PHAC: Promptable Human Amodal Completion

Supplementary Material

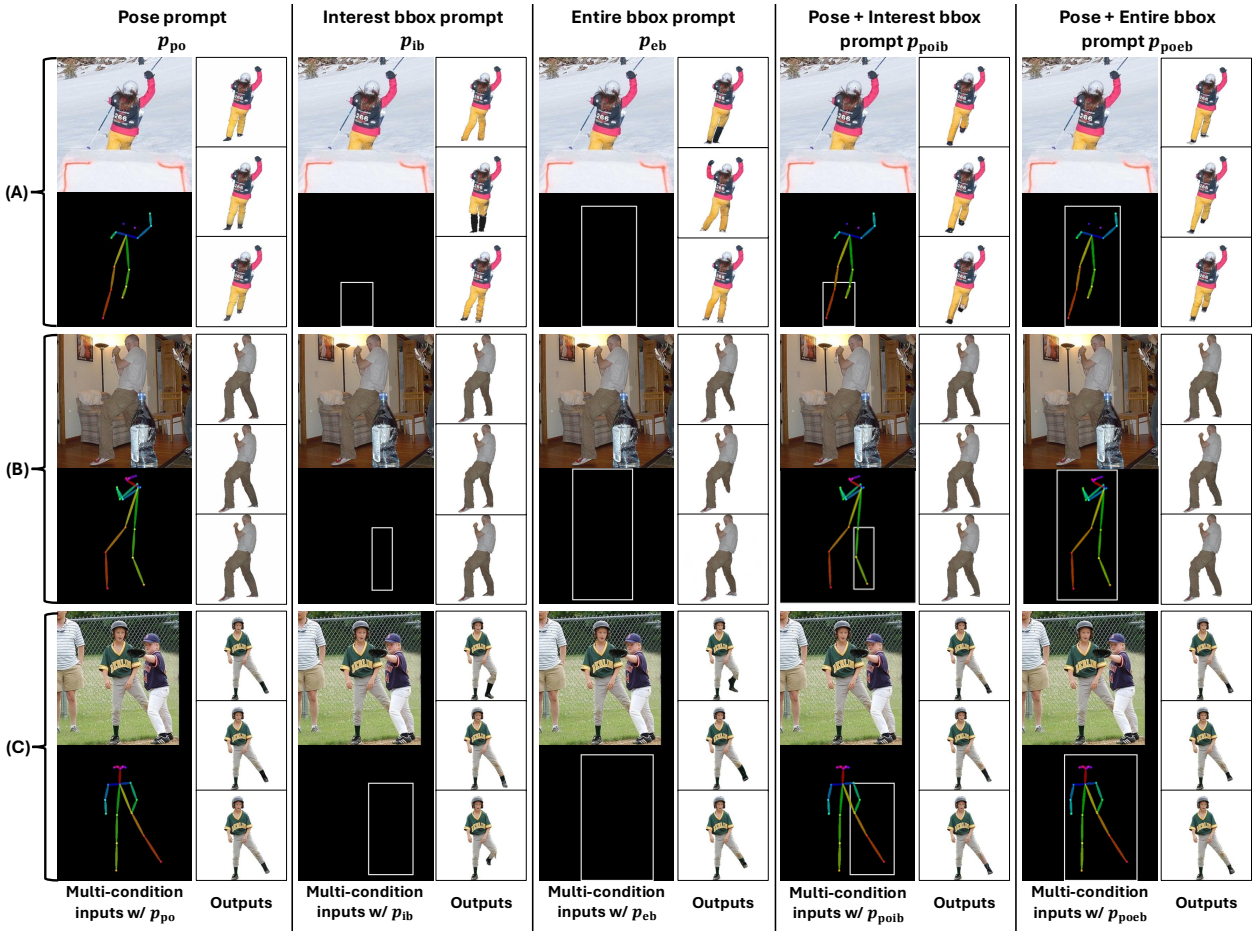

Figure S1. **Qualitative results with different user prompts.** Examples (A, B, C) are from the AHP real test dataset. For each prompt type, three samples are generated using the same random seeds for a fair comparison. In (A), the three samples in the p_{po} column are generated using seeds 42, 616, and 2026, and the same set of seeds is reused for the remaining prompt types (p_{ib}, p_{eb}, p_{poib}, p_{poeb}), ensuring that images within the same row are generated with identical seeds.

S1. Additional Analysis of User Prompts

Table 2 in the main paper reports the user prompt ablation results on OccThuman2.0. In this section, we extend the evaluation by presenting additional quantitative and qualitative results on the AHP real test dataset, including the KID metric. Based on these results, we analyze each prompt in terms of performance and user effort by reporting the marginal gain in joint error per additional input point, enabling a normalized comparison across prompts.

