DREAMHMR: FROM VIDEO GENERATION TO HUMAN MOTION MODELING VIA STRUCTURAL FOCUSING

Anonymous authors
Paper under double-blind review

ABSTRACT

Human mesh recovery (HMR) provides a geometric foundation for understanding human motion, but accurately reconstructing both body motion and fine hand articulation from videos remains challenging, especially in real-world scenes where occlusion and partial visibility are common. Video generation pretraining offers rich spatial and temporal priors for this task, but its focus on visual synthesis leaves a gap between these representations and explicit human geometry. We present DreamHMR, a general framework that adapts video generation priors for complete body-and-hand reconstruction. It follows a global-to-local strategy, first grounding video representations in human geometry and then refining body and hand detail with reliable local evidence. Global Body Representation Transfer trains the video backbone to predict complete human renderings, turning its native prediction task into a means of learning surface correspondence and posed geometry. Local Joint Feature Enhancement combines contextual estimates with reliable local evidence to recover fine body and hand detail, while limiting interference from occluders and background. Experiments on five datasets with four video backbones demonstrate strong reconstruction performance across backbone families. Under a unified evaluation protocol, our best model outperforms the best existing methods on all reported metrics of 3DPW, EMDB-1, and RICH, reducing body MPJPE by 4.2%, 4.7%, and 6.1%, respectively. In more challenging occlusion settings on held-out BEDLAM and EgoBody sequences, DreamHMR reduces overall body and hand MPJPE by 7.0% and 10.0%, respectively. Our code will be made publicly available.

1 INTRODUCTION

Human mesh recovery (HMR) reconstructs the 3D human body from visual observations, providing a geometric foundation for understanding and modeling human motion. Its estimates support a wide range of downstream applications, from structured motion analysis for action recognition (Goel et al., 2023) to motion references for humanoid robot teleoperation and imitation learning (He et al., 2024; Fu et al., 2025; Zhao et al., 2025; Luo et al., 2026). The field has advanced rapidly, with progress in monocular parametric body estimation (Kanazawa et al., 2018; Kolotouros et al., 2019), temporal modeling from video (Kocabas et al., 2020), and large-scale learning from diverse imagery (Goel et al., 2023). Despite these advances, recovering complete body and hand motion from real-world videos remains challenging, as occlusion and partial visibility leave incomplete and uneven visual evidence across body parts and frames.

Recovering motion from such observations requires both inferring missing structure and identifying reliable local evidence. Information from visible body parts and neighboring frames can help constrain unobserved body and hand configurations. However, local features sampled near a joint may correspond to an occluder or the background rather than the target person, so incorporating them indiscriminately can introduce misleading cues. A reconstruction model therefore needs to combine contextual inference of the complete body with selective use of local observations for detailed body and hand recovery.

Video generation pretraining offers a promising starting point for learning such contextual representations (Team Wan et al., 2025). By learning to synthesize videos, these models capture spatial appearance and temporal dependencies within a shared representation, offering potentially useful

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.

cues for reasoning about body and hand motion under partial observations. However, their training objective targets visual synthesis rather than explicit human pose and surface geometry. Transferring these representations to human reconstruction therefore requires adapting them to the underlying structure of the human body.

To this end, we present DreamHMR, a framework that adapts video generation representations for complete body-and-hand reconstruction through geometric supervision and evidence-gated local refinement. First, *Global Body Representation Transfer* trains the video backbone to predict renderings of complete SMPL-X bodies (Pavlakos et al., 2019), turning its video prediction objective into a proxy for learning human structure. Canonical surface colors encode correspondence across poses, while shading conveys pose-dependent geometry, providing supervision for body regions even when they are absent from the observations. Second, *Local Joint Feature Enhancement* refines contextual body and hand estimates using features sampled around predicted joints. Supervised evidence scores predict target-person support at these locations and weight the local updates, limiting the influence of features drawn from occluders or background.

Experiments spanning five datasets and four video backbones demonstrate strong reconstruction performance, with gains extending across backbone families. On standard benchmarks, our Wan-5B model outperforms the best existing methods across all reported metrics under a unified evaluation protocol, reducing body MPJPE by 4.2%, 4.7%, and 6.1% on 3DPW, EMDB-1, and RICH, respectively. In challenging occlusion settings on held-out BEDLAM and EgoBody sequences, it reduces overall body and hand MPJPE by 7.0% and 10.0%, respectively.

DreamHMR makes the following contributions:

- We bring pretrained video generation priors to complete body-and-hand reconstruction through *Global Body Representation Transfer*, offering a new approach to recovering human motion from occluded and incomplete observations.
- We introduce *Local Joint Feature Enhancement* to recover fine body and hand detail from reliable local evidence, complementing global reconstruction within a general framework applicable across different video backbones.
- Extensive experiments on five datasets and four video backbones demonstrate that DreamHMR outperforms the best existing methods on standard HMR benchmarks, with body and hand MPJPE reduced by 7.0% and 10.0%, respectively, in challenging occlusion settings.

2 Related Work

Human reconstruction from images and videos. Human mesh recovery estimates parametric bodies from images (Kanazawa et al., 2018; Kolotouros et al., 2019), with video models adding temporal context (Kocabas et al., 2020; Goel et al., 2023). PromptHMR-Vid combines full-image features with a box-prompted temporal transformer (Wang et al., 2025); GENMO learns a motion distribution for generation and estimation (Li et al., 2025). Whole-body methods recover articulated bodies and hands (Pavlakos et al., 2019; Lin et al., 2023b; Li et al., 2023; Baradel et al., 2024; Moon, 2026), scale training data and model capacity (Cai et al., 2023; Yin et al., 2026; Yang et al., 2026); dedicated models recover detailed hand meshes (Pavlakos et al., 2024; Potamias et al., 2025). DreamHMR uses a shared video generation representation for body-and-hand recovery.

