SAM-Body4D: Training-Free 4D Human Body Mesh Recovery from Videos

Mingqi Gao [1] **Yunqi Miao** [2] **Jungong Han** [3]

Abstract

Human Mesh Recovery (HMR) aims to reconstruct 3D human pose and shape from 2D observations and is fundamental to human-centric understanding in real-world scenarios. While recent image-based HMR methods such as SAM 3D Body achieve strong robustness on in-the-wild images, they rely on per-frame inference when applied to videos, leading to temporal inconsistency and degraded performance under occlusions. We address these issues without extra training by leveraging the inherent human continuity in videos. We propose SAM-Body4D, a training-free framework for temporally consistent and occlusion-robust HMR from videos. We first generate identity-consistent masklets using a promptable video segmentation model, then refine them with an Occlusion-Aware module to recover missing regions. The refined masklets guide SAM 3D Body to produce consistent full-body mesh trajectories, while a padding-based parallel strategy enables efficient multi-human inference. Experimental results demonstrate that SAM-Body4D achieves improved temporal stability and robustness in challenging in-the-wild videos, without any retraining. Our code and demo are available at: https://github.com/gaomingqi/sam-body4d.

1. Introduction

Human Mesh Recovery (HMR) aims to reconstruct 3D human pose and shape from 2D visual observations, offering explicit representations of human geometry. These representations are valuable for numerous human-centric applications, including human–robot interaction, behaviour analysis, sports performance understanding, immersive VR/AR environments, and embodied AI.

[1]University of Sheffield, Sheffield, UK [2]University of Warwick, Coventry, UK [3]Department of Automation, Tsinghua University, Beijing, China. Correspondence to: Jungong Han <jungonghan77@gmail.com>.

Preprint. December 9, 2025.

Recent advances in image-based HMR have significantly improved generalization in in-the-wild conditions. Most notably, SAM 3D Body (Yang et al., 2025) achieves strong robustness through a scalable data engine, a separate optimization strategy for body and hands, and a decoupled skeleton/shape representation via Momentum Human Rig (MHR (Ferguson et al., 2025)), resulting in more reliable performance than SMPL/SMPL-X-based methods on diverse and complex images. However, when extended to videos, most image-based HMR methods operate in a frame-by-frame manner, relying heavily on independent human detection results for each input image. As per-frame detections lack temporal continuity, the reconstructed human meshes often fluctuate and fail to remain stable in video scenarios (Fig. 1(c)). This becomes particularly problematic in in-the-wild settings where dynamic camera motion, background clutter, and frequent occlusions often cause mixed identities and tracking breakdowns.

Although video-based HMR methods attempt to ensure temporal continuity by modeling temporal information (Kocabas et al., 2020) or incorporating tracking mechanisms, (Wang et al., 2024), they are fundamentally optimization based, which demands large annotated video datasets and carefully crafted objectives. Such reliance restricts their scalability and weakens the robustness to diverse and unpredictable in-the-wild human motions and scene dynamics.

Building upon this insight, we propose SAM-Body4D, a training-free framework for temporally consistent HMR from videos. Given an input video, SAM-Body4D generates identity-preserving mesh trajectories for the target humans. We first track and segment the target pixels using a promptable video segmentation model, producing identity-consistent masklets that carry temporal continuity. These masklets then serve as prompts to guide SAM 3D Body, transferring the inherent temporal coherence of videos to the reconstructed 4D human meshes. Furthermore, an Occlusion-Aware Refiner is introduced to recover missing or corrupted regions caused by occlusions, preventing hallucinated predictions. Additionally, a padding-based parallel strategy enables efficient multi-person and multi-frame inference without modifying pre-trained models. An overview of our framework is shown in Fig. 1.