S1.1. Additional Qualitative Analysis

Qualitative results for the same input image under different user prompts are shown in Fig. S1. While the pose prompt p_{po} provides joint-level pose constraints, it does not directly constrain the synthesis region. As shown in the p_{po} column of Fig. S1(A), the occluded right leg generally follows the 2D pose map, but its reconstructed length varies across samples.

The interest-region bbox prompt p_{ib} provides a direct constraint on the synthesis region but offers less control over pose compared with the pose prompt p_{po}. In the p_{ib} column of Fig. S1(A, C), the occluded leg is synthesized within the bbox, and the resulting poses show diverse variations within this constraint. The entire-region bbox prompt p_{eb} provides an even weaker constraint on the occluded regions than the interest-region bbox p_{ib}. Consequently, the

User Prompt	OccThuman2.0 test dataset (synthetic)						AHP test dataset (real)					
	LPIPS* ↓	SSIM ↑	KID* ↓	MSE* ↓	PSNR ↑	Joint Err. ↓	LPIPS* ↓	SSIM ↑	KID* ↓	MSE* ↓	PSNR ↑	Joint Err. ↓
Pose (p_{po})	49.47	**0.948**	6.12	**4.37**	**25.86**	23.33	38.77	0.970	**1.25**	**3.41**	**26.93**	6.37
Interest bbox (p_{ib})	51.83	0.942	6.89	5.69	24.99	24.01	41.03	0.964	1.44	4.87	25.99	11.18
Entire bbox (p_{eb})	52.28	0.941	6.32	5.66	25.07	28.23	41.35	0.963	1.37	5.04	25.78	11.84
Pose & Interest bbox (p_{poib})	**49.35**	0.947	**6.03**	4.56	25.69	22.15	38.67	0.970	1.31	3.42	26.60	6.21
Pose & Entire bbox (p_{poeb})	49.42	0.946	6.12	4.89	25.49	**21.96**	**38.51**	**0.971**	1.33	3.48	26.74	**6.15**

Table S1. **Quantitative comparison across different user prompts.**

	p_{po}	p_{ib}	p_{eb}	p_{poib}	p_{poeb}
Δ JE	6.43	5.75	1.53	7.61	**7.80**
Δ JE pp	1.07	**2.86**	0.77	0.95	0.98

Table S2. **Quantitative comparison of prompt efficiency.** We report the absolute gain in joint error (ΔJE) and the per-point gain (ΔJE pp) for different user prompts. While pose-based prompts yield larger absolute gains, the interest-region bbox prompt p_{ib} achieves the largest per-point gain, indicating efficient improvements with minimal user input.

model sometimes generates a left arm within the entire-region bbox (the p_{eb} column of Fig. S1(A)), and when the input already satisfies this bbox constraint, it may fail to reconstruct the occluded leg (the p_{eb} column of Fig. S1(B)). Similar to the interest-region bbox p_{ib}, the poses remain variable within the bbox, as seen in the p_{eb} column of Fig. S1(C).

This issue can be mitigated by providing both a bbox prompt and a pose prompt as the user prompt. Regardless of bbox type, combining a pose with either an interest-region or entire-region bbox $(p_{\text{poib}}, p_{\text{poeb}})$ yields images that are consistent across samples and satisfy both the pose and bbox constraints simultaneously.

S1.2. Additional Quantitative Results

For a more detailed analysis, we evaluate our method on both the OccThuman2.0 and AHP test datasets and report the results for all metrics described in Sec. 4.3 of the main paper in Table S1.

Across both datasets, bbox prompts $(p_{\text{ib}}, p_{\text{eb}})$ yield higher joint error than the pose prompt p_{po}. This pattern is consistent with the qualitative examples in Fig. S1, where bbox prompting allows diverse poses within the specified region. In addition, the entire-region bbox prompt p_{eb} generally performs worse than the interest-region bbox prompt p_{ib}, as shown in Fig. S1, where the less specific entire-region constraint can lead to incomplete reconstructions of occluded parts or spurious content within the bbox.

Combining pose and bbox prompts $(p_{\text{poib}}, p_{\text{poeb}})$ yields overall performance comparable to the pose-only setting. In addition, it provides a small but consistent reduction in joint error across both datasets, regardless of bbox type.

S1.3. Prompt Efficiency under User Effort

Different prompts provide distinct levels of structural guidance and require different amounts of user input. The pose prompt p_{po} provides joint-level structural cues, but under severe occlusion it may require many input points to specify a reliable pose. In contrast, bbox prompts $(p_{\text{ib}}, p_{\text{eb}})$ require only two input points, providing a lightweight way to localize the target synthesis region.

