ANATOGUARD: ANATOMY-GUIDED POSE–SEMANTIC DIFFUSION FOR HUMAN AMODAL COMPLETION

Anonymous authors
Paper under double-blind review

ABSTRACT

Human amodal instance segmentation aims to recover the complete extent of a person, including body regions hidden by occlusion. This task is challenging because invisible regions cannot be directly observed. We observe that the anatomical structure of the human body allows a detected full-body pose to serve as an explicit spatial prior for locating occluded body parts. However, pose alone provides limited information about appearance, scene context, and the continuation of hidden body regions, and its reliability may decrease under severe occlusion.

We introduce ANATOGUARD, an anatomy-guided pose-semantic diffusion framework for human amodal completion. The detected pose is rendered as a spatial condition and jointly provided with the occluded image and visible-person mask. To complement unreliable pose predictions, a Qwen-based module extracts typed semantic attributes related to appearance, contextual relations, and hidden-body continuation. These attributes are localized using pose-derived body-part supports and transformed into spatial semantic residuals, providing complementary evidence beyond pose. The proposed framework therefore combines anatomical structure with semantic cues for robust amodal completion. We evaluate pose-only and pose-semantic variants through controlled comparisons and component ablations, with particular emphasis on severe occlusion.

1 INTRODUCTION

Amodal instance segmentation (AIS) predicts the complete spatial extent of an object, including the part hidden from direct view. Compared with traditional instance segmentation Manh Hoang (2026); Kim et al. (2026), AIS enables a more complete geometric and semantic scene understanding. It asks a model to infer a whole from incomplete evidence, while occluded regions have no direct appearance evidence, this ambiguity is especially pronounced for people. Human bodies are articulated, and a plausible completion must connect limbs and remain compatible with the visible posture and scene. Human-centric AIS supports a variety of real-world applications, including autonomous driving Qi et al. (2019), human–computer interaction Chi et al. (2025), augmented and virtual reality systems Taravat et al. (2023), and image editing Zhan et al. (2020).

Recently, diffusion models have emerged as a powerful paradigm for amodal instance segmentation. Beyond achieving competitive segmentation accuracy, diffusion-based frameworks naturally support image content completion, making them attractive for downstream applications such as image editing and amodal reconstruction. Despite these advantages, existing diffusion-based AIS methods remain fragile when applied to human instances under severe occlusion. As shown in Fig. 1, when a person is largely occluded, some body parts are missing in the predicted results of pix2gestalt Ege et al. (2024). Similar errors can also be observed in cases of under-segmentation in modal segmentation. Although several advanced modal segmentation methods have recently been proposed Alexander et al. (2023); Fu et al. (2025), obtaining accurate visible masks remains challenging in practical applications.

This limitation arises from a fundamental reliance on appearance-driven generation. When large portions of the human body are occluded, visual cues alone become insufficient to constrain the diffusion process, leading to ambiguous or inconsistent completions. Some approaches attempt to mitigate this issue by exploiting temporal consistency in videos Chen et al. (2025); Fan et al. (2023); however, their applicability is inherently limited in single-image settings or scenarios where occlu-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

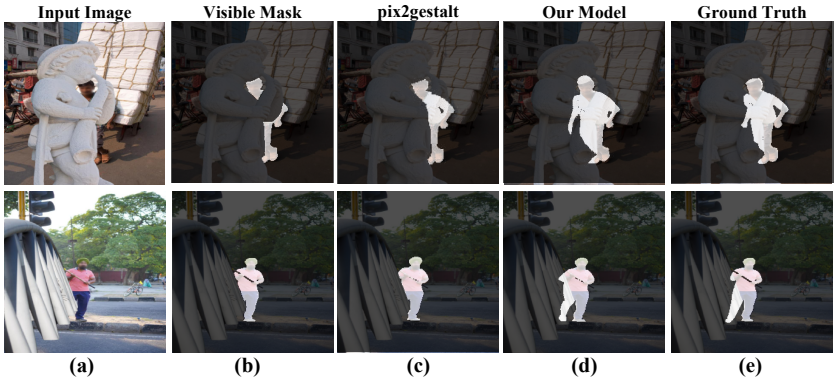


Figure 1: Existing diffusion-based AIS methods remain fragile when applied to human instances under severe occlusion. Our method combines a detected full-body pose, which provides an anatomical spatial prior, with localized semantic cues that compensate for unreliable pose estimates and guide the diffusion process toward anatomically plausible amodal masks. (a) the input image; (b) visible mask; (c) result of pix2gestalt Ege et al. (2024); (d) result of our method; (e) the ground truth.

sions persist across frames. Consequently, recovering structurally coherent human shapes from a single image remains an open challenge. We observe that humans exhibit strong and well-defined anatomical structures that are poorly captured by appearance statistics alone but can be effectively represented by pose configurations. This insight motivates us to introduce full-body human pose as an explicit anatomical prior for diffusion-based amodal instance segmentation. However, pose estimates can become unreliable under severe occlusion, and pose alone provides limited information about appearance, scene context, and the continuation of hidden body regions. Semantic attributes can therefore provide complementary evidence, helping the model assess and compensate for imperfect pose predictions.

Based on this insight, we introduce ANATOGUARD, a pose- and semantic-conditioned diffusion framework for human-centric amodal instance segmentation. A language model fine-tuned on Pix2Human predicts full-body keypoints from occluded images, which are rendered into a spatial pose condition. In parallel, Qwen2.5-VL extracts typed semantic attributes describing appearance, contextual relations, and hidden-body continuation. These attributes are spatially grounded using pose-derived body-part supports and transformed into a localized residual that is injected into the scene latent, enabling semantic cues to refine pose-guided generation.

- We introduce an anatomy-guided diffusion formulation that uses detected full-body pose as a spatial structural prior for human amodal instance segmentation.
- We propose localized pose-semantic residual conditioning, where typed semantic attributes provide complementary evidence about appearance, scene relations, and hidden-body continuation, while mitigating the effect of erroneous poses.
- We conduct controlled ablation experiments to quantify the individual and joint benefits of anatomical pose conditioning and semantic verification, particularly under severe occlusion.