Our main contributions are summarized as follows:

Figure 1. Illustration of temporally consistent Human Mesh Recovery (HMR) from videos. (a) Input video frames. (b) Identity-consistent human masks, where each person is highlighted with a unique and consistent colour across frames. (c) Vanilla image-to-video HMR baseline using SAM 3D Body with automatic human detection and per-frame inference. Note that only the meshes corresponding to the masks in (b) are visualised here; if a mesh does not appear in a certain frame, it indicates that the corresponding person is not detected in that frame. (d) Our spatial-temporal consistent HMR, where the temporal continuity in masklets is directly propagated into the 4D human meshes. (e) Our full SAM-Body4D with occlusion-aware refinement. Across the 2nd–5th columns, SAM-Body4D recovers plausible and temporally stable reconstructions under occlusion. As these humans are heavily occluded, their complete meshes are visualised in the bottom-left corner for clearer observation.

- We present SAM-Body4D, a training-free and scalable framework for temporally consistent and robust human mesh recovery from in-the-wild videos.
- We achieve identity-consistent 4D mesh trajectories across complex and dynamic scenes using temporally aligned masklets as prompts.
- We propose an occlusion-aware refinement mechanism that improves reconstruction quality under occlusions.

2. Related Work

Human Mesh Recovery. Image-based human mesh recovery (HMR) predicts 3D human body meshes directly from a RGB image. Existing approaches can be broadly categorized into regression-based methods and token-based methods. The former directly regresses parameters of 3D human models, such as SMPL-X (Pavlakos et al., 2019), from image features (*e.g.*, HMR (Kanazawa et al., 2018), SPIN (Pavlakos et al., 2018), and PromptHMR (Wang et al., 2025)) while the latter represent joints or mesh vertices as learnable tokens and employ transformer reasoning to model their relationships, enabling more flexible and expressive structured prediction (*e.g.* TokenHMR (Dwivedi et al., 2024) and MEGA (Fiche et al., 2025)).

Despite the strong performance of image-based HMR methods, they do not naturally adapt to video settings, where challenges such as temporal consistency and occlusions become critical. To enforce temporal smoothness, feature-level temporal modeling approaches, such as VIBE (Kocabas et al., 2020) and TRAM (Wang et al., 2024), build upon image-based feature encoders and introduce temporal modules such as GRU and transformers to aggregate frame-wise representations and learn coherent motion cues over time. 4DHumans (Goel et al., 2023), on the other hand, jointly performs mesh reconstruction and identity-consistent tracking, enabling stable human modeling under occlusions and rapid pose variation. In addition, these approaches are fundamentally optimization-based, requiring large amounts of manually annotated video data and carefully designed loss functions, which restrict their generalizability and scalability. In contrast, our method is entirely training-free, enforcing temporal coherence from the source by directly leveraging the pixel-level continuity of humans in video, rather than relying on feature- or pose-space temporal modelling where such continuity may already be lost.

Video Object Segmentation (VOS) focuses on tracking and segmenting a target object across video frames, typically based on a few pixel/box annotations (Gao et al., 2023) or text prompts (Zhou et al., 2022). Prior VOS meth-

ods generally follow a memory-based paradigm (Oh et al., 2019; Cheng & Schwing, 2022; Cheng et al., 2024), where historical predictions are leveraged to maintain temporal consistency. However, these approaches still struggle to generalize to in-the-wild scenarios due to limited model capacity and training diversity.

Benefiting from billion-scale training data and strong transformer architectures, SAM (Kirillov et al., 2023) demonstrates high segmentation accuracy, strong generalisation to in-the-wild images, and rich prompting modalities. By introducing a memory mechanism, SAM 2 (Ravi et al., 2024) extends these strengths to the VOS setting. More recently, SAM 3 (Carion et al., 2025) incorporates text prompts and an independent object perceiver, enabling more user-friendly interactions and stronger robustness to common challenges such as disappearance and reappearance in complex videos.

Despite strong performance on visible regions, SAM-based VOS methods cannot recover occluded body parts, resulting in incomplete visual cues for downstream HMR and causing hallucinated geometry during occlusion. To address this limitation, we introduce an occlusion-aware refinement module that reconstructs hidden regions and provides full-body references for robust and consistent HMR in videos.

3. Preliminary

SAM 3 (Carion et al., 2025) is a promptable object segmentation model supporting both images and videos. Given a video $\mathcal{V} = \{I_t\}_{t=1}^{T}$, $I_t \in \mathbb{R}^{H\times W\times 3}$, and user-defined prompts $\mathcal{P}$, SAM 3 predicts a mask sequence $\mathcal{M} = \{M_t\}_{t=1}^{T}$, $M_t \in \{0,1\}^{H\times W}$.