Generative representations for reconstruction. Image diffusion features support correspondence estimation (Tang et al., 2023; Stracke et al., 2025), while video diffusion features support motion and geometry tasks (Vélez et al., 2025; Nam et al., 2025; Huang et al., 2026). DPMesh conditions a text-to-image diffusion backbone on 2D observations for occluded mesh recovery (Zhu et al., 2024), and ADHMR optimizes diffusion-based mesh predictions with preference data (Shen et al., 2025). DreamHMR adapts video generation features with complete human renderings to encode surface correspondence and posed geometry for body-and-hand recovery.

Reconstruction under occlusion. Occluded HMR addresses body reconstruction with incomplete observations (Khirodkar et al., 2022). PARE aggregates neighboring regions with part-guided attention (Kocabas et al., 2021), and Divide and Fuse merges part meshes from partially visible im-

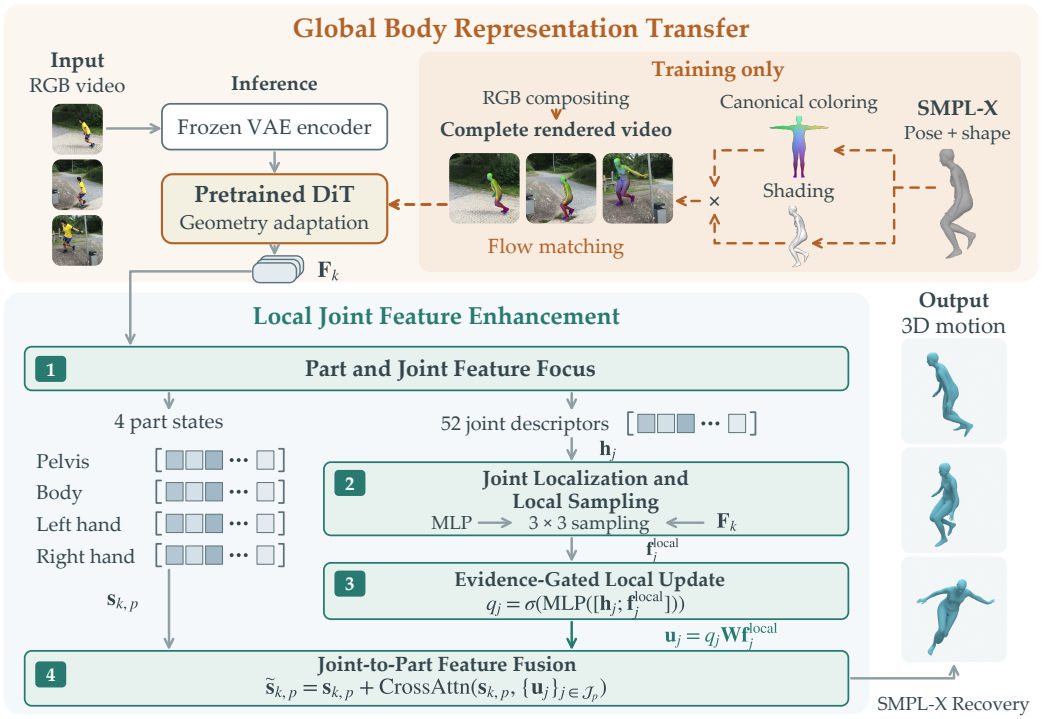

Figure 1: **DreamHMR architecture.** Global Body Representation Transfer adapts a video DiT with complete SMPL-X renderings. Local Joint Feature Enhancement samples joint-local features, gates their corrections by target-pixel support, and fuses them into body-and-hand part states.

ages (Luan et al., 2024). Human and joint tokens structure decoding (Sun et al., 2024), while local image-aligned features refine hands (Potamias et al., 2025). Nearby features may describe background or another person; DreamHMR supervises target-pixel support to weight their updates.

3 Method

DreamHMR recovers complete body-and-hands motion through Global Body Representation Transfer, which adapts video features with complete rendered human structure, and Local Joint Feature Enhancement, which refines anatomical states with local image evidence.

3.1 Overall framework

Given target-person RGB $I_{1:T}$, binary observation masks $M_{1:T}$, and frozen-tracker boxes $B_{1:T}$, DreamHMR predicts complete SMPL-X motion, body shape, and presence:

$$\left(\widehat{\boldsymbol{\phi}}_{1:T}, \widehat{\boldsymbol{\theta}}_{1:T}, \widehat{\mathbf{t}}_{1:T}, \widehat{\boldsymbol{\beta}}, \widehat{a}_{1:T}\right) = f(I_{1:T}, M_{1:T}, B_{1:T}), \tag{1}$$

where $\boldsymbol{\phi}_t$ is the root orientation, $\boldsymbol{\theta}_t$ contains the local rotations of the 21 body joints and 15 joints in each hand, and $\mathbf{t}_t$ is the camera-frame pelvis position. The shape vector $\boldsymbol{\beta} \in \mathbb{R}^{10}$ is shared across the window, and a_t denotes the target person's presence in the observed canvas.

As shown in Figure 1, DreamHMR uses complete geometry supervision to adapt the video backbone and locally supported refinement to recover details that are only visible in parts of the image. Global Body Representation Transfer renders a complete shaded human video whose colors encode a canonical surface map and whose shading reflects posed geometry. Local Joint Feature Enhancement then samples features around predicted joints, gates them by target-pixel support, and fuses the resulting corrections into four anatomical states.