2 RELATED WORK

2.1 AMODAL INSTANCE SEGMENTATION

Amodal instance segmentation (AIS) predicts both visible and occluded regions of an object. Early methods extend Mask R-CNN with additional prediction heads or layered reasoning modules to estimate amodal masks Patrick et al. (2019); Ke et al. (2021); Xiao et al. (2021). To reduce the dependence on costly amodal annotations, PCNet Zhan et al. (2020) uses partially visible masks as self-supervision. Transformer-based methods formulate AIS as query-based prediction, with AISFormer Tran et al. (2022) modeling visible, occluded, and amodal regions using separate queries,

OAFormer Li et al. (2023) incorporating occlusion-rate estimation, and ShapeFormer Tran et al. (2024) improving the transition between visible and amodal masks. More recently, SAMEO En et al. (2025) and SAMBA Liu et al. (2025) adapt SAM as a universal mask decoder for AIS.

Generative models provide another promising solution by jointly completing object appearance and predicting amodal masks Kiana et al. (2018); Ling et al. (2020); Zhan et al. (2024); Xu et al. (2024); Ege et al. (2024). Early approaches such as Amodal-VAE Ling et al. (2020) recover complete masks with variational autoencoders, while later diffusion-based methods exploit pretrained diffusion features or modify the denoising process to improve amodal completion Zhan et al. (2024); Xu et al. (2024); Ege et al. (2024).

Despite their strong object-completion priors, existing diffusion-based AIS methods rely mainly on appearance cues. For articulated humans, these cues are insufficient under severe occlusion, often causing missing limbs or distorted body shapes. Full-body pose provides an explicit anatomical prior for locating occluded parts, but its accuracy may also degrade under heavy occlusion. Our method therefore combines pose-based structural guidance with complementary semantic evidence for robust human amodal segmentation.

2.2 Joint Pose Estimation and Segmentation

Pose estimation and human parsing are closely related and have therefore been jointly modeled in multi-task frameworks Liang et al. (2018); Nie et al. (2018); Zhang et al. (2020b). For visible human instance segmentation, Pose2Instance Tripathi et al. (2017), PersonLab Papandreou et al. (2018), and Pose2Seg Zhang et al. (2019) use keypoints to improve instance separation, especially in crowded scenes. Pose2Seg further introduces affine alignment to normalize each person before mask prediction. However, these methods mainly model visible body regions and do not explicitly recover occluded parts.

Recent studies extend pose-guided segmentation to the amodal setting. Zhou et al. Zhou et al. (2021) use human parsing and pose templates to reconstruct occluded regions, but such fixed priors may lack diversity for complex poses. POISE Dutta et al. (2024) introduces a self-supervised framework that uses pretrained pose estimators and visible segmentation networks to provide geometric and shape supervision. PHAC Noh & Chang (2026) formulates human de-occlusion as controllable diffusion-based generation using ControlNet prompts. However, it relies on user-provided keypoint or region prompts, limiting fully automatic application. Our method instead predicts full-body pose from the occluded image and uses it to guide amodal generation without requiring user interaction.

Existing joint pose-estimation and segmentation methods typically focus on visible-person understanding and do not use full-body pose as an anatomical prior for amodal completion. We develop an automatic pose-to-amodal framework that injects pose-derived spatial guidance into diffusion-based segmentation. To alleviate the effect of inaccurate poses, semantic attributes describing appearance, scene context, and hidden-body continuation are spatially localized and used to refine the generation process.

3 Methodology

3.1 Problem Formulation and Overview

Given an occluded image $\mathbf{X}$ and its visible-person mask $\mathbf{M}_v$, we seek to recover the complete person appearance $\hat{\mathbf{I}}_a$ and obtain the corresponding amodal mask $\hat{\mathbf{M}}_a$. As in prior generative AIS formulations, the mask is extracted from the completed image using a fixed post-processing protocol; the diffusion model itself is trained to reconstruct the complete image. Human completion is under-constrained by appearance when large body regions are hidden. We therefore introduce a predicted full-body pose $\hat{\mathbf{K}}$ as an explicit spatial condition and typed semantic attributes A as localized complementary cues. Our factorization is

$$p(\hat{\mathbf{I}}_a, \hat{\mathbf{K}}, A \mid \mathbf{X}, \mathbf{M}_v) = p(\hat{\mathbf{I}}_a \mid \mathbf{X}, \mathbf{M}_v, \hat{\mathbf{K}}, A)\, p(\hat{\mathbf{K}} \mid \mathbf{X}, \mathbf{M}_v)\, p(A \mid \mathbf{X}, \mathbf{M}_v), \tag{1}$$

where $\hat{\mathbf{K}}$ is the estimated full-body keypoint configuration and A is a set of target-grounded semantic descriptions. The pose model estimates $p(\hat{\mathbf{K}} \mid \mathbf{X}, \mathbf{M}_v)$, and the attribute extractor produces $p(A \mid$