To quantify prompt efficiency with respect to user effort, we report both the absolute gain in joint error (ΔJE) and the gain per input point (ΔJE pp) in Table S2. Pose-based prompts $(p_{\text{po}}, p_{\text{poib}}, p_{\text{poeb}})$ yield larger absolute gains, reflecting the benefit of stronger structural constraints. The interest-region bbox prompt p_{ib} achieves the highest per-point gain, indicating efficient improvements with minimal user input. In contrast, the entire-region bbox prompt p_{eb} provides the weakest constraint, as discussed in Secs. S1.1 and S1.2. This encourages diverse plausible generations within the bbox; consequently, p_{eb} yields the lowest gains in both ΔJE and ΔJE pp when evaluated against the ground-truth pose. Finally, combining pose and bbox prompts $(p_{\text{poib}}, p_{\text{poeb}})$ results in higher absolute gains than single-prompt settings, but lower per-point gains than p_{po} and p_{ib}, reflecting the increased user effort required by the combined prompts.

Overall, supporting multiple prompt types enables flexible trade-offs between performance and user effort, allowing users to select prompts that best match the difficulty of occlusion and the desired level of interaction.

S2. Detailed Implementations

S2.1. OccThuman2.0 Dataset

As discussed in Sec. 4.1 of the main paper, we construct the OccThuman2.0 dataset from THuman2.0 [16] for training and evaluation. Fig. S2 illustrates the dataset construction pipeline for 10 views of a single subject. For each subject, we first render 10 views from the corresponding 3D mesh, following prior work on clothed human reconstruction [5, 6, 14]. From the rendered images, we estimate 2D human poses using OpenPose [3]. We then apply the same augmentations, including horizontal flip, vertical flip, rotation, and scaling, to both the images and the poses.

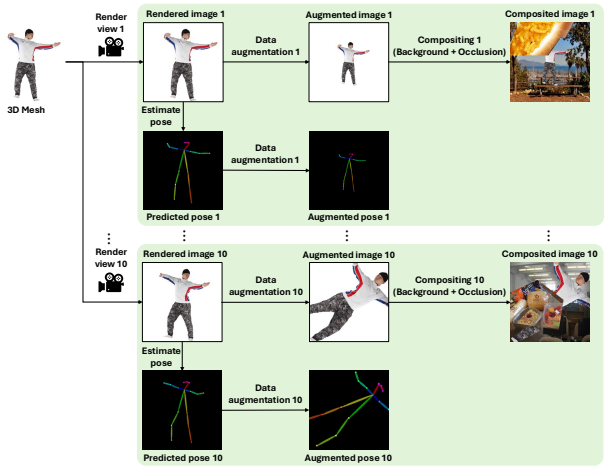

Figure S2. **Construction pipeline of the OccThuman2.0 dataset.** For each THuman2.0 mesh, we render 10 views and apply per-view data augmentations and background-occlusion compositing to generate 10 composited images. Repeating this process across all 526 meshes produces a total of 5,260 composited images.

Each augmentation is applied independently with a probability of 30%. When rotation is applied, we sample a single angle θ' from a uniform distribution $\theta' \sim \mathcal{U}[0°, 360°)$. When scaling is applied, we sample a single scale factor $s' \sim \mathcal{U}[0.5, 1.5]$ to prevent the person from becoming excessively small or large. At this stage, the rendered images cover diverse person orientations and scales but contain no backgrounds or occlusions.

We composite backgrounds using images randomly sampled from Places2 [19] and synthesize occlusions by overlaying randomly sampled MSCOCO [7] object instances onto the images, following previous work on amodal completion [10, 11, 17]. Since the last four subjects in THuman2.0 (0522-0525) are scans of the same person, we use the corresponding 40 images (4 subjects $\times$ 10 views) for testing and the remaining 5,220 images corresponding to 522 subjects (0000-0521) for training. For each subject, we sample a target occlusion ratio from a Gaussian distribution and ensure that the final set of ratios follows this target distribution. We use the GT complete images (before background and occlusion compositing) to train the coarse image generation diffusion model ϵ_{cig} and the ControlNet Φ_{CN}, and the GT invisible masks to train the invisible mask prediction U-Net $\mathcal{U}_{\text{iv}}$.