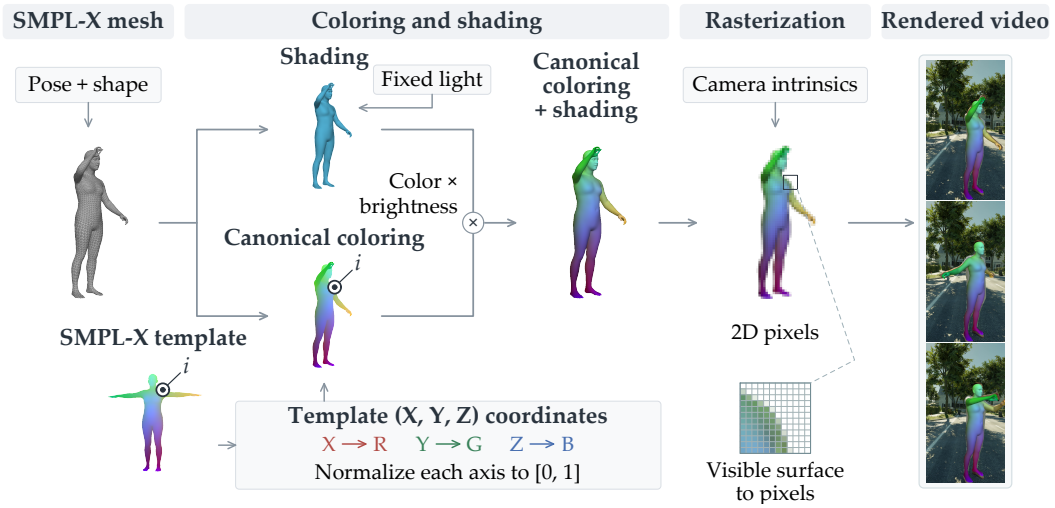

Figure 2: **From SMPL-X to rendered video.** Fixed template coordinates provide canonical colors; posed normals determine shading. Their product is rasterized and composited with source RGB across frames. Vertex colors persist across poses; a neutral shape and enlarged pixels aid illustration.

3.2 Global Body Representation Transfer

Global Body Representation Transfer injects complete projected human structure into the video representation over the full image canvas, including regions outside the observation mask.

Rendering complete human structure. We construct the target video from the annotated SMPL-X pose and shape. For each frame, we place the posed mesh's pelvis at $\mathbf{t}_t$ to obtain camera-frame vertices $\mathbf{V}_t$. The camera projection Π_t uses the input calibration and the same resized-and-padded canvas, aligning the rendered mesh with the RGB input.

We render a canonical surface map (Kulkarni et al., 2019): every visible pixel is colored by the canonical surface coordinates of its corresponding template point (Figure 2). The image-space encoding of these coordinates follows dense human pose estimation (Güler et al., 2018). Let $\mathbf{T}_i$ be vertex i of a neutral SMPL-X mesh in its template pose, with per-axis extrema $\mathbf{T}_{\min}$ and $\mathbf{T}_{\max}$. We normalize these template coordinates componentwise to obtain the R, G, and B channels:

$$\mathbf{C}_i = \frac{\mathbf{T}_i - \mathbf{T}_{\min}}{\mathbf{T}_{\max} - \mathbf{T}_{\min}} \in [0,1]^3. \tag{2}$$

The canonical surface map is computed once: vertex i retains $\mathbf{C}_i$ as pose and shape deform the mesh. The rendered color therefore identifies the same body surface location across poses, people, and frames, supplying dense correspondence supervision. This correspondence is defined on the template surface rather than by a point's image location. As the body moves, its projected location can change while the underlying canonical color assigned to that surface point remains the same.

Canonical coloring and shading define the appearance of the same posed mesh: template coordinates supply the surface colors, while current surface normals and a fixed light determine brightness. Their product combines stable surface correspondence with pose-dependent geometry. Rasterization projects the triangles onto the image, resolves visibility, and evaluates this appearance at covered pixels. Compositing with the source RGB yields the three-channel target I_t^{rend}; repeating the process across frames produces the supervision video. Rendering covers the full canvas independently of M_t, retaining projected human pixels outside the observed region; Appendix C details the renderer.

Learning from the rendered video. A frozen VAE encoder E maps the rendered video to the clean target latent $\mathbf{z}_0 = E(I_{1:T}^{\text{rend}})$ and input RGB to $\mathbf{z}_I = E(I_{1:T})$. The condition $\mathcal{C}$ comprises the RGB latent $\mathbf{z}_I$, observation masks, box tracks, and a fixed text context. Following the linear flow-matching path used by Wan (Team Wan et al., 2025), we interpolate between the target latent

and Gaussian noise at noise level $\tau \in [0, 1]$:

$$\mathbf{z}_\tau = (1 - \tau)\mathbf{z}_0 + \tau\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{3}$$

with $\tau = 0$ denoting the clean target latent. The target velocity is $\partial\mathbf{z}_\tau/\partial\tau = \boldsymbol{\epsilon} - \mathbf{z}_0$. We train the DiT velocity predictor $v(\mathbf{z}_\tau, \tau; \mathcal{C})$ with the conditional flow-matching objective

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{\mathbf{z}_0, \mathcal{C}, \tau, \boldsymbol{\epsilon}}\left[\left\|v(\mathbf{z}_\tau, \tau; \mathcal{C}) - (\boldsymbol{\epsilon} - \mathbf{z}_0)\right\|_2^2\right]. \tag{4}$$

The expectation is over paired target latents and conditions, noise levels, and Gaussian noise; in training, the squared error is normalized by the number of latent elements. The VAE and text encoder remain frozen. This stage adapts the DiT to predict complete projected human structure and provides the features used by the structured decoder.

3.3 Local Joint Feature Enhancement

The structured decoder combines global anatomical states with local joint descriptors: states summarize complete body-part context, while descriptors locate joints and retrieve nearby image evidence.