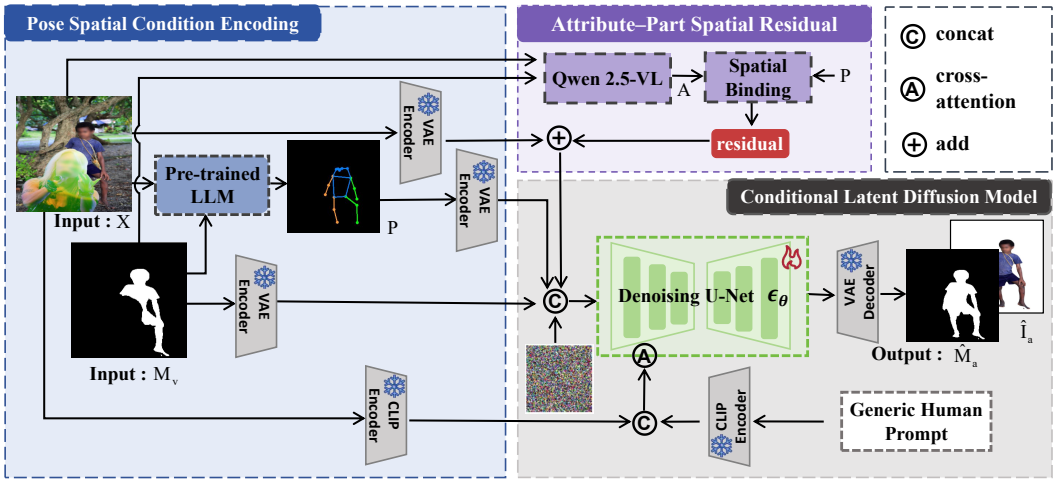


Figure 2: Overview of the proposed ANATOGUARD. Pre-trained LLM predicts a full-body pose from the occluded input, which is rendered as an anatomical spatial condition. The input image, visible mask, and pose map are encoded as latent conditions, while image-context and generic-human features are introduced through cross-attention. Qwen2.5-VL extracts semantic attributes that are localized using pose-derived body-part supports and injected as a residual into the scene latent. These pose and semantic conditions jointly guide the conditional latent diffusion model to generate an anatomically plausible amodal person, from which the final amodal mask is obtained.

$\mathbf{X}, \mathbf{M}_v)$; the conditional generator then models $p(\hat{\mathbf{I}}_a \mid \mathbf{X}, \mathbf{M}_v, \hat{\mathbf{K}}, A)$. The pose estimator and image generator are trained in separate stages. As shown in Fig. 2, the image, visible mask, and rendered skeleton enter the diffusion model as spatial latent conditions. Image and generic-human embeddings provide cross-attention context, while typed attributes are bound to pose-derived part supports and injected as a spatial residual.

3.2 FULL-BODY POSE ESTIMATION

We first predict a full-body pose from the occluded observation. LocLLM formulates keypoint localization as language-conditioned coordinate prediction (Wang et al., 2024). We fine-tune LocLLM on Pix2-Human with occluded-person inputs and full-body keypoint targets, so that supervision includes joints not visible in the input. Given image $\mathbf{X}$, visible mask $\mathbf{M}_v$, and keypoint instruction $\mathcal{Q}$, the model predicts

$$\hat{\mathbf{K}} = f_\phi(\mathbf{X}, \mathbf{M}_v, \mathcal{Q}), \tag{2}$$

where each keypoint is represented by its normalized image coordinates. A fixed renderer $\mathcal{R}$ converts the coordinates and predefined limb connections into an RGB skeleton condition:

$$\mathbf{P} = \mathcal{R}(\hat{\mathbf{K}}). \tag{3}$$

3.3 POSE-CONDITIONED LATENT DIFFUSION

A latent diffusion model reconstructs the complete person image. Let E and D be the frozen VAE encoder and decoder. We encode the target image, occluded input, visible mask, and rendered pose as

$$\mathbf{z}_x = E(\mathbf{I}_a), \qquad \mathbf{z}_X = E(\mathbf{X}), \qquad \mathbf{z}_V = E(\mathbf{M}_v), \qquad \mathbf{z}_P = E(\mathbf{P}). \tag{4}$$

The input image, mask, and pose latents are concatenated channel-wise:

$$\mathbf{C}_{\text{concat}} = [\mathbf{z}_X; \mathbf{z}_V; \mathbf{z}_P] \in \mathbb{R}^{12\times h\times w}. \tag{5}$$

At timestep t, the forward process forms $\mathbf{z}_t = \alpha_t \mathbf{z}_x + \sigma_t \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. The denoising U-Net receives the noisy target latent concatenated with the 12-channel spatial condition and predicts the injected noise:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{X},\epsilon,t}\left[\|\epsilon - \epsilon_\theta([\mathbf{z}_t; \mathbf{C}_{\text{concat}}], t, \mathbf{C}_{\text{cross}})\|_2^2\right]. \tag{6}$$

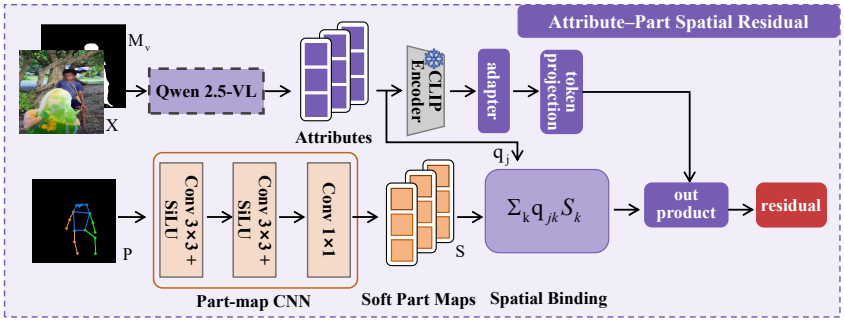


Figure 3: Detailed architecture of the proposed Attribute-Part Spatial Residual. Qwen2.5-VL extracts typed attributes from the input image $\mathbf{X}$ and visible mask $\mathbf{M}_v$, which are encoded into latent features. In parallel, a part-map CNN converts the predicted skeleton $\mathbf{P}$ into soft body-part supports. Spatial binding combines each attribute with its pose-derived support to produce the localized residual R.

Cross-attention context. In addition to spatial concatenation, a frozen CLIP image encoder represents the input image, and a trainable projection maps it to the U-Net context dimension:

$$\mathbf{h}_X = \mathbf{W}_X \, \mathrm{CLIP}_{\mathrm{img}}(\mathbf{X}). \tag{7}$$

A fixed generic-human prompt is encoded by the frozen CLIP text encoder and trainable attribute adapter, producing a second context token h_G. The two tokens form the cross-attention context:

$$\mathbf{C}_{\mathrm{cross}} = [\mathbf{h}_X; \mathbf{h}_G]. \tag{8}$$

Sample-specific Qwen attributes do not enter this global context; they are injected through the localized residual below. This separates broad text context from spatially grounded attribute conditioning. This separation distinguishes global human priors from spatially grounded, instance-specific semantic conditioning. Detailed attribute acquisition and processing are provided in Appendix B.

