**93.27**	**88.34**
20	21.21	61.65	**0.99**	**1.74**	**0.24**	18.59	2.15	28.5	4.91	37.10	17.56	26.35	93.64	88.87
25	**21.49**	**62.73**	0.86	1.48	**0.24**	19.69	2.19	30.28	4.82	39.55	17.84	26.64	93.78	88.69
30	20.34	59.34	0.85	1.46	0.23	**16.31**	2.19	**24.94**	4.51	**32.38**	**18.09**	**26.85**	94.48	88.80
35	20.66	60.32	0.83	1.47	0.20	17.29	**2.07**	26.53	**4.33**	34.61	17.83	25.92	94.52	89.05
40	21.38	62.35	0.89	1.56	0.23	18.79	2.17	28.89	4.92	37.61	17.75	25.78	94.86	89.25

Table 5. Complete result of analysis on top-μ at scale $\mathcal{S} = (2)$.

Scale	AP↑	AP_t↑	AP_o↑	AP_{o+}↑	AP_{o-}↑	MAE↓	MAE_t↓	MAE_o↓	MAE_{o+}↓	MAE_{o-}↓	mIoU↑	Recall↑	CE↓	SE↓
(1)	21.02	61.28	**0.89**	**1.59**	0.19	20.31	2.25	31.32	5.14	40.85	**18.34**	25.25	96.36	89.57
(2)	19.06	55.74	0.72	1.27	0.17	21.28	2.11	32.81	5.32	42.82	17.60	23.67	96.62	90.34
(4)	19.93	58.62	0.58	1.03	0.12	21.87	2.22	33.77	5.15	44.19	17.41	23.09	96.60	90.41
$(4, 2)$	19.20	56.27	0.67	1.11	0.22	**13.64**	2.27	**20.68**	4.82	39.55	16.88	25.34	98.66	94.06
$(2, 1)$	**21.49**	**62.73**	0.86	1.48	**0.24**	19.69	2.19	30.28	4.82	39.55	17.84	26.64	**93.78**	**88.69**
$(4, 2, 1)$	19.96	58.30	0.79	1.42	0.16	**14.94**	**2.05**	22.86	**4.63**	**29.49**	18.26	23.91	98.25	93.06

Table 6. Complete result of analysis on multiple-scale.

Num of Samples	AP↑	AP_t↑	AP_o↑	AP_{o+}↑	AP_{o-}↑	MAE↓	MAE_t↓	MAE_o↓	MAE_{o+}↓	MAE_{o-}↓	mIoU↑	Recall↑	CE↓	SE↓
1	**9.07**	**24.01**	1.59	2.01	**1.17**	24.03	3.25	36.25	6.75	46.99	18.41	24.91	**93.68**	92.56
2	7.75	20.13	1.56	2.23	0.89	14.16	2.50	20.95	4.57	26.92	19.98	**25.27**	95.43	91.06
5	5.89	15.02	1.32	1.86	0.78	**13.75**	2.17	**20.45**	3.71	**26.54**	20.47	25.15	96.18	90.39
8	5.51	12.64	**1.94**	2.01	**1.17**	14.16	2.08	21.11	3.58	27.50	**20.53**	25.07	96.35	90.24
10	4.93	11.54	1.62	**2.47**	0.76	14.69	2.07	21.94	**3.48**	28.67	**20.53**	25.03	96.41	**90.18**

Table 7. Analysis of the number of outpainting samples.

Figure 7. Outpainted example from SDXL [17] + BLIP2 [12]. These three examples show that the outpainted example can be bottlenecked by any one component, and the randomness of the outpainted result. The middle example demonstrates that when both components collaborate well, the left and right example shows the bottleneck made by either VLM or the outpainting model.