For the t-th frame, the mask is obtained by combining two complementary modules: `propagate` and `detect`.

$$\begin{aligned} \hat{M}_t &= \text{propagate}(M_{t-1}), \\ O_t &= \text{detect}(I_t, \mathcal{P}), \\ M_t &= \text{match\_and\_update}(\hat{M}_t, O_t). \end{aligned} \tag{1}$$

The `propagate` module leverages spatial-temporal correspondences between historical predictions (stored in memory) and the current frame, enabling reliable transfer of mask labels across time and preserving target continuity throughout the video. In contrast, the `detect` module focuses on semantic associations between the prompt and objects in the current frame, which is particularly effective for challenging cases such as disappearance and reappearance in complex scenes. By integrating the outputs of the two complementary modules through `match_and_update`, SAM 3 achieves substantially higher accuracy than SAM 2 and other VOS methods on complex videos.

SAM 3D Body (Yang et al., 2025) is a promptable model supporting Human Mesh Recovery (HMR) from in-the-wild images. It accepts prompts at both the encoder and decoder stages. Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, SAM 3D Body performs feature encoding as:

$$F = \text{ImgEncoder}(I, \mathcal{P}_{\text{enc}}), \tag{2}$$

where $\mathcal{P}_{\text{enc}}$ denotes optional encoder prompts (e.g., 2D keypoints or segmentation masks) that help the model focus on the target human. With encoded features, the decoder predicts full-body tokens as:

$$O = \text{Decoder}(F, \mathcal{P}_{\text{dec}}), \tag{3}$$

where $\mathcal{P}_{\text{dec}}$ includes optional prompts such as keypoint, camera, or MHR tokens. Then the first output token O_0 is passed through an MLP to obtain MHR parameters:

$$\theta = \{P, S, C, S_k\} = \text{MLP}(O_0), \tag{4}$$

where P, S, C, and S_k denote pose, shape, camera pose, and skeleton parameters.

During inference, body and hand are optimised separately due to different optimisation strategies and later fused. Unless otherwise specified, θ indicates the final full-body mesh parameters containing both body and hand information.

4. Methodology

4.1. Overview

The framework of SAM-Body4D is illustrated in Fig. 2. Given an input video $\mathcal{V} = \{I_t\}_{t=1}^{T}$ and prompts $\mathcal{P} = \{\mathcal{P}^{h_i}\}_{i=1}^{N}$ indicating N target humans, SAM-Body4D estimates temporally consistent human mesh parameters $\Theta = \{\theta_t^{h_i}\}_{t=1,i=1}^{T,N}$ for all selected persons.

SAM-Body4D consists of three key components. A *Masklet Generator* produces identity-consistent masklets as temporal tracking cues. An *Occlusion-Aware Masklet Refiner* enhances these masklets by recovering missing regions when occlusion occurs. A *Mask-Guided HMR module* then predicts per-frame mesh parameters θ_t. Since each mesh is aligned with its corresponding mask over time, the temporal continuity in masklets is naturally propagated to the reconstructed human meshes.

4.2. Masklet Generator

For each target human h_i specified by prompts $\mathcal{P}$, we apply SAM 3 (Carion et al., 2025) over the video $\mathcal{V}$ to obtain spatio-temporally aligned masklets $\mathcal{M} = \{M_t^{h_i}\}$. Following the hybrid propagation-detection formulation in Eq. 1, identity consistency is maintained across frames. Note that other video segmentation models capable of producing temporally aligned instance masks can also be adopted here.