S2.2. Invisible Mask Prediction U-Net

To train the invisible mask prediction U-Net $\mathcal{U}_{\text{iv}}$, we first train the coarse image generation DM and the ControlNet. Following Sec. 4.2 of the main paper, we generate 16 coarse complete images I_{cc} for each training input image I_{ic}. During the training of $\mathcal{U}_{\text{iv}}$, we randomly select one of these 16 coarse complete images as input. We use the AdamW optimizer with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. A single invisible mask prediction step takes approximately 0.02 seconds per image (about 50 fps) on a single RTX 3090 GPU. During training, each term of the loss in Eq. (13) of the main paper is computed as follows:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[p_i^* \log p_i + (1-p_i^*)\log(1-p_i)\right], \quad \text{(S1)}$$

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} p_i p_i^* + \delta}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} p_i^* + \delta}, \quad \text{(S2)}$$

where $p_i = M_{\text{iv}}(i)$ and $p_i^* = M_{\text{iv}}^*(i)$ denote the predicted and GT invisible mask values at pixel i, respectively, and N denotes the number of pixels. $\delta = 1 \times 10^{-6}$ is a small positive constant added for numerical stability to prevent division by zero.

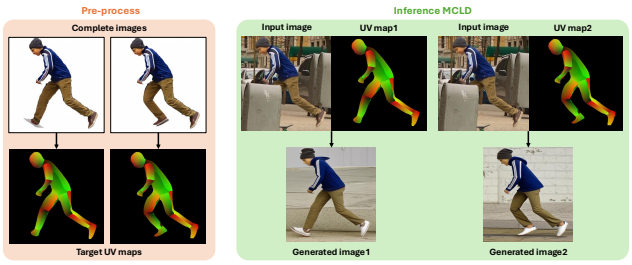

Figure S3. **Prompt image for MCLD.** MCLD uses a DensePose-based UV map as its conditioning input instead of a 2D pose map. We obtain the target UV map from the occlusion-free complete image and perform MCLD inference conditioned on this UV map.

S2.3. Refinement Network

We use a pre-trained Stable Diffusion XL (SDXL) model [12] as the refinement network Φ_{RF}. Since SDXL is optimized for a resolution of 1024×1024, we resize the baseline composite image I_{base} to 1024×1024 before feeding it into Φ_{RF}. For evaluation, all generated images are resized to a fixed resolution to ensure a fair comparison across methods. We fix the number of refinement steps to 20 and set the noise strength to $s = 0.5$, based on the ablation study in Table 3 of the main paper. With these settings, denoising begins at timestep $t_0 = 380$. We set the classifier-free guidance scale to 1.5 and use the `diffusers/stable-diffusion-xl-1.0-inpainting-0.1` checkpoint from the Diffusers library [13]. The refinement step takes approximately 4 seconds per image (about 0.25 fps) on a single RTX 3090 GPU.

S2.4. Prompt Image

As described in Sec. 3.3 of the main paper, our approach takes as input simple, user-friendly prompts P consisting of

Method	OccThuman2.0		AHP	
	mIoU ↑	L1* ↓	mIoU ↑	L1* ↓
pix2gestalt [11]	48.26	39.74	48.97	22.04
SDHDO [10]	51.86	33.58	51.47	21.35
Ours	87.84	6.04	77.87	7.89

Table S3. **Quantitative results for invisible mask prediction.** Previous amodal completion methods [10, 11] do not use the coarse complete image I_{cc} as input, so their results are not directly comparable.

only a few points. For the pose prompt p_{po}, the input consists of the indices and 2D coordinates of additional joints beyond the visible ones. Given these user-specified joints, we convert the 2D pose from the OpenPose keypoint format to the COCO format [7] and, following ControlNet [18], render it as a 2D pose map with a fixed color mapping and parent-child relations. This pose map is used as the prompt image I_p.

For the bbox prompts p_{ib} and p_{eb}, the input consists of two points corresponding to the top-left and bottom-right corners. We construct the corresponding axis-aligned bbox and visualize its boundary with a fixed thickness, setting boundary pixels to 255 (white) and all other pixels to 0 (black). When the thickness is set to 1, the resulting bbox image has a one-pixel boundary. In practice, we fix the thickness to 4 for both training and inference. To improve robustness to noisy user-provided bbox prompts, we augment the bbox during training by randomly scaling it and jittering its corner coordinates with 50% probability. When using both pose and bbox prompts (p_{poib}, p_{poeb}), we concatenate their corresponding prompt images along the channel dimension to form the final prompt image I_p.

As noted in Sec. S1.1, because most previous methods accept a 2D pose map as input, we use only the prompt image derived from the pose prompt p_{po} for comparison. Because MCLD [8] is conditioned on a UV map [4], we use the UV map extracted from the occlusion-free complete image as the prompt image I_p. Fig. S3 shows the UV map prompt used to generate the MCLD results in Fig. 1 of the main paper. The UV map provides not only pose information but also regional coverage, supplying cues that help the model generate a plausible human shape. Such regional information is not available in a 2D pose map.