Part and Joint Feature Focus. Let $\mathbf{F}_k$ be the adapted DiT feature map at latent frame k, projected to the decoder dimension $d = 512$, with $K = 21$ latent frames. Four learned part queries and 52 learned joint queries extract part and joint information from shared features through cross-attention:

$$\mathbf{s}_{k,p} = \mathrm{CrossAttn}_{\mathrm{part}}(\mathbf{F}_k), \qquad \mathbf{h}_{k,j} = \mathrm{CrossAttn}_{\mathrm{joint}}(\mathbf{F}_k, \mathbf{s}_{k,p}). \tag{5}$$

Here $p \in \{1, \ldots, 4\}$ indexes the pelvis, remaining body, left hand, and right hand, while $j \in \{1, \ldots, 52\}$ indexes body and hand joints. Part-state vectors $\mathbf{s}_{k,p} \in \mathbb{R}^d$ represent the overall context of each anatomical region, while joint descriptors guide localization and local sampling. Assigning joint corrections to their corresponding parts structures the subsequent feature fusion. A box encoder supplies geometry, confidence, and tracking status as tokens injected at the corresponding times.

Joint Localization and Local Sampling. Within each latent frame, each descriptor predicts an image location and retrieves a local feature:

$$\widehat{\mathbf{x}}_j = \mathrm{MLP}_{\mathrm{loc}}(\mathbf{h}_j), \qquad \mathbf{f}_j^{\mathrm{local}} = \mathrm{LocalPool}(\mathbf{F}_k, \widehat{\mathbf{x}}_j). \tag{6}$$

LocalPool bilinearly samples a fixed 3×3 neighborhood around the predicted location and averages the valid samples. Available ground-truth joints supervise localization.

Evidence-Gated Local Update. An evidence head predicts target-pixel support for each joint and uses the resulting score to weight its local correction:

$$q_j = \sigma\big(\mathrm{MLP}([\mathbf{h}_j; \mathbf{f}_j^{\mathrm{local}}])\big), \qquad \mathbf{u}_j = q_j \mathbf{W}\mathbf{f}_j^{\mathrm{local}}, \tag{7}$$

where q_j is the evidence score, σ is the sigmoid, $[\cdot;\cdot]$ denotes concatenation, $\mathbf{W}$ projects the local feature, and $\mathbf{u}_j$ is the gated correction.

Evidence Supervision. We supervise the evidence score q_j with a binary target-pixel support label q_j^{GT}. The label is 1 when the ground-truth joint projection lies within both the observation region and the target person's visible instance mask, and 0 otherwise. We train the evidence head with

$$\mathcal{L}_{\mathrm{support}} = \mathrm{BCE}(q_j, q_j^{\mathrm{GT}}). \tag{8}$$

The balanced BCE averages the losses within three groups—supported joints inside the observation region, unsupported joints inside it, and joints outside it—and then equally averages the nonempty groups. Supervision uses available mask and projection annotations at the original frames corresponding to latent endpoints, with group averages computed over the training batch. Reconstruction gradients also train q_j through the gated local update.

Joint-to-Part Feature Fusion. Gated corrections update the corresponding anatomical state. Let $\mathcal{J}_p$ contain the joints assigned to part p. Cross-attention fuses their corrections into the part state:

$$\widetilde{\mathbf{s}}_{k,p} = \mathbf{s}_{k,p} + \mathrm{CrossAttn}\big(\mathbf{s}_{k,p}, \{\mathbf{u}_j\}_{j \in \mathcal{J}_p}\big), \tag{9}$$

where $\mathbf{s}_{k,p}$ serves as the query and $\{\mathbf{u}_j\}_{j \in \mathcal{J}_p}$ serves as keys and values. Prediction heads map the refined states to root orientation, local rotations, pelvis position, shape, and presence. Temporal recovery maps the K latent endpoints to all T frames, while $\boldsymbol{\beta}$ remains shared across the window.

3.4 Training and Inference

We first adapt the DiT with rendered-video flow matching, then jointly train the backbone and structured decoder for parameter recovery using

$$\mathcal{L}_{\text{recovery}} = \mathcal{L}_{\text{SMPLX}} + \lambda_{\text{support}}\mathcal{L}_{\text{support}}, \tag{10}$$

where $\mathcal{L}_{\text{SMPLX}}$ supervises pose, position, shape, joint localization, temporal velocity, and presence. Training mixes full, static, and moving observations; Appendix A details losses and freezing.

At inference, the adapted DiT processes the VAE-encoded RGB once at $\tau = 0$, conditioned on masks, box tracks, and fixed text context. The structured decoder maps its features to SMPL-X parameters and presence, following the same feature path used in parameter-recovery training.

4 Experiments

We evaluate DreamHMR on standard human mesh recovery benchmarks and complete body-and-hand recovery under occlusion across four video backbones. Ablations examine rendering supervision, local evidence mechanisms, and temporal context. Our code is available in the supplementary material.

4.1 Experimental Setup

We train on BEDLAM (Black et al., 2023), BEDLAM2 (Tesch et al., 2025), EgoBody (Zhang et al., 2022), ARCTIC (Fan et al., 2023), InterCap (Huang et al., 2022b), and IDEA400 (Lin et al., 2023a), using full and spatially masked 81-frame windows. Occlusion evaluation and ablations use separate holdouts of 90 BEDLAM and 30 EgoBody sequences each, disjoint from training and validation.

We compare PromptHMR-Vid (Wang et al., 2025), GENMO (Li et al., 2025), SMPLest-X (Yin et al., 2026), 3DB-DINOv3 (Yang et al., 2026), and Hand4Whole++ (Moon, 2026), and instantiate DreamHMR with Cosmos-2B (NVIDIA et al., 2025), LTX-2B (HaCohen et al., 2024), Wan-1.3B, and Wan-5B (Team Wan et al., 2025). Methods share observed RGB and target tracks; video methods use full windows, while image methods use track crops without temporal smoothing. We report MPJPE, Procrustes-aligned MPJPE (PA-MPJPE), and per-vertex error (PVE) on 3DPW test (von Marcard et al., 2018), EMDB-1 (Kaufmann et al., 2023), and RICH test (Huang et al., 2022a). Occlusion results report pelvis-relative body and wrist-relative hand MPJPE. Errors are in millimeters. Appendices A and B detail backbones, observation construction, alignment, and aggregation.