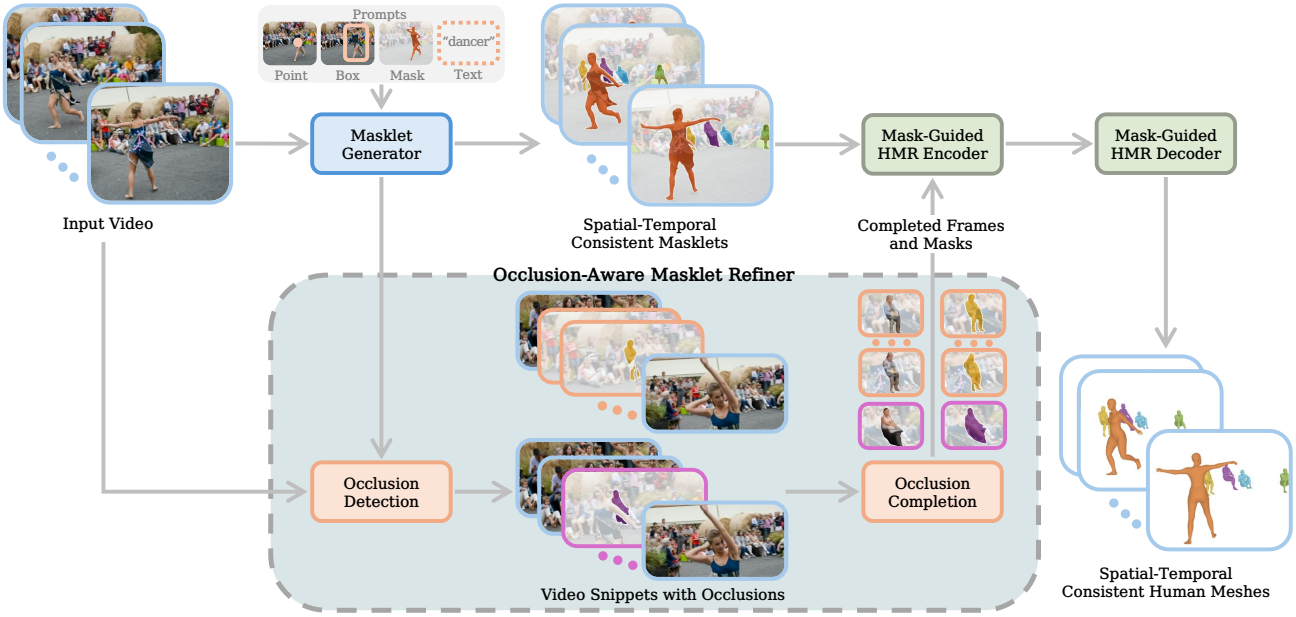

Figure 2. Overall framework of the proposed SAM-Body4D. Given an input video with human prompts, SAM-Body4D operates on three main modules in a training-free manner. The Masklet Generator derives identity-consistent temporal masklets from the video to provide spatio-temporal tracking cues. The Occlusion-Aware Masklet Refiner enriches these masklets by recovering invisible body regions and stabilizing temporal alignment. Finally, the Mask-Guided HMR module uses refined masklets as spatial prompts to predict accurate and temporally coherent human meshes across the entire sequence.

4.3. Occlusion-Aware Masklet Refiner

In in-the-wild videos, humans frequently undergo severe occlusions, where even state-of-the-art segmentation methods like SAM 3 can only capture visible body regions. Such incomplete masklets provide insufficient visual evidence for human mesh estimation and may lead to hallucinated predictions with unrealistic pose and shape. To resolve this issue, we introduce an occlusion-aware refinement module to recover missing regions and offer complete visual references for the subsequent HMR stage.

We first obtain mask completion results by feeding video frames $\mathcal{V}$ and masklets $\mathcal{M}$ into a mask completion model (Diffusion-VAS (Chen et al., 2025)), producing completed masklets $\tilde{\mathcal{M}} = \{\tilde{M}_t^{h_i}\}$. For each human h_i at frame t, occlusions are detected when the completed mask area becomes larger while the overlap remains low:

$$\text{occ}(t, h_i) = \mathbb{1}\begin{pmatrix} |\tilde{M}_t^{h_i}| > |M_t^{h_i}| \\ \wedge\ \text{IoU}(\tilde{M}_t^{h_i}, M_t^{h_i}) < 0.7 \end{pmatrix} \quad (5)$$

To obtain accurate visual evidence for the occluded regions, the detected frames are temporally grouped and re-fed to Diffusion-VAS to recover missing pixels (see the yellow module in Fig. 2). We then update the corresponding frames and masks as:

$$I_t^{h_i} \leftarrow \tilde{I}_t^{h_i}, \quad M_t^{h_i} \leftarrow \tilde{M}_t^{h_i}, \quad (6)$$

yielding refined video $\tilde{\mathcal{V}}$ and refined masklets $\tilde{\mathcal{M}}$ with full-body visual cues and improved temporal stability, which provide more reliable supervision for the subsequent HMR module and enable more accurate and consistent meshes across challenging frames.