S3. Additional Results

S3.1. Invisible Mask

In this section, we present quantitative and qualitative results for the invisible mask prediction U-Net $\mathcal{U}_{iv}$. To evaluate performance, we report invisible mask prediction results on both OccThuman2.0 and the AHP real test dataset, measuring performance using mean Intersection over Union

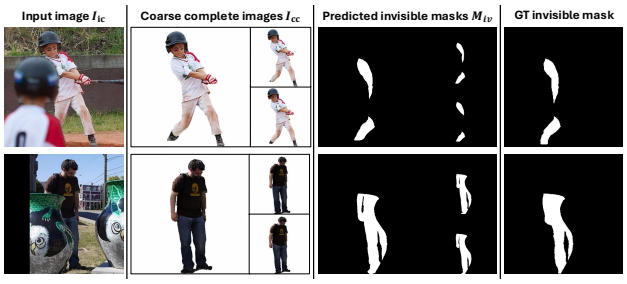

Figure S4. **Qualitative results for invisible mask prediction.** Given the coarse complete image I_{cc}, $\mathcal{U}_{iv}$ predicts invisible masks M_{iv} that closely match the ground truth.

(mIoU) and L1 loss. Table S3 compares our method with previous amodal completion methods [10, 11]. These methods do not take the coarse complete image I_{cc} as input; therefore, their results are not directly comparable and are provided for reference only. Predicting the invisible mask M_{iv} from the incomplete image I_{ic} and the visible mask M_v alone is under-constrained, whereas coarse completion provides I_{cc} as a hypothesis over the invisible regions. Conditioning M_{iv} on I_{cc} therefore encourages consistency between coarse completion and invisible-region localization.

As shown in Fig. S4, when using the coarse complete image I_{cc} as input, $\mathcal{U}_{iv}$ predicts invisible masks that closely match the GT invisible mask. To improve robustness to noisy or structurally imperfect coarse completions, we train $\mathcal{U}_{iv}$ with multiple randomly sampled I_{cc} candidates per GT mask. For a single input image I_{ic}, the stochastic sampling scheme generates multiple distinct coarse complete images I_{cc}, and we predict an invisible mask M_{iv} for each sample. Since the input image is identical across samples, the GT invisible mask is the same for all of them.

S3.2. Refinement Process

Our refinement process addresses two issues: degradation of the visible appearance in the coarse complete image I_{cc} and boundary artifacts introduced by naïve compositing to form the base composite image I_{base}. The degradation in the visible region can be partially mitigated by compositing the coarse complete image I_{cc} with the input image I_{ic} using the visible region mask M_v to obtain the base composite image I_{base}, as in Eq. (11) of the main paper. However, the base composite image I_{base} still suffers from boundary artifacts and fails to blend the appearance smoothly across regions. To address these issues, we adopt an inpainting-based refinement process.

As shown in the blue bboxes in Fig. S5, the coarse complete image I_{cc} exhibits degradation in fine-grained facial details. In contrast, the refined complete image I_{rc} produced by the refinement network Φ_{RF} closely matches the input in the facial region and preserves fine-grained details. As shown in the red bboxes in Fig. S5, the base composite im-

Method	User Prompt	OccThuman2.0 test dataset (synthetic)						AHP test dataset (real)					
		LPIPS* ↓	SSIM ↑	KID* ↓	MSE* ↓	PSNR ↑	Joint Err. ↓	LPIPS* ↓	SSIM ↑	KID* ↓	MSE* ↓	PSNR ↑	Joint Err. ↓
pix2gestalt [11]	-	90.11	0.911	16.51	7.58	22.63	36.65	75.73	0.942	5.98	5.22	24.06	10.96
pix2gestalt [11]†	2D pose map	88.58	0.914	16.75	6.93	22.94	31.37	75.25	0.943	6.35	4.87	24.25	10.42
SDHDO [10]	2D pose map	81.39	0.924	16.41	7.05	23.80	43.49	64.19	0.956	6.05	6.05	24.45	9.24
Ours	-	53.19	0.939	6.23	5.69	25.03	29.76	41.99	0.964	1.60	4.79	25.86	13.23
Ours	2D pose map	**49.47**	**0.948**	**6.12**	**4.37**	**25.86**	**23.33**	**38.77**	**0.970**	**1.25**	**3.41**	**26.93**	**6.37**

Table S4. **Quantitative comparison for PHAC.** This table largely mirrors Table 1 of the main paper, but additionally includes the results of our method evaluated without a user prompt. To avoid redundancy, PGPIS results are omitted. † indicates the use of the same user prompt injected through a pre-trained ControlNet.