4.2 Evaluation on Widely Used HMR Benchmarks

Table 1: **HMR benchmark comparison.** Errors in mm; PA denotes PA-MPJPE. Methods share RGB, targets, joint regressors, alignment, and sequence averaging. Blue bold, underlined, and black bold values mark the best, second-best, and best external result per column, respectively.

Method	3DPW			EMDB-1			RICH		
	PA ↓	MPJPE ↓	PVE ↓	PA ↓	MPJPE ↓	PVE ↓	PA ↓	MPJPE ↓	PVE ↓
PromptHMR-Vid	39.7	59.8	67.8	39.8	68.5	78.7	37.8	57.8	65.3
GENMO	34.4	**54.2**	65.5	42.8	65.5	75.3	36.4	60.2	64.9
SMPLest-X	33.8	54.5	64.1	41.1	63.0	81.0	**28.9**	**52.5**	**58.7**
3DB-DINOv3	**33.5**	55.1	**63.2**	**38.5**	**61.2**	**72.9**	30.6	54.0	59.8
Hand4Whole++	37.5	57.0	69.6	45.6	70.2	81.4	35.3	62.0	67.2
DreamHMR / Cosmos-2B	<u>31.9</u>	53.6	62.9	<u>37.2</u>	<u>60.5</u>	72.9	30.5	53.6	<u>57.4</u>
DreamHMR / LTX-2B	33.8	53.8	63.5	38.5	62.8	73.1	31.2	53.9	58.5
DreamHMR / Wan-1.3B	32.4	<u>53.0</u>	<u>62.3</u>	37.9	60.9	<u>71.2</u>	<u>28.0</u>	<u>51.0</u>	58.0
DreamHMR / Wan-5B	**30.9**	**51.9**	**61.9**	**36.2**	**58.3**	**69.8**	**27.2**	**49.3**	**55.2**

DreamHMR attains the lowest error on every metric of the widely used human mesh recovery benchmarks. In Table 1, Wan-5B improves on every metric: MPJPE by 2.3 mm on 3DPW, 2.9 mm on

EMDB-1, and 3.2 mm on RICH, with PVE reductions of 1.3, 3.1, and 3.5 mm. All four DreamHMR backbones outperform the best external 3DPW body MPJPE result.

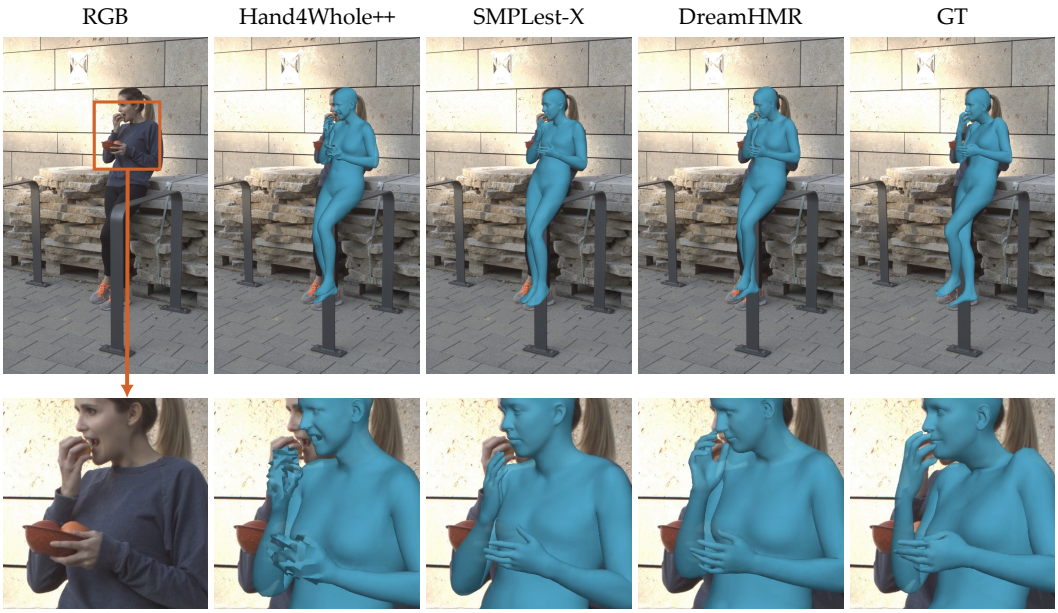

Figure 3: **Real-scene occlusion and hand detail.** Wan-1.3B on a RICH eating sequence with foreground railings. The box and arrow mark the hand-and-container region enlarged below. Methods share the crop and scale; GT denotes ground-truth meshes.

Figure 3 shows how this advantage appears in real scenes. In the boxed close-up, SMPLest-X (Yin et al., 2026) lays the fingers flat against the torso, leaving the bowl unheld, while DreamHMR keeps the fingers wrapped around the food and the bowl rim. See Appendix D.3 for temporal comparisons.

4.3 Body-and-Hand Reconstruction under Occlusion

Table 2: **Body-and-hand reconstruction under occlusion.** MPJPE (mm); Overall weights BEDLAM/EgoBody at 75%/25% and static/dynamic masks equally. Blue bold and underlined values mark the best and second-best results per column; – denotes no articulated-hand output.