4.4. Training-Free Mask-Guided HMR

With refined video $\tilde{\mathcal{V}}$ and refined masklets $\tilde{\mathcal{M}}$, we perform training-free HMR. For each target human h_i, the corresponding mask $\tilde{M}_t^{h_i}$ is used as the encoder prompt $\mathcal{P}_{\text{enc}}$ in Eq. 2 and Eq. 3, enabling the model to focus on the correct identity. This produces per-frame mesh parameters $\Theta = \{\theta_t^{h_i}\}_{t=1,i=1}^{T,N}$ consistent with the refined masklets over time. Importantly, the entire pipeline operates in a **training-free** manner without any task-specific finetuning.

We further improve the efficiency of the SAM 3D Body HMR stage. The original pipeline performs per-frame sequential inference, and the number of visible humans may vary across frames, making naive batching infeasible. We introduce a simple padding mechanism to unify the batch shape so that all humans within the same frame batch can be processed jointly in a single forward pass. This parallelisation eliminates redundant per-human inference and yields a substantial speed-up while preserving the temporal alignment of the predicted meshes.

To further enhance motion stability, we apply lightweight test-time temporal smoothing to the Momentum Human Rig pose and hand parameters, reducing jitter and promoting

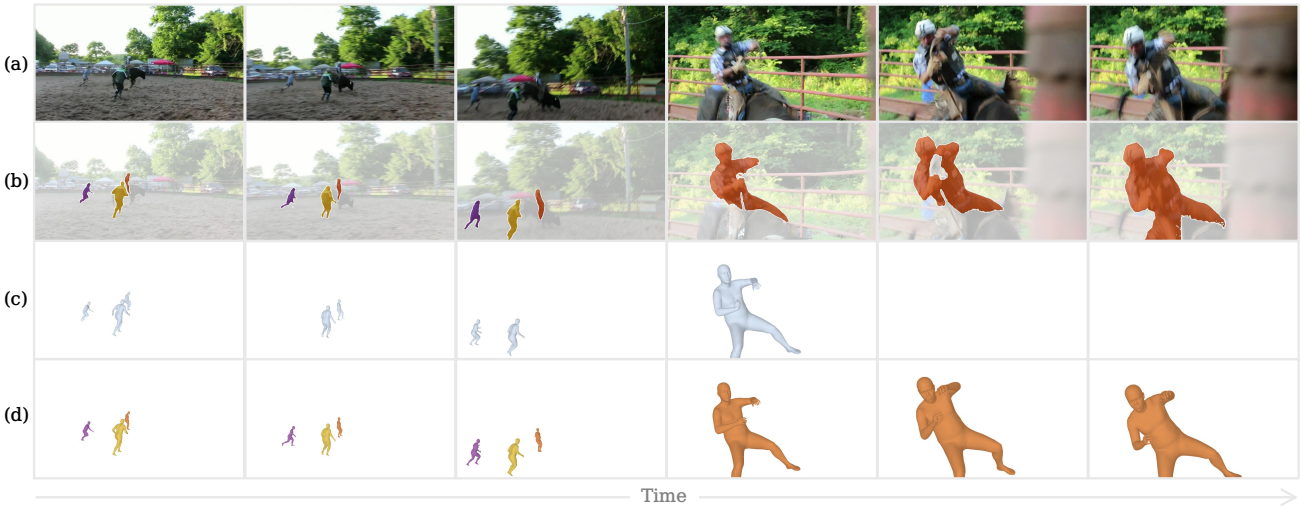

Figure 3. Visualised comparisons between the vanilla image-to-video extension of SAM 3D-Body and our SAM-Body4D. (a) Input video frames. (b) Identity-consistent human masks. (c) Vanilla per-frame HMR results using SAM 3D-Body with automatic human detection, where missed detections lead to missing meshes. (d) Our SAM-Body4D maintains temporally continuous and identity-preserving mesh trajectories throughout the video by leveraging spatial-temporal masklet guidance.