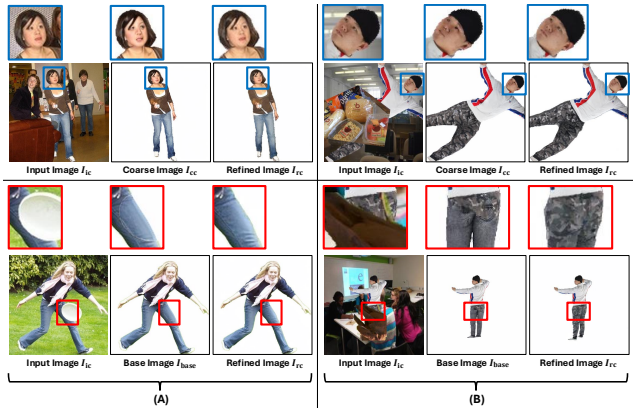

Figure S5. **Qualitative results for the refinement process.** (A) Results on the AHP dataset and (B) results on the OccThuman2.0 dataset. The blue bboxes compare the facial regions of the input image I_{ic}, the coarse complete image I_{cc}, and the refined image I_{rc} to assess preservation of visible appearance. The red bboxes compare the boundary regions of the input image I_{ic}, the base composite image I_{base}, and the refined image I_{rc} to verify smooth blending without boundary artifacts.

Method	OccThuman2.0				AHP			
	LPIPS* ↓	SSIM ↑	MSE* ↓	PSNR ↑	LPIPS* ↓	SSIM ↑	MSE* ↓	PSNR ↑
PIDM [2]	88.43	0.862	16.16	19.31	120.85	0.835	20.40	17.51
MCLD [8]	76.77	0.894	11.48	20.73	101.23	0.880	11.82	20.05
pix2gestalt [11]	48.49	0.976	1.37	29.05	52.81	0.971	1.55	28.55
SDHDO [10]	38.08	0.996	0.32	35.23	37.71	0.994	0.32	35.27
Ours	**13.72**	**0.998**	**0.17**	**37.93**	**15.65**	**0.998**	**0.14**	**38.85**

Table S5. **Quantitative comparison for PHAC in the visible region.** We compute LPIPS, SSIM, MSE, and PSNR on the visible (non-occluded) regions of the input image to measure fidelity to the input.

Method	OccThuman2.0				AHP			
	LPIPS* ↓	SSIM ↑	MSE* ↓	PSNR ↑	LPIPS* ↓	SSIM ↑	MSE* ↓	PSNR ↑
PIDM [2]	51.99	0.908	11.50	21.37	27.44	0.966	4.92	26.03
MCLD [8]	47.67	0.927	7.65	23.25	23.94	**0.974**	3.34	27.24
pix2gestalt [11]	45.79	0.932	6.36	24.32	23.96	0.972	3.67	27.43
SDHDO [10]	47.14	0.926	6.90	24.22	27.57	0.960	5.73	25.39
Ours	**36.85**	**0.950**	**4.29**	**26.15**	**23.05**	0.973	**3.28**	**27.60**

Table S6. **Quantitative comparison for PHAC in the invisible region.** Using the same metrics as Table S5, we compute them on the invisible (occluded) regions to measure reconstruction fidelity, conditioned on the visible appearance of the input image.

age I_{base} produces boundary artifacts and does not blend the appearance smoothly across regions. In contrast, the refined image I_{rc} mitigates boundary artifacts and achieves seamless blending across regions. The refinement network Φ_{RF} takes the base composite image I_{base} as input, rather than the coarse complete image I_{cc}.

S3.3. Without Prompt Conditioning

Existing amodal completion [9, 11, 15] and HAC [10, 20] approaches are not designed to take user prompts as input. Consequently, although the proposed method accepts simple, user-friendly prompts, it is restrictive to assume that they are always available. To reduce prompt dependency, we apply prompt dropout during training, where the prompt image I_p is replaced with an all-zero image with a probability of 5%. With this training strategy, our method can perform HAC without a user prompt.

As shown in Table S4, performance without a user prompt is slightly worse across all metrics than in the 2D pose-prompted setting, with the largest degradation observed in joint error. Nevertheless, our approach achieves the best performance on most metrics compared with previous methods that use a 2D pose map as the user prompt. As Table S4 is nearly identical to Table 1 of the main paper, we do not repeat the PGPIS results in this table. The detailed results for the PGPIS methods are reported in Table 1 of the main paper.

S3.4. Region-wise Quantitative Results

Table 1 of the main paper reports quantitative results over the entire image, covering both visible and invisible regions. We additionally report region-wise quantitative results for the visible and invisible regions in Tables S5 and S6, respectively. Since we evaluate the visible and invisible regions separately, whole-image metrics such as KID and joint error are not applicable. We therefore report only reconstruction metrics (MSE, PSNR) and perceptual quality metrics (LPIPS, SSIM).