3.4 Attribute-Part Spatial Residual

Even a pose predictor trained with full-body targets can make errors when joints are heavily occluded or scene evidence is ambiguous. We therefore supplement the pose raster with typed attributes that describe visible evidence, scene context, and plausible hidden continuation. As illustrated in Fig. 3, the proposed module consists of two complementary streams. The attribute stream uses Qwen2.5-VL to obtain sample-specific descriptions and converts them into latent attribute features, while the pose stream transforms the predicted skeleton into soft body-part supports. Spatial binding combines each attribute with its relevant support, yielding a localized residual that is added to the image latent before diffusion denoising.

Qwen2.5-VL generates six attribute groups offline: visible geometry, pose state, scene relation, hidden-completion cue, occlusion relation, and appearance. The appearance query uses the visible target crop; other queries use the configured scene and target context. Let s_j denote attribute text of type j, with presence indicator m_j.

For each slot, a frozen CLIP text encoder produces token features. Mean pooling, a trainable adapter, and a type embedding yield a 768-dimensional attribute representation:

$$\mathbf{u}_j = \mathrm{LN}\left(\mathrm{Adapter}(\mathrm{CLIP}_{\mathrm{text}}(s_j)) + \mathbf{e}_{t_j}\right) m_j. \tag{9}$$

The skeleton image is passed through a learned convolutional mapper g_ψ to produce seven soft support maps,

$$\mathbf{S} = g_\psi(\mathbf{P}) \in [0,1]^{7\times h\times w}, \tag{10}$$

representing the whole person, upper body, lower body, left/right arm, and left/right leg. A deterministic phrase-to-part rule maps each attribute to a binary part selector $\mathbf{q}_j \in \{0,1\}^7$. Thus, language selects the relevant body region while the learned mapper provides its spatial support.

Table 1: Performance comparison on Pix2-Human. In each column, the best is in **bold**, and the second-best is underlined

Method	Publication	mIoU$_{amodal}$(%)	mIoU$_{occ}$(%)
MaskRCNN He et al. (2017)	ICCV'17	79.21	65.18
ORCNN Patrick et al. (2019)	WACV'19	72.07	62.13
DeepSnake Peng et al. (2020)	CVPR'20	67.16	54.66
BCNet Ke et al. (2021)	CVPR'21	76.89	62.91
ShapeDict Xiao et al. (2021)	AAAI'21	76.32	62.30
E2EC Zhang et al. (2022)	CVPR'22	76.25	61.70
C2F Gao et al. (2023)	AAAI'23	83.33	62.38
pix2gestalt Ege et al. (2024)	CVPR'24	85.78	67.12
SAMEO-style En et al. (2025)	CVPR'25	84.79	66.67
Ours	-	**87.24**	**70.02**

The attribute representation is projected to the four latent channels, $\mathbf{v}_j = \tanh(\mathbf{W}_a \mathrm{LN}(\mathbf{u}_j))$. The spatial residual is obtained by combining the projected attribute with its selected support:

$$\mathbf{R} = \gamma \frac{\sum_j m_j \, \mathbf{v}_j \otimes (\sum_k q_{jk} S_k)}{\max\left(1, \sum_j m_j \sum_k q_{jk} S_k\right)}, \tag{11}$$

where γ is a learned scalar initialized to zero and $\otimes$ broadcasts the four-channel vector over spatial locations. The residual is added to the input-image latent:

$$\tilde{\mathbf{z}}_X = \mathbf{z}_X + \mathbf{R}, \qquad \mathbf{C}_{\text{concat}} = [\tilde{\mathbf{z}}_X; \mathbf{z}_V; \mathbf{z}_P]. \tag{12}$$

4 Experimental Protocol

We conduct extensive experiments on the Pix2-Human and 3DOH50K benchmarks to evaluate the proposed pose-guided diffusion framework and compare it with state-of-the-art methods in Section 4.2. Ablation study is presented in Section 4.3.

4.1 Implementation Details

Datasets. Pix2gestalt Ege et al. (2024) is a recent large dataset for amodal completion, which contains both person and a variety of other everyday-object categories. To evaluate our developed method, we collect all images with persons to establish Pix2-Human, which includes 30,000 images in total. Among them, 8,000 images are used to fine-tune the pose estimation network LocLLM. Following the setup in Ege et al. (2024), 90% of the remaining images are used for training the segmentation network and 10% are used for testing. Based on the whole person images provided by the dataset, we use a recent large-scale pretrained human-centric foundation model, Sapiens-0.3B Rawal et al. (2025), to automatically detect the pseudo-labels.

To test the generalization ability of the method, we also perform evaluations on a subset of the 3DOH50K dataset, which contains images collected from real-world scenarios. KINS Qi et al. (2019) and COCO-A Yan et al. (2015) are not used due to the limited number of human samples. The 3DOH50K dataset was originally developed for 3D human shape and pose estimation Zhang et al. (2020a). To adapt it for human-centric amodal segmentation, we derive amodal masks from the mesh annotations. Under the zero-shot setting, 5,000 samples are used for testing.

Evaluation Metrics. To evaluate the quality of the amodal masks, we employ mean Intersection-over-Union (mIoU) as our metric on both Pix2-Human and 3DOH50K. Specifically, we report mIoU over the entire amodal mask (mIoU$_{amodal}$) as well as over occluded regions only (mIoU$_{occ}$). Occluded masks are obtained by subtracting visible masks from amodal masks. While mIoU$_{amodal}$ reflects overall segmentation performance, mIoU$_{occ}$ directly measures the model's ability to recover invisible regions.