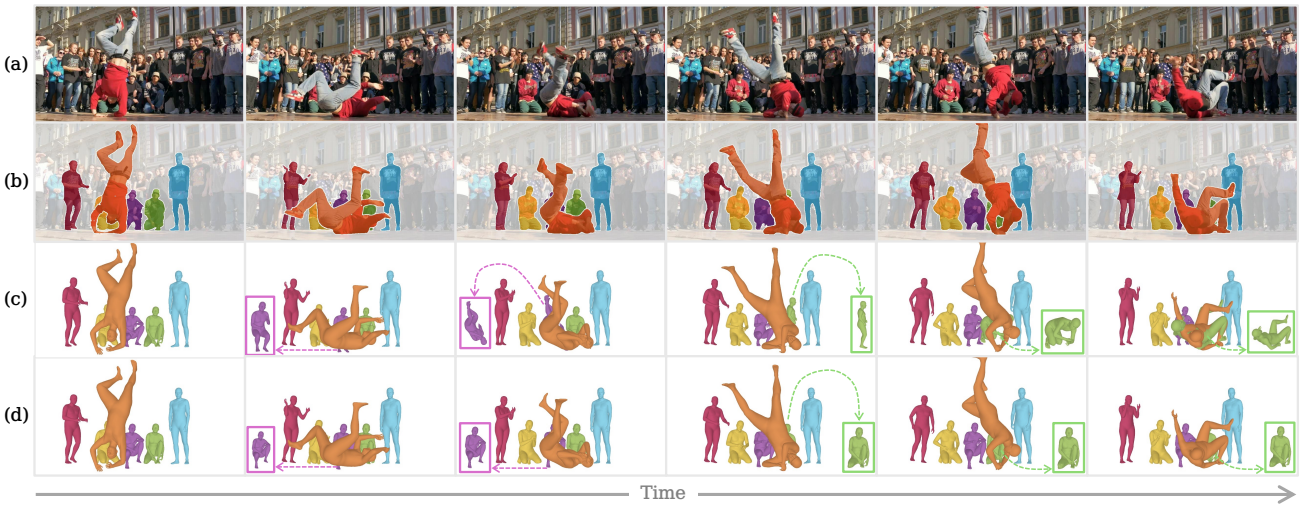

Figure 4. Visualised comparisons between SAM-Body4D w/o and w/ Occlusion-Aware Masklet Refiner. (a) Input video frames; (b) Temporally consistent human masks, where each person is highlighted with a unique and consistent color across frames; (c) SAM-Body4D without Occlusion-Aware Masklet Refiner; (d) SAM-Body4D with Occlusion-Aware Masklet Refiner. Across the 2nd–6th columns, SAM-Body4D produces more robust reconstructions under occlusion (e.g., the blue-rendered person in the 2nd column, the purple-rendered people in the 3rd/4th column, and the green-rendered people in the 5th and 6th columns). Since these subjects are heavily occluded, their meshes without occlusion are shown at the bottom-left/bottom-right for clearer observation.

smooth transitions. In addition, for each target human, the scale and shape parameters from the first visible frame are reused across the entire sequence to maintain consistent body proportions and avoid identity drift. These operations require no learning and introduce negligible computational overhead, keeping the entire framework fully **training-free**.

The full procedure of SAM-Body4D is summarised in Algorithm 1, which complements the structural illustration in Fig. 2 by explicitly detailing the execution flow, including identity-consistent masklet generation, occlusion-aware refinement, and our parallel mask-guided HMR strategy.

5. Experiments

This section presents qualitative results to demonstrate the effectiveness of our training-free framework for video HMR. Our implementation combines SAM 3 (Carion et al., 2025) as the Masklet Generator, Diffusion-VAS (Chen et al., 2025) for occlusion detection and refinement, and SAM 3D Body (Yang et al., 2025) for Mask-Guided HMR.