Method	Overall		BEDLAM		EgoBody	
	Body ↓	Hands ↓	Body ↓	Hands ↓	Body ↓	Hands ↓
PromptHMR-Vid	121.2	–	120.6	–	122.8	–
GENMO	118.8	–	116.8	–	124.6	–
SMPLest-X	115.6	61.6	113.9	60.7	120.5	64.3
3DB-DINOv3	115.2	61.2	114.4	59.4	117.6	66.6
Hand4Whole++	118.3	63.1	117.0	62.4	122.2	65.2
DreamHMR / Cosmos-2B	<u>109.7</u>	60.7	<u>108.2</u>	59.8	114.2	63.4
DreamHMR / LTX-2B	113.2	62.1	111.8	61.5	117.4	63.8
DreamHMR / Wan-1.3B	111.3	<u>57.5</u>	110.5	<u>55.8</u>	<u>113.7</u>	<u>62.6</u>
DreamHMR / Wan-5B	**107.1**	**55.1**	**105.2**	**53.4**	**112.6**	**60.0**

DreamHMR improves complete body-and-hand reconstruction under occlusion across both evaluation sources. In Table 2, Wan-5B reaches 107.1/55.1 mm body/hand MPJPE, improving on the strongest external baseline, 3DB-DINOv3, by 8.1/6.1 mm. It also achieves the lowest error in every source-specific column, including EgoBody's real interaction videos.

The gains extend across video backbone families: all four improve on the strongest external overall body result, with Cosmos-2B ranking second, while both Wan models provide the strongest joint body-and-hand results. We use Wan-1.3B for the design analyses below.

4.4 Ablation Studies

We evaluate Wan-1.3B on the independent ablation holdout, reporting full-window body/hand MPJPE for Full (without additional masking) and Occlusion inputs. Rendering comparisons vary the first-stage target while fixing complete projected coverage and measure body PVE on shared SMPL-X vertices with pelvis alignment, no Procrustes fitting, and common GT-valid frames and vertices. Decoder and gating variants share adapted backbone initialization and reconstruction settings.

Table 3: **Rendering supervision.** Wan-1.3B errors (mm) over all 81 frames on the ablation holdout. Geometry targets share complete projected coverage. Body/Hands report MPJPE; body PVE uses shared SMPL-X vertices and pelvis alignment. RGB supervision is the reference.

Supervision target	Full			Occlusion		
	Body ↓	Hands ↓	PVE ↓	Body ↓	Hands ↓	PVE ↓
RGB	62.3	39.2	73.4	120.0	68.4	140.6
Shading	55.9	32.4	64.2	109.3	64.8	126.7
Canonical coloring	56.1	32.8	66.8	111.6	60.7	129.1
Shading + skeleton	54.7	31.3	62.5	110.5	64.2	127.4
Shading + canonical coloring	**53.8**	**30.7**	**61.3**	**106.3**	**57.8**	**121.8**



Figure 4: **Rendering targets in Table 3.** The same pose with RGB, shading, canonical coloring, a body-skeleton overlay, or blended shading and canonical colors.

Global Body Representation Transfer. Complete geometry supervision improves recovery under occlusion (Table 3). Replacing RGB supervision with canonical coloring and shading reduces body/hand MPJPE from 120.0/68.4 to 106.3/57.8 mm and body PVE from 140.6 to 121.8 mm. The combined target also yields the lowest Full-input errors: 53.8/30.7 mm MPJPE and 61.3 mm PVE.

Canonical coloring and shading provide complementary cues, outperforming either alone and shading with a skeleton. Figure 4 illustrates how canonical coloring identifies surface locations, shading expresses posed geometry, and the skeleton supplies sparse joint structure. The combined target thus grounds the video representation in dense correspondence and geometry.

Structured Decoding and Local Sampling. Table 4 ablates Local Joint Feature Enhancement (Section 3.3). The Structured Query Decoder replaces ordinary attention among flat task queries with part states, joint descriptors, and joint-to-part fusion (Equation 9). Combined with ungated local updates, it reduces occlusion body/hand MPJPE from 124.4/65.0 to 116.9/60.5 mm. Within this decoder, local sampling (Equation 6) reduces errors from 119.3/63.9 to 116.9/60.5 mm. Ungated updates fix the weight in Equation 7 to 1; the no-update variant removes this branch.

Evidence Gating and Supervision. Both evidence-based weighting and target-pixel support supervision improve local updates. With evidence supervision, gating reduces occlusion body/hand MPJPE from 113.1/59.6 to 106.3/57.8 mm. With gating, adding evidence supervision (Equation 8)

Table 4: **Local Joint Feature Enhancement ablation.** Wan-1.3B full-window body/hand MPJPE (mm) on the ablation holdout. Local updates are absent (None), weighted by one (Ungated), or by predicted support (Evidence-Gated). Evidence supervision uses target-pixel support labels.

Decoder	Local update	Evidence supervision	Full		Occlusion	
			Body ↓	Hands ↓	Body ↓	Hands ↓
Flat Query	None	No	55.8	32.2	124.4	65.0
Structured Query	None	No	55.1	31.8	119.3	63.9
Structured Query	Ungated	No	<u>54.0</u>	31.0	116.9	60.5
Structured Query	Ungated	Yes	54.1	<u>30.9</u>	113.1	59.6
Structured Query	Evidence-Gated	No	54.3	31.2	<u>110.7</u>	<u>59.0</u>
Structured Query	Evidence-Gated	Yes	**53.8**	**30.7**	**106.3**	**57.8**

reduces errors from 110.7/59.0 to 106.3/57.8 mm. The supervised, ungated variant trains the evidence head as an auxiliary task without using its scores for local updates; gating without evidence supervision sets $\lambda_{\text{support}} = 0$ (Equation 10), learning weights through reconstruction gradients. These two variants distinguish learning an evidence predictor as an auxiliary task from using its predictions to weight the local corrections. The full model combines both mechanisms, applying the predicted support to local features and supervising those predictions with the target-pixel labels. The full model also achieves the lowest Full-input errors, with larger gains under occlusion.

4.5 Temporal Context under Occlusion

Table 5: **Temporal context under occlusion.** Center-frame body/hand MPJPE (mm) on the ablation holdout with full-budget Wan-1.3B. ✓: original video; ✗: center RGB and mask duplicated over the window, retaining the original tracks and timestamps.