Table 2: Performance comparison on 3DOH50K. In each column, the best is in **bold**, and the second-best is underlined.

Method	Publication	mIoU$_{amodal}$(%)	mIoU$_{occ}$(%)
MaskRCNN He et al. (2017)	ICCV'17	69.55	11.77
ORCNN Patrick et al. (2019)	WACV'19	13.73	7.78
DeepSnake Peng et al. (2020)	CVPR'20	29.88	0.66
BCNet Ke et al. (2021)	CVPR'21	69.36	12.59
ShapeDict Xiao et al. (2021)	AAAI'21	45.93	13.90
E2EC Zhang et al. (2022)	CVPR'22	15.49	3.62
C2F Gao et al. (2023)	AAAI'23	1.84	1.25
pix2gestaltEge et al. (2024)	CVPR'24	74.23	17.04
SAMEO-style En et al. (2025)	CVPR'25	73.28	15.67
Ours	-	**78.24**	**20.17**

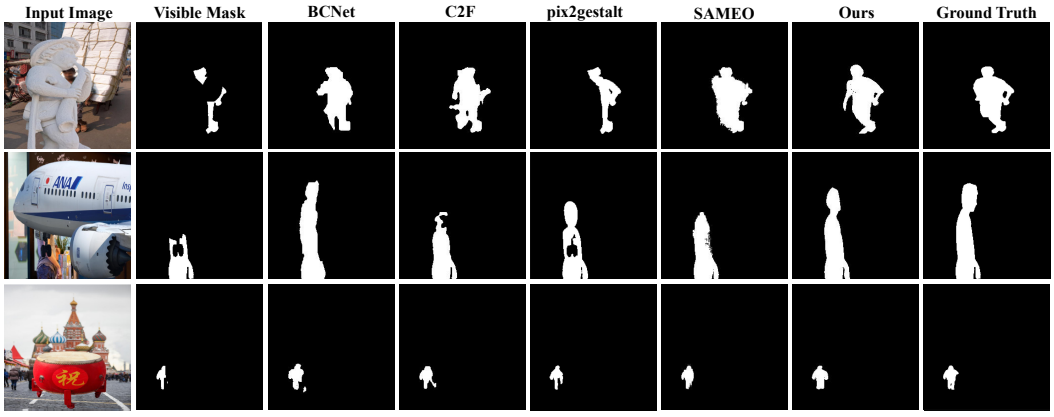


Figure 4: Qualitative comparisons on the Pix2-Human dataset. Under heavy occlusions, existing methods often fail to recover complete human structures, producing missing parts or fragmented shapes. Our method generates anatomically coherent amodal masks guided by full-body pose skeletons.

4.2 Comparison with amodal segmentation methods

4.2.1 Quantitative Comparison

We first compare our method with state-of-the-art amodal instance segmentation approaches on Pix2-Human. Quantitative results are summarized in Table 1. For Pix2-Human, all methods are trained on the dataset under their respective implementations and settings. As the official implementation of SAMEO En et al. (2025) is unavailable, we reproduce its setting by fine-tuning EfficientSAM Xiong et al. (2024) on Pix2-Human. Our method achieves the best performance, reaching 87.24% amodal mIoU and 70.02% occluded-region mIoU. Compared with pix2gestalt Ege et al. (2024), the strongest competing baseline, it improves amodal and occluded-region mIoU by 1.46 and 2.90 percentage points, respectively. The larger gain on occluded regions supports the benefit of explicit full-body pose conditioning for recovering body structures with limited visual evidence.

We further assess cross-dataset generalization by directly applying Pix2-Human-trained models to 3DOH50K without target-domain fine-tuning. The results are shown in Table 2. Our method again obtains the best results, achieving 78.24% amodal mIoU and 20.17% occluded-region mIoU. It surpasses pix2gestalt by 4.01 and 3.13 percentage points on the two metrics, respectively. These improvements indicate that the learned pose-conditioned prior transfers beyond the training distribution and provides more reliable structural guidance under unseen occlusion patterns.

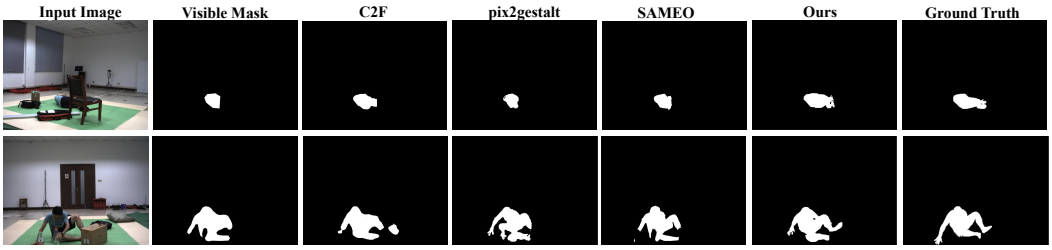


Figure 5: Qualitative comparison on the 3DOH50K dataset under segmentation-training setting. Although no fine-tuning on 3DOH50K, the developed pose estimation network successfully infers postures, which in turn leads to improved amodal instance segmentation.