As shown in Table S5, PIDM and MCLD tend to overfit the training set, failing to preserve the visible appearance and resulting in weak performance in the visible re-

Figure S6. **Additional qualitative comparison for PHAC.** (A, B): AHP test examples, where previous methods either hallucinate appearance, lose fine-grained details, or fail to complete occluded regions. In contrast, our method preserves the visible appearance and aligns with the pose condition. (C, D): OccThuman2.0 test examples under severe occlusions, where previous methods often produce implausible or pose-misaligned results. In contrast, our method synthesizes occluded regions that remain consistent with both the visible appearance and the pose condition.

gion across datasets. Amodal completion methods better preserve the visible regions than PGPIS methods, with SDHDO substantially outperforming pix2gestalt. With the refinement network, our method further improves performance and outperforms all previous methods on all metrics.

As shown in Table S6, the gap between PGPIS and amodal completion methods is comparatively modest in the invisible region, in contrast to the larger gap observed in Table S5. On OccThuman2.0, we obtain the best overall performance, with a clear margin over the top three competing methods [8, 10, 11], while on AHP our approach leads on three of the four metrics.

S3.5. Additional Qualitative Results

In this section, we present additional qualitative results for PHAC in Figs. S6–S8. Fig. S6 shows results on the AHP test dataset (A, B) and the OccThuman2.0 test dataset (C, D). Fig. S7 presents in-the-wild results on MSCOCO [7] and MPII [1], and Fig. S8 provides additional examples under severe occlusions. As described in Sec. S2.4, we construct the prompt images from the multi-condition inputs in Fig. S6, and all baselines are conditioned on the same 2D pose map except that MCLD uses a DensePose UV map and pix2gestalt does not use an explicit conditioning input.

As shown in Fig. S6(A, B), PIDM and MCLD hallucinate appearance rather than preserving the visible appearance in the input image. In contrast, pix2gestalt often fails to complete occluded regions (Fig. S6(A)) and generates images that lack fine-grained detail, particularly in the facial region (Fig. S6(B)). SDHDO preserves the visible appearance more faithfully, but produces blurry images and may still fail to reconstruct occluded regions, similar to pix2gestalt. Overall, our method consistently produces images that align with the pose condition while preserving the visible appearance.

Fig. S6(C, D) presents OccThuman2.0 results under

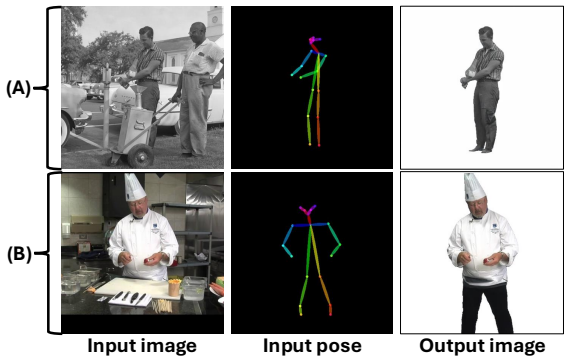

Figure S7. **In-the-wild qualitative results.** We show in-the-wild results of our method on a grayscale MSCOCO example (A) and an unseen-identity MPII example (B).

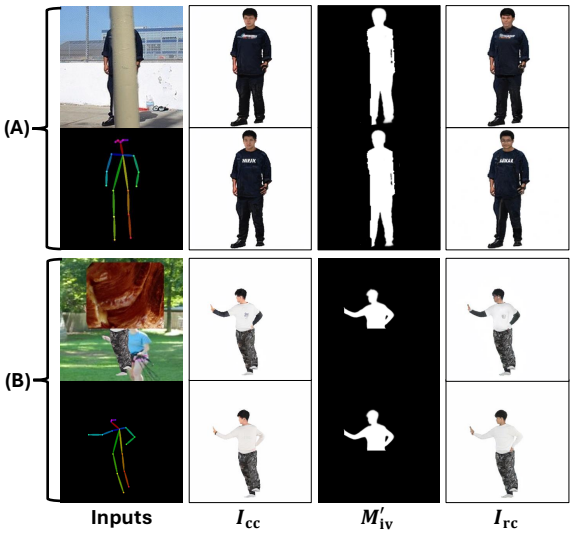

Figure S8. **Severe-occlusion qualitative results.** We show intermediate outputs (I_{cc}, M'_{iv}) and the final output (I_{rc}) of our method under severe occlusions. (A): AHP real example (76% occlusion). (B): OccThuman2.0 example (53% occlusion). Each row uses a different random seed.

more severe occlusions than (A, B). Under heavy occlusions, PIDM often fails to produce a plausible person image. Due to errors in UV map estimation, MCLD may transfer occluder appearance into the output and produce implausible results. SDHDO and pix2gestalt yield either implausible (Fig. S6(C)) or pose-misaligned outputs (Fig. S6(D)). In contrast, our method reconstructs occluded regions consistent with the visible appearance and closely aligns the synthesized content with the pose condition, producing plausible person images even under severe occlusions. For example, in Fig. S6(D), other methods fail to reconstruct the right hand correctly, whereas our method synthesizes it consistently with the pose condition.