Algorithm 1 Training-Free Mask-Guided Video HMR

Input: Video $\mathcal{V} = \{I_t\}_{t=1}^{T}$, prompts $\mathcal{P} = \{\mathcal{P}^{h_i}\}_{i=1}^{N}$
Output: Mesh parameters $\Theta = \{\theta_t^{h_i}\}_{t=1,i=1}^{T,N}$

// 1. Masklet Generation (SAM 3)
Obtain initial masklets $\mathcal{M} = \{M_t^{h_i}\}$ using SAM 3 via Eq. 1.
// 2. Occlusion-Aware Refinement (Diffusion-VAS)
Complete masks using Diffusion-VAS to obtain $\tilde{\mathcal{M}} = \{\tilde{M}_t^{h_i}\}$.
for each frame t and human h_i **do**
  if occlusion $\text{occ}(t, h_i)$ detected by Eq. 5 **then**
    Recover missing pixels and update video and masks via Eq. 6.
  end if
end for
Obtain refined video $\tilde{\mathcal{V}}$ and refined masklets $\tilde{\mathcal{M}}$.

// 3. Parallel Mask-Guided HMR (SAM 3D Body)
Construct frame batches and pad missing humans to a fixed batch shape.
for each frame batch **do**
  Use $\tilde{M}_t^{h_i}$ as encoder prompts $\mathcal{P}_{\text{enc}}$.
  Compute $\theta_t^{h_i}$ via Eq. 2 and Eq. 3 in a single forward pass.
end for

// 4. Temporal Smoothing in MHR Space
for each human h_i **do**
  Fix scale and shape to those from the first visible frame of h_i.
  Apply Kalman smoothing to the per-frame pose and hand parameters of h_i.
end for

return Θ

The detection and refinement stages operate at a spatial resolution of 512×1024, while lower resolutions can further improve efficiency with a moderate loss in visual fidelity.

The pipeline supports efficient deployment on a single GPU. Without the occlusion refiner, our parallel multi-frame inference achieves substantial speed improvements over sequential per-frame HMR. For instance, on an NVIDIA A100-SXM4-80GB (96GB system memory), processing a 480×854 video (90 frames, 5 persons) runs approximately $2\times$ faster with a parallel batch size of 32. When the refiner is enabled, both memory usage and runtime increase depending on the duration of the occlusion and the number of persons being refined.

5.1. Comparison with Vanilla Image-to-Video Extension

We evaluate the vanilla image-to-video extension of SAM 3D Body, where mesh prediction is performed independently on each frame without any temporal enforcement. As shown in Fig. 3, such a per-frame strategy cannot ensure continuous mesh trajectories for the same person across the video. When the target becomes small, suffers motion blur, or is heavily occluded, the detector may fail to localise the human, leading to missing meshes.

In contrast, SAM-Body4D leverages temporally aligned masklets to provide consistent target localisation throughout the sequence. These masklets are used as element-wise prompts for HMR, effectively transferring pixel-level continuity to 4D human meshes. Owing to the one-to-one correspondence between mask regions and reconstructed meshes, identity association remains stable, even when visibility temporarily degrades.

This comparison highlights that enforcing spatial-temporal continuity at the pixel level is essential for reliable and stable video HMR. By preserving consistent localisation cues, SAM-Body4D maintains identity association and smooth mesh evolution across frames, enabling coherent 4D reconstruction throughout the video.

5.2. Effectiveness of Occlusion-Aware Refinement

Occlusions occur frequently in real-world videos, where major body areas are temporarily hidden by objects or other people. In such cases, per-frame HMR relies on incomplete visual evidence and easily hallucinates implausible body structures. To demonstrate the benefit of our refinement, we present visual comparisons under challenging occlusion scenarios in Fig. 4.

When only a small portion of the body is occluded (e.g., first column of Fig. 4), the vanilla SAM 3D Body can still produce reasonable predictions. However, once most of the body becomes invisible, the baseline depends solely on the limited visible pixels and generates distorted and unstable meshes. In contrast, our occlusion-aware refiner restores missing human regions before HMR inference, enabling SAM-Body4D to preserve plausible pose and consistent body structure throughout the occluded frames.