Table 3: Ablation study on Pix2-Human. Fine-tune: fine-tuned on the Pix2-Human dataset. Generic Prompt: fixed human-prior cross-attention token. Spatial Residual: pose-bound Qwen attribute residual injected into the scene latent. In each column, the best is in **bold**, and the second-best is underlined

Method	$\text{mIoU}_{amodal}(\%)\uparrow$	$\text{mIoU}_{occ}(\%)\uparrow$
Baseline Ege et al. (2024)	85.78	67.12
+ LocLLM pose(off-the-shelf)	85.87 (+0.09)	67.43 (+0.31)
+ LocLLM pose w/Fine-tune	86.99 (+1.21)	69.40 (+2.28)
+ LocLLM pose w/Fine-tune + Generic Prompt	<u>87.14</u> (+1.36)	<u>69.64</u> (+2.52)
+ LocLLM pose w/Fine-tune + Spatial Residual + Generic Prompt	**87.24** (+1.46)	**70.02** (+2.90)

4.2.2 Qualitative Comparison

Fig. 4 and Fig. 5 present qualitative comparisons on Pix2-Human and 3DOH50K, respectively. Under heavy occlusions, existing methods often fail to recover complete human structures, producing missing parts or fragmented shapes. Our method generates anatomically coherent amodal masks guided by full-body pose skeletons, the localized semantic residual further refines regions where pose alone is ambiguous. For example, in the first row of Fig. 4, large portions of the person are heavily occluded, causing existing methods to either miss the invisible regions or generate incorrect predictions. In contrast, our approach correctly estimates the underlying human pose, resulting in more accurate amodal segmentation results. In the bottom row of Fig. 4, our method demonstrates its advantages even for small human figures. Although the pose estimation network is not fine-tuned on 3DOH50K, the developed pose estimation network successfully infers unusual postures, which in turn leads to improved amodal segmentation, as shown in Fig. 5.

4.3 Ablation Study

Table 3 evaluates the effects of pose adaptation, the generic prompt, and the spatial residual. Off-the-shelf LocLLM pose provides only limited gains, while fine-tuning it on Pix2-Human improves amodal and occluded mIoU to 86.99% and 69.40%, respectively. This confirms that full-body pose estimates adapted to occlusion provide stronger structural guidance.

The generic prompt alone has little effect, but improves the fine-tuned pose model to 87.14% amodal mIoU and 69.64% occluded mIoU. The spatial residual further injects sample-specific Qwen attributes into pose-related regions. Combining all components achieves the best results, with 87.24% amodal mIoU and 70.02% occluded mIoU. These results indicate that fine-tuned pose supplies the main structural cue, while the generic prompt and localized attributes provide complementary global and semantic priors.

Pose representation ablation. As illustrated in Fig. 6, we compare four pose-aware conditioning strategies for the diffusion model. The first two encode the predicted keypoints as sequences using an MLP or a Transformer, whereas the latter two convert them into spatial images and encode them with a VAE, using either a heatmap or a skeleton representation. As reported in Table 4, the

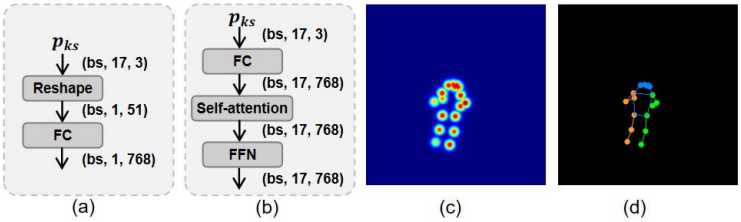


Figure 6: Four types of pose representation for the diffusion model. (a) MLP-based sequence embedding; (b) Transformer-based sequence embedding; (c) Heatmap-based VAE embedding; (d) Skeleton-based VAE embedding.

sequence-based embeddings provide only marginal improvements over the baseline, increasing the amodal mIoU from 85.78% to 85.82% and 85.81%, respectively. In contrast, the image-based representations yield substantially better performance: the heatmap VAE reaches 86.84% $\text{mIoU}_{\text{amodal}}$ and 68.97% mIoU_{occ}, while the skeleton VAE achieves the best results, with 86.99% and 69.40%, respectively. These results suggest that spatial pose representations are better aligned with the pixel-level structure of amodal segmentation than sequence embeddings. The skeleton representation performs best, likely because its explicit limb connectivity provides a concise anatomical scaffold for completing occluded human regions. All comparisons are conducted in the pose-only setting, without the generic prompt or the attribute-part spatial residual.

Table 4: Ablation study on four types of pose representation. In each column, the best is in **bold**, and the second-best is underlined

Method	$\text{mIoU}_{amodal}(\%)$	$\text{mIoU}_{occ}(\%)$
Baseline Ege et al. (2024)	85.78	67.12
+MLP-based embedding	85.82	67.31
+Transformer-based embedding	85.81	67.25
+Heatmap VAE embedding	<u>86.84</u>	<u>68.97</u>
+Skeleton VAE embedding	**86.99**	**69.40**

5 Conclusion

We introduced AnatoGuard, a pose-semantic diffusion framework for human amodal instance segmentation. The key idea is to treat full-body human pose as an explicit anatomical spatial prior rather than merely as an auxiliary prediction target. A pose predictor trained on Pix2-Human estimates complete-body keypoints from occluded images, and the resulting pose representation guides latent diffusion toward anatomically coherent completions. To address the unreliability of pose estimation under severe occlusion, Qwen2.5-VL extracts typed semantic attributes describing appearance, scene context, and hidden-body continuation. These attributes are localized using pose-derived body-part supports and injected as spatial residuals, providing complementary evidence for pose-guided generation.

This design unifies geometric structure and semantic context within a single amodal completion framework. The pose-only and pose-semantic variants also provide a direct way to study the individual contributions of anatomical and semantic conditioning. Future work will focus on improving pose prediction under extreme occlusion, calibrating the interaction between pose and semantic cues, and extending the framework to diverse articulated objects and more challenging real-world scenes.

AI use statement

In this work, we used generative AI tools for language editing and proofreading, including improving grammar, clarity, and readability. They were not used to generate research ideas, develop the proposed methodology, conduct experiments, analyze results, or formulate scientific conclusions.

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Kirillov Alexander, Mintun Eric, Ravi Nikhila, Mao Hanzi, Rolland Chloe, Gustafson Laura, Xiao Tete, Whitehead Spencer, Berg Alexander C., and Lo Wan Yen. Segment anything. In *Int. Conf. Comput. Vis.*, pp. 3992–4003, 2023.

K Chen, D Ramanan, and T Khurana. Using diffusion priors for video amodal segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025.

Seunggeun Chi, Enna Sachdeva, Pin Hao Huang, and Kwonjoon Lee. Contact-aware amodal completion for human-object interaction via multi-regional inpainting. 2025.