Fig. S7 reports in-the-wild results of our method, where (A) shows a grayscale MSCOCO example and (B) shows an unseen-identity MPII example. These results demonstrate that our method can perform PHAC on in-the-wild data while preserving the visible appearance of the input image and remaining consistent with the input pose.

	LPIPS* ↓	KID* ↓	MSE* ↓	PSNR ↑	Joint Err. ↓
w/o dil	49.86	6.54	4.59	25.63	23.73
w/ dil	**49.47**	**6.12**	**4.37**	**25.86**	**23.33**

Table S7. **Quantitative comparison of our method with and without dilation.** Invisible mask dilation is used as a conservative margin for boundary uncertainty and has a marginal impact on metrics.

We additionally provide qualitative results under severe occlusions in Fig. S8. For each example, we visualize the multi-condition inputs as well as the intermediate and final outputs of our method: the coarse complete image I_{cc}, the dilated invisible mask M'_{iv}, and the refined complete image I_{rc}. Fig. S8(A) shows an AHP real example where approximately 76% of the human body is occluded, and Fig. S8(B) shows an OccThuman2.0 example with approximately 53% occlusion. Each row corresponds to a different random seed during inference. Even in these severe-occlusion cases, our method robustly aligns with the input pose and produces plausible I_{cc}, M'_{iv}, and I_{rc}. Due to the limited visible-appearance cues under heavy occlusion, varying the seed can lead to different texture realizations in the invisible regions, while maintaining pose consistency. For reference, the OccThuman2.0 examples in Fig. S6(C, D) correspond to approximately 58% and 43% occlusion, respectively.

S3.6. Effect of Invisible Mask Dilation

As discussed in Sec. 3.4, we use invisible mask dilation only as a conservative margin to account for boundary uncertainty in the predicted invisible region. Table S7 reports an ablation comparing our method with and without dilation on the OccThuman2.0 test dataset. Dilation yields consistent but marginal improvements across metrics, indicating that it has limited impact on overall performance while providing a safer margin around uncertain occlusion boundaries.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3686–3693, 2014. 6

[2] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5968–5976, 2023. 5

[3] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estima-

tion using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2

[4] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7297–7306, 2018. 4

[5] Hsuan-I Ho, Jie Song, and Otmar Hilliges. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 538–549, 2024. 2

[6] Jin Gyu Hong, Seung Young Noh, Hee Kyung Lee, Won Sik Cheong, and Ju Yong Chang. 3d clothed human reconstruction from sparse multi-view images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 677–687, 2024. 2

[7] 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 *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 3, 4, 6

[8] Jiaqi Liu, Jichao Zhang, Paolo Rota, and Nicu Sebe. Multi-focal conditioned latent diffusion for person image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16019–16028, 2025. 4, 5, 6

[9] Zhengzhe Liu, Qing Liu, Chirui Chang, Jianming Zhang, Daniil Pakhomov, Haitian Zheng, Zhe Lin, Daniel Cohen-Or, and Chi-Wing Fu. Object-level scene deocclusion. In *ACM SIGGRAPH 2024 Conference*, pages 1–11, 2024. 5

[10] Seung Young Noh and Ju Yong Chang. Stable diffusion-based approach for human de-occlusion. In *ACM International Conference on Multimedia*, pages 10044–10053, 2025. 3, 4, 5, 6

[11] Ege Ozguroglu, Ruoshi Liu, Dídac Surís, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. pix2gestalt: Amodal segmentation by synthesizing wholes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3931–3940. IEEE Computer Society, 2024. 3, 4, 5, 6

[12] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3

[13] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 3

[14] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 2

[15] Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Amodal completion via progressive mixed context diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9099–9109, 2024. 5

[16] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5746–5756, 2021. 2

[17] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene de-occlusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3784–3792, 2020. 3

[18] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 4

[19] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017. 3

[20] Qiang Zhou, Shiyin Wang, Yitong Wang, Zilong Huang, and Xinggang Wang. Human de-occlusion: Invisible perception and recovery for humans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3691–3701, 2021. 5