These results highlight that recovering occluded body evidence is essential to mitigate hallucinated predictions and achieve reliable 4D human reconstruction in challenging in-the-wild videos.

6. Conclusion

We presented SAM-Body4D, a training-free framework for temporally consistent Human Mesh Recovery (HMR) from videos. By leveraging identity-consistent masklets and an occlusion-aware refinement module, our approach effectively transfers pixel-level continuity into coherent 4D human mesh reconstruction. Without requiring any additional training or architectural modification to SAM-3D-Body, SAM-Body4D improves temporal stability and robustness in challenging in-the-wild scenarios. Our parallel multi-frame inference strategy further enables efficient and scalable deployment in practical applications.

References

Carion, N., Gustafson, L., Hu, Y.-T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K. V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.-H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H. K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., and Feichtenhofer, C. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

Chen, K., Ramanan, D., and Khurana, T. Using diffusion priors for video amodal segmentation. In *CVPR*, pp. 22890–22900, 2025.

Cheng, H. K. and Schwing, A. G. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, pp. 640–658. Springer, 2022.

Cheng, H. K., Oh, S. W., Price, B., Lee, J.-Y., and Schwing, A. Putting the object back into video object segmentation. In *CVPR*, pp. 3151–3161, 2024.

Dwivedi, S. K., Sun, Y., Patel, P., Feng, Y., and Black, M. J. Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In *CVPR*, pp. 1323–1333, 2024.

Ferguson, A., Osman, A. A., Bescos, B., Stoll, C., Twigg, C., Lassner, C., Otte, D., Vignola, E., Bogo, F., Santesteban, I., et al. Mhr: Momentum human rig. *arXiv preprint arXiv:2511.15586*, 2025.

Fiche, G., Leglaive, S., Alameda-Pineda, X., and Moreno-Noguer, F. Mega: Masked generative autoencoder for human mesh recovery. In *CVPR*, pp. 5366–5378, 2025.

Gao, M., Zheng, F., Yu, J. J., Shan, C., Ding, G., and Han, J. Deep learning for video object segmentation: a review. *Artificial Intelligence Review*, 56(1):457–531, 2023.

Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., and Malik, J. Humans in 4d: Reconstructing and tracking humans with transformers. In *ICCV*, pp. 14783–14794, 2023.

Kanazawa, A., Black, M. J., Jacobs, D. W., and Malik, J. End-to-end recovery of human shape and pose. In *CVPR*, pp. 7122–7131, 2018.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. In *ICCV*, pp. 4015–4026, 2023.

Kocabas, M., Athanasiou, N., and Black, M. J. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, pp. 5253–5263, 2020.

Oh, S. W., Lee, J.-Y., Xu, N., and Kim, S. J. Video object segmentation using space-time memory networks. In *ICCV*, pp. 9226–9235, 2019.

Pavlakos, G., Zhu, L., Zhou, X., and Daniilidis, K. Learning to estimate 3d human pose and shape from a single color image. In *CVPR*, pp. 459–468, 2018.

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A. A., Tzionas, D., and Black, M. J. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al. Sam 2: Segment anything in images and videos. In *ICLR*, 2024.

Wang, Y., Wang, Z., Liu, L., and Daniilidis, K. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *ECCV*, pp. 467–487. Springer, 2024.

Wang, Y., Sun, Y., Patel, P., Daniilidis, K., Black, M. J., and Kocabas, M. Prompthmr: Promptable human mesh recovery. In *CVPR*, pp. 1148–1159, 2025.

Yang, X., Kukreja, D., Pinkus, D., Sagar, A., Fan, T., Park, J., Shin, S., Cao, J., Liu, J., Ugrinovic, N., Feiszli, M., Malik, J., Dollar, P., and Kitani, K. Sam 3d body: Robust full-body human mesh recovery. *arXiv preprint; identifier to be added*, 2025.

Zhou, T., Porikli, F., Crandall, D. J., Van Gool, L., and Wang, W. A survey on deep learning technique for video segmentation. *IEEE TPAMI*, 45(6):7099–7122, 2022.