Arindam Dutta, Rohit Lal, Dripta S. Raychaudhuri, Calvin-Khang Ta, and Amit K. Roy-Chowdhury. Poise: Pose guided human silhouette extraction under occlusions. In *IEEE Winter Conf. Appl. Comput. Vis.*, 2024.

Ozguroglu Ege, Ruoshi Liu, Dian Chen, Dave Achal, Tokmakov Pavel, and Carl Vondrick. pix2gestalt: Amodal segmentation by synthesizing wholes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3931–3940, 2024.

Tai Wei En, Shih Yu Lin, Sun Cheng, Wang Yu Chiang Frank, and Chen Hwann Tzong. Segment anything, even occluded. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3009–3018, 2025.

Ke Fan, Jingshi Lei, Xuelin Qian, Miaopeng Yu, Tianjun Xiao, Tong He, Zheng Zhang, and Yanwei Fu. Rethinking amodal video segmentation from learning supervised signals with object-centric representation. In *Int. Conf. Comput. Vis.*, pp. 1272–1281, 2023.

Yunxiang Fu, Meng Lou, and Yizhou Yu. Segman: Omni-scale context modeling with state space models and local attention for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025.

Jianxiong Gao, Xuelin Qian, and Yikai Wang. Coarse-to-fine amodal segmentation with shape prior. In *Int. Conf. Comput. Vis.*, pp. 1262–1271, 2023.

Kaiming He, Georgia Gkioxari, Piotr Doll´ar, and Ross Girshick. Mask r-cnn. In *IEEE/CVF International Conference on Computer Vision*, pp. 2961–2969, 2017.

Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 4019–4028, 2021.

Ehsani Kiana, Roozbeh Mottaghi, and Ali Farhadi. Segan: Segmenting and generating the invisible. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 6144–6153, 2018.

Junho Kim, Byung-Kwan Lee, and Yong Man Ro. Causal unsupervised semantic segmentation. *Pattern Recognition*, 171:112173, 2026.

Zhixuan Li, Ruohua Shi, Tiejun Huang, and Tingting Jiang. Oaformer: Learning occlusion distinguishable feature for amodal instance segmentation. In *ICASSP*, 2023.

Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing pose estimation network and a new benchmark. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):871–885, 2018.

H. Ling, D. Acuna, K. Kreis, S. W. Kim, and S. Fidler. Variational amodal object completion. In *Adv. Neural Inform. Process. Syst.*, pp. 16246–16257, 2020.

Zhaochen Liu, Limeng Qiao, Xiangxiang Chu, Lin Ma, and Tingting Jiang. Towards efficient foundation model for zero-shot amodal segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20254–20264, 2025.

Cuong Manh Hoang. Unsupervised instance segmentation with superpixels. *Pattern Recognition*, 172:112402, 2026.

Xuecheng Nie, Jiashi Feng, and Shuicheng Yan. Mutual learning to adapt for joint human parsing and pose estimation. In *Eur. Conf. Comput. Vis.*, pp. 502–517, 2018.

Seung Young Noh and Ju Yong Chang. Phac: Promptable human amodal completion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2026.

George Papandreou, Tyler Zhu, Liang Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. *Springer, Cham*, 2018.

Follmann Patrick, Knig Rebecca, Haertinger Philipp, Klostermann Michael, and Bttger Tobias. Learning to see the invisible: End-to-end trainable amodal instance segmentation. In *IEEE Winter Conference on Applications of Computer Vision*, pp. 1328–1336, 2019.

Sida Peng, Wen Jiang, Huaijin Pi, Xiuli Li, and Hujun Bao. Deep snake for real-time instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3009–3018, 2019.

Khirodkar Rawal, Bagautdinov Timur, Martinez Julieta, Zhaoen Su, James Austin, Selednik Peter, Anderson Stuart, and Saito Shunsuke. Sapiens: Foundation for human vision models. In *Eur. Conf. Comput. Vis.*, pp. 206–228, 2025.

Anvari Taravat, Park Kyoungju, and Kim Ganghyun. Upper body pose estimation using deep learning for a virtual reality avatar. *Applied Sciences*, 13(4), 2023.

Minh Tran, Vo Khoa, Yamazaki Kashu, Fernandes Arthur, Kidd Michael, and Le Ngan. Aisformer: Amodal instance segmentation with transformer. In *Brit. Mach. Vis. Conf.*, 2022.

Minh Tran, Bounsavy Winston, Vo Khoa, Anh Nguyen, Nguyen Tri, and Le Ngan. Shapeformer: Shape prior visible-to-amodal transformer-based amodal instance segmentation. In *IEEE International Joint Conference on Neural Networks*, 2024.

Subarna Tripathi, Maxwell Collins, Matthew Brown, and Serge Belongie. Pose2instance: Harnessing keypoints for person instance segmentation. 2017.

Dongkai Wang, Shiyu Xuan, and Shiliang Zhang. Locllm: Exploiting generalizable human keypoint localization via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Yuting Xiao, Yanyu Xu, Ziming Zhong, Weixin Luo, Jiawei Li, and Shenghua Gao. Amodal segmentation based on visible region segmentation and shape prior. In *AAAI*, pp. 2995–3003, 2021.

Yunyang Xiong, Xiaoyu Xiang Fanyi Xiao Chenchen Zhu Bala Varadarajan, Lemeng Wu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, Raghuraman Krishnamoorthi, and Vikas Chandra. Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024.

Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Amodal completion via progressive mixed context diffusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 9099–9109, 2024.

Zhu Yan, Yuandong Tian, Dimitris N. Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3001–3009, 2015.

Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. Amodal ground truth and completion in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 28003–28013, 2024.

Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene de-occlusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3784–3792, 2020.

Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi Min Hu. Pose2seg: Detection free human instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 889–898, 2019.

Tao Zhang, Shiqing Wei, and Shunping Ji. E2ec: An end-to-end contour-based method for high-quality high-speed instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 4443–4452, 2022.

Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 7374–7383, 2020a.

Ziwei Zhang, Chi Su, Liang Zheng, and Xiaodong Xie. Correlating edge, pose with parsing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 8900–8909, 2020b.

Qiang Zhou, Shiyin Wang, Yitong Wang, Zilong Huang, and Xinggang Wang. Human de-occlusion: Invisible perception and recovery for humans. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3690–3700, 2021.

A Implementation Details

Diffusion architecture. The model takes three RGB inputs at 256×256: the occluded image, visible-person mask, and rendered pose raster. A frozen VAE encodes each input into a four-channel 32×32 latent, yielding a 12-channel spatial condition. This condition is concatenated with the four-channel noisy target latent, so the U-Net input has 16 channels. The U-Net uses 320 base channels, channel multipliers $(1, 2, 4, 4)$, two residual blocks per level, eight attention heads, and cross-attention at downsampling factors 4, 2, and 1. The forward diffusion schedule contains 1,000 steps with linear endpoints 0.00085 and 0.0120.

The frozen image CLIP encoder produces an image token that is mapped to 768 dimensions by a trainable projection. A frozen CLIP ViT-L/14 text encoder, followed by a trainable type-aware adapter, provides the generic-human token and the six attribute embeddings. Thus, cross-attention receives exactly two tokens: the image token and the fixed generic-human-prior token. The six sample-specific Qwen attributes never enter cross-attention directly.

Optimization and sampling. We optimize the diffusion U-Net with AdamW at a base learning rate of 10^{-4}. The image projection, attribute adapter, and attribute-part residual branch use a ten-times larger learning rate. The configuration uses a batch size of 64, 48 data-loader workers, a 100-step linear warm-up, and selects the three checkpoints with the lowest validation EMA simple loss. During training, the code applies classifier-free condition dropout with the inherited parameter 0.05 and applies attribute dropout of 0.15 to the entire sample-specific attribute context. The fixed generic-human token inference uses deterministic DDIM ($\eta = 0$), 200 steps, guidance scale 3.0, and one output per input.

B Attribute Conditioning and Spatial Binding

Each sample may contain six fixed attribute slots: *visible geometry*, *pose state*, *scene relation*, *hidden-completion cue*, *occlusion relation*, and *appearance*. The dataset groups all valid phrases by slot and, by default, concatenates phrases of the same type. Empty slots have a zero presence mask and make no contribution to the residual. The original text strings are tokenized by CLIP with a maximum length of 77; the loader's auxiliary word-token fields are used for metadata and fixed-slot bookkeeping rather than as the CLIP input.

The pose raster is fed to a three-layer convolutional mapper (two 3×3 convolution-SiLU blocks followed by a 1×1 output layer) that predicts seven sigmoid support maps: whole person, upper body, lower body, left/right arm, and left/right leg. A deterministic lexical rule selects one or more supports for each attribute. Mentions of arms, legs, upper-body clothing/head terms, and lower-body clothing/footwear terms activate the corresponding maps. A trainable projection maps each 768-dimensional attribute feature to four latent channels. The selected support and projected feature are multiplied and normalized by total support area, then scaled by a learnable scalar initialized to zero.

Evidence-first Qwen attribute acquisition. We generate sample-specific attributes offline with Qwen2.5-VL-7B-Instruct using a target-anchored, evidence-first question-answering protocol. The main pass receives the original occluded scene and a synthetic target-position anchor derived from that reference image. The anchor is obtained by thresholding the difference from white and applying a 9×9 max-filter dilation; it marks only the visible spatial support of the target and contains no appearance information. The visible reference and anchor identify the target in the full scene, preventing nearby people or occluders from being described as the target.

The main pass requests five attribute types: visible geometry (Q1), pose state (Q2), scene relation (Q3), hidden-completion cue (Q4), and occlusion relation (Q5). It first records sample-specific evidence about visible body regions, pose cues, scene objects, and occlusion cues, then produces concise answers grounded in this evidence.

Visible-only appearance pass. Appearance (Q6) is generated in a separate Qwen pass that receives *only* the visible target-reference RGB image. It can describe visible clothing category, color, coverage, pattern, and texture, but cannot access the occluded scene or an occluder image. The ap-

Table 5: Qwen attribute schema and information sources. Q1–Q5 use target-anchored scene reasoning; Q6 is isolated from scene context.

Type	Question	Permitted input evidence
Visible geometry	Q1	visible target reference and position anchor
Pose state	Q2	target reference, optional skeleton, local context
Scene relation	Q3	target reference and scene context
Hidden-completion	Q4	visible endpoints and optional skeleton
Occlusion relation	Q5	target body evidence and occlusion context
Appearance	Q6	visible target reference

pearance output from the main pass is discarded even if produced accidentally; only the visible-only pass is retained. This hard source separation reduces the risk that an occluder's color, contour, or texture is transferred to the target appearance.

Validation and serialization. Answers are constrained to at most 14 words; the per-type limits are three for visible geometry and two for every other type. The processing script removes placeholders, duplicates, unsafe identity content, and type-inconsistent answers, and limits the candidate pool to at most ten attributes. Candidate validation requires at least four retained attributes by default; samples that fail this criterion are regenerated with an evidence-focused repair instruction for up to two retries. Although Qwen produces grounding, evidence source, and confidence fields for auditing, the diffusion dataset consumes only the compact `type` and `text` fields for each retained attribute.

Fixed generic-human prior. The generic human description is not generated by Qwen and is not part of the six attribute slots. It is the same for every sample: *"A complete person, realistic human body structure, natural pose, natural clothing, consistent lighting."* The frozen CLIP text backbone and trainable adapter encode this description as a dedicated generic-prior token. It is concatenated with the image token in cross-attention, whereas the six Qwen attributes are used exclusively by the attribute-part spatial residual. This separation assigns global human plausibility to the generic prior and sample-specific spatial refinement to Qwen attributes.