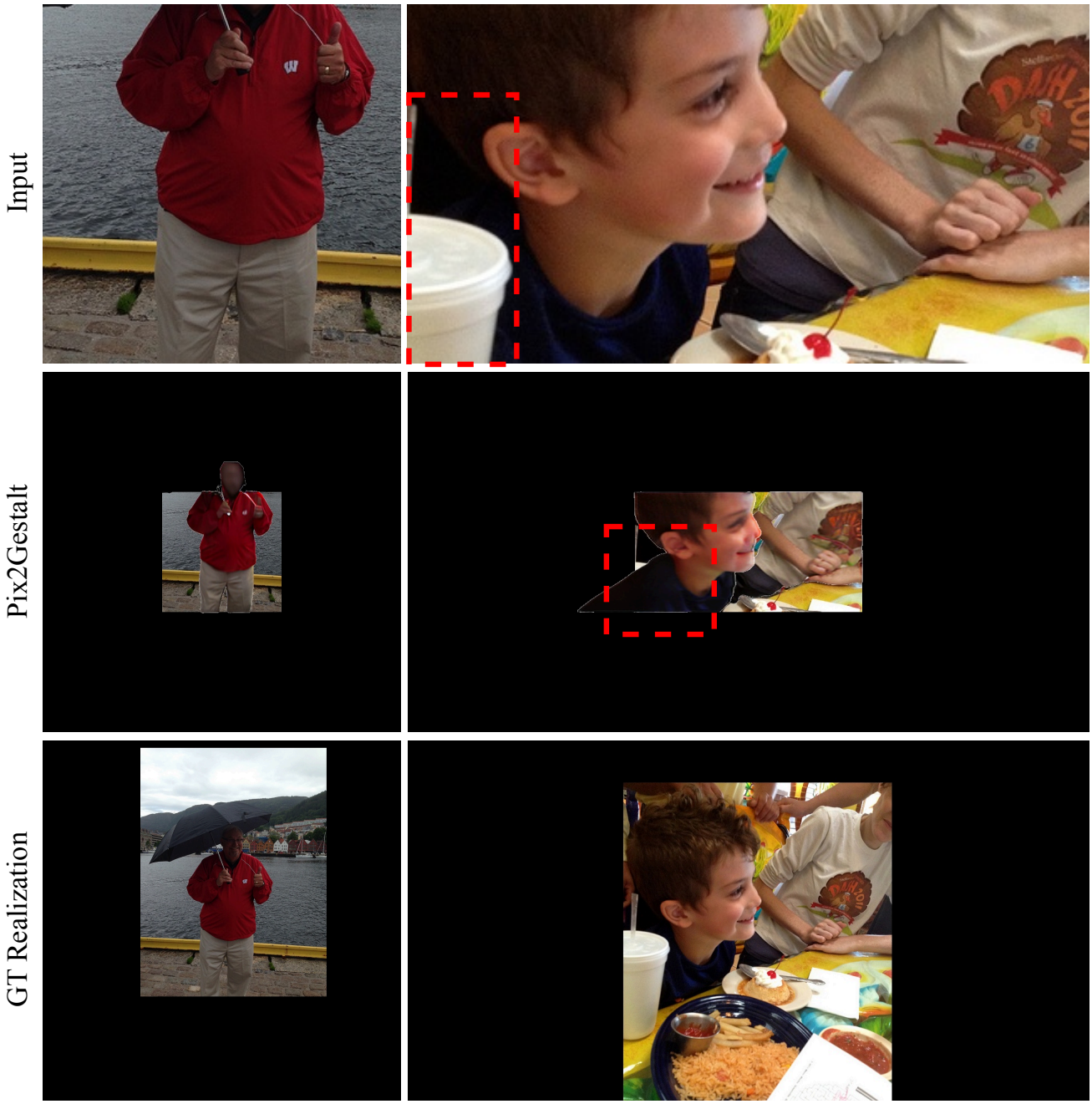

Figure 8. Completion examples with Pix2Gestalt [15]. The first example shows that the model struggles to complete out-of-frame regions despite strong visual evidence, while the second demonstrates effective in-frame occluder removal. Together, these cases highlight the distinction between in-frame completion and out-of-frame completion: strong performance on the former does not necessarily transfer to the latter.