	Body MPJPE ↓		Hands MPJPE ↓	
Temporal Context	✓	✗	✓	✗
Full observation	56.44	59.70 (+3.26)	29.57	33.15 (+3.58)
Occlusion	108.92	121.63 (+12.71)	50.26	59.53 (+9.27)

DreamHMR uses cross-frame RGB context to improve reconstruction. Table 5 evaluates the full-budget Wan-1.3B checkpoint at the same center frames after replacing every RGB frame and observation mask with the center-frame input. This intervention increases body/hand MPJPE by 12.71/9.27 mm under occlusion, compared with 3.26/3.58 mm without additional masking. The larger effect under occlusion supports the use of neighboring observations for reconstruction. This diagnostic measures the response to an inference-time change in temporal context and input distribution. The original tracks and timestamps are retained in both conditions, so the repeated sequence removes changing RGB evidence while preserving the tracking input. We score only the center frame to keep the evaluation target fixed across conditions; these errors therefore should not be compared directly with the full-window averages in the other ablations. Appendix Figure 6 shows reconstructions through successive phases of a real action.

5 Conclusion

DreamHMR adapts video generation representations for complete body-and-hand recovery under occlusion. Complete human renderings transfer surface correspondence and posed geometry into the backbone, while evidence-gated local features refine contextual body-part estimates. Across four video backbones, DreamHMR improves reconstruction under occlusion and attains the lowest errors on the evaluated HMR benchmarks.

AI USE STATEMENT

In this work, we used generative AI tools to assist with method implementation, translation, and dataset cleaning and reformatting. We did not use LLMs to develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, support qualitative or thematic data analysis, or interpret results. Synthetic dataset generation, formulation of mathematical claims, and development or writing of mathematical proofs are not applicable to this work. Additionally, we used LLMs to create or modify scientific figures, edit software code, create research artifacts, improve manuscript readability, identify relevant literature, and search for information. The authors checked AI-assisted code and data processing for correctness, verified figures against the underlying results, and reviewed translations, language edits, and suggested references for accuracy. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégie, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-HMR: Multi-person whole-body human mesh recovery in a single shot. In *European Conference on Computer Vision*, 2024.

Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. SMPLer-X: Scaling up expressive human pose and shape estimation. In *Advances in Neural Information Processing Systems*, 2023.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. HumanPlus: Humanoid shadowing and imitation from humans. In *Proceedings of the Conference on Robot Learning*, 2025.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris M. Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2024.

Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022a.

Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *German Conference on Pattern Recognition*, 2022b.

Zixuan Huang, Xiang Li, Zhaoyang Lv, and James M. Rehg. How much 3D do video foundation models encode? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan Zarate, and Otmar Hilliges. EMDB: The electromagnetic database of global 3D human pose and shape in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Nilesh Kulkarni, Abhinav Gupta, David F. Fouhey, and Shubham Tulsiani. Canonical surface mapping via geometric cycle consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Jiefeng Li, Siyuan Bian, Chao Xu, Zhicun Chen, Lixin Yang, and Cewu Lu. HybrIK-X: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. GENMO: A GENeralist model for human MOtion. *arXiv preprint arXiv:2505.01425*, 2025.

Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-X: A large-scale 3D expressive whole-body human motion dataset. In *Advances in Neural Information Processing Systems*, 2023a.

Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3D whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023b.

Tianyu Luan, Zhongpai Gao, Luyuan Xie, Abhishek Sharma, Hao Ding, Benjamin Planche, Meng Zheng, Ange Lou, Terrence Chen, Junsong Yuan, and Ziyan Wu. Divide and fuse: Body part mesh recovery from partially visible human images. In *European Conference on Computer Vision*, 2024.

Zhengyi Luo, Ye Yuan, Tingwu Wang, et al. SONIC: Supersizing motion tracking for natural humanoid whole-body control. *Science Robotics*, 11(117), 2026. doi: 10.1126/scirobotics.aed4592.

Gyeongsik Moon. Enhancing hands in 3D whole-body pose estimation with conditional hands modulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Jisu Nam, Soowon Son, Dahyun Chung, Jiyoung Kim, Siyoon Jin, Junhwa Hur, and Seungryong Kim. Emergent temporal correspondences from video diffusion transformers. In *Advances in Neural Information Processing Systems*, 2025.

NVIDIA et al. World simulation with video foundation models for physical AI. *arXiv preprint arXiv:2511.00062*, 2025.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10975–10985, 2019.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. WiLoR: End-to-end 3D hand localization and reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Wenhao Shen, Wanqi Yin, Xiaofeng Yang, Cheng Chen, Chaoyue Song, Zhongang Cai, Lei Yang, Hao Wang, and Guosheng Lin. ADHMR: Aligning diffusion-based human mesh recovery via direct preference optimization. In *Proceedings of the International Conference on Machine Learning*, 2025.

Oriane Simeoni et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Nick Stracke, Stefan Andreas Baumann, Kolja Bauer, Frank Fundel, and Björn Ommer. CleanDIFT: Diffusion features without noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Qingping Sun, Yanjun Wang, Ailing Zeng, Wanqi Yin, Chen Wei, Wenjia Wang, Haiyi Mei, Chi Sing Leung, Ziwei Liu, Lei Yang, and Zhongang Cai. AiOS: All-in-one-stage expressive human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Advances in Neural Information Processing Systems*, 2023.

Team Wan et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Joachim Tesch, Giorgio Becherini, Prerana Achar, Anastasios Yiannakidis, Muhammed Kocabas, Priyanka Patel, and Michael J. Black. BEDLAM2.0: Synthetic humans and cameras in motion. In *Advances in Neural Information Processing Systems*, 2025.

Pedro Vélez, Luisa F. Polanía, Yi Yang, Chuhan Zhang, Rishabh Kabra, Anurag Arnab, and Mehdi S. M. Sajjadi. From image to video: An empirical study of diffusion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision (ECCV)*, 2018.

Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J. Black, and Muhammed Kocabas. PromptHMR: Promptable human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, Matt Feiszli, Jitendra Malik, Piotr Dollar, and Kris Kitani. SAM 3D Body: Robust full-body human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Wanqi Yin, Zhongang Cai, Ruisi Wang, Ailing Zeng, Chen Wei, Qingping Sun, Haiyi Mei, Yanjun Wang, Hui En Pang, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Atsushi Yamashita, Lei Yang, and Ziwei Liu. SMPLest-X: Ultimate scaling for expressive human pose and shape estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(2):1778–1794, 2026. doi: 10.1109/TPAMI.2025.3618174.

Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. EgoBody: Human body shape and motion of interacting people from head-mounted devices. In *European Conference on Computer Vision*, 2022.

Wenshuai Zhao, Yi Zhao, Joni Pajarinen, and Michael Muehlebach. Bi-level motion imitation for humanoid robots. In *Proceedings of the Conference on Robot Learning*, 2025.

Yixuan Zhu, Ao Li, Yansong Tang, Wenliang Zhao, Jie Zhou, and Jiwen Lu. DPMesh: Exploiting diffusion prior for occluded human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

A Implementation Details

Backbone interface. We use Cosmos2.5-Predict-2B (Cosmos-2B), LTX-Video-2B-0.9.6 (LTX-2B), Wan2.1-T2V-1.3B (Wan-1.3B), and Wan2.2-TI2V-5B (Wan-5B). Each video backbone encodes the observed RGB with its frozen VAE and supplies the final Transformer features before the native velocity head. A linear projection followed by LayerNorm maps these features to the shared 512-D decoder space. Table 6 summarizes the interfaces. LTX features are resampled along time by nearest-neighbor interpolation to match the 21 decoder endpoints; their spatial grid is retained. All backbones use the fixed text prompt *A full-body video of a person moving naturally.*

Table 6: Backbone interfaces for an 81-frame input. Endpoints refer to temporal positions before and after alignment to the shared decoder.

Backbone	Feature dimension	Native endpoints	Decoder endpoints
Wan-1.3B	1536	21	21
Wan-5B	3072	21	21
LTX-2B	2048	11	21
Cosmos-2B	2048	21	21

Structured decoder. The part and joint query branches each use two attention layers with eight heads, a 512-D hidden space, and a 2048-D feed-forward network. A two-layer MLP encodes box geometry, confidence, and tracking status at the corresponding endpoints. Each joint descriptor predicts a 2D location in normalized canvas coordinates. Local sampling retrieves a 3×3 neighborhood with offsets of one backbone feature cell, averages valid bilinear samples, and returns zero when the center or all samples are invalid. Valid samples lie within the observation region and the feature grid. The evidence head is a $1024 \to 256 \to 1$ MLP with GELU and a final sigmoid. Its score weights the local correction before joint-to-part fusion.

Parameter heads and temporal recovery. Part-specific MLPs predict 6D rotations, camera-frame pelvis position, and a presence logit at 21 endpoints. For consecutive endpoint timestamps s_k, s_{k+1}, the output vector $\mathbf{y}_k$ is interpolated at each original timestamp s as

$$\overline{\mathbf{y}}(s) = (1-\alpha)\mathbf{y}_k + \alpha \mathbf{y}_{k+1}, \qquad \alpha = \frac{s - s_k}{s_{k+1} - s_k}. \tag{11}$$

A residual temporal block, consisting of a kernel-3 depthwise convolution, GELU, and a pointwise convolution, refines the interpolated sequence. Its final projection is initialized to zero. The resulting 6D rotation vectors are orthonormalized into rotation matrices. A separate head predicts the shared 10D shape from part states averaged over parts and endpoints.

Optimization. The first stage trains the conditioned backbone and velocity head with the rendered-video flow-matching objective. The second stage jointly trains the backbone and structured decoder for parameter recovery while freezing the velocity head. VAE and text encoders remain frozen throughout. Table 7 gives the settings; training samples full, static, and dynamic observations in a 2:1:1 ratio and uses the terminal checkpoint.

Recovery objective. The parameter objective in Equation 10 expands as

$$\begin{aligned}\mathcal{L}_{\mathrm{SMPLX}} = \lambda_{\mathrm{rot}}\mathcal{L}_{\mathrm{rot}} + \lambda_{\mathrm{joint}}\mathcal{L}_{\mathrm{joint}} + \lambda_{\mathrm{root}}\mathcal{L}_{\mathrm{root}} + \lambda_{\mathrm{shape}}\mathcal{L}_{\mathrm{shape}} \\ + \lambda_{\mathrm{2D}}\mathcal{L}_{\mathrm{2D}} + \lambda_{\mathrm{vel}}\mathcal{L}_{\mathrm{vel}} + \lambda_{\mathrm{pres}}\mathcal{L}_{\mathrm{pres}}.\end{aligned} \tag{12}$$

We use hats for predictions and the superscript GT for annotations. For scalar errors e_i with availability mask m_i, and vectors of dimension d, define

$$\mathrm{Mean}_m(e) = \frac{\sum_{i:m_i=1} e_i}{\max(1, \sum_i m_i)}, \qquad \ell_1(\mathbf{u}, \mathbf{v}) = \frac{1}{d}\|\mathbf{u} - \mathbf{v}\|_1. \tag{13}$$

Rotation, relative joint, localization, and relative velocity losses average within the body, left hand, and right hand before equally averaging the supervised parts:

$$\mathrm{PartMean}_m(e) = \frac{1}{\max(1, |\mathcal{P}_m|)} \sum_{p \in \mathcal{P}_m} \mathrm{Mean}_{m_p}(e). \tag{14}$$

Table 7: Training settings. Task modules comprise the trainable conditioning projections and, during recovery, the structured decoder.

Setting	Geometry adaptation	Parameter recovery
Optimizer steps	10,000	47,500
Optimizer	AdamW	AdamW
Backbone / task learning rate	10^{-6} / 10^{-4}	10^{-6} / 10^{-4}
Weight decay	0.01	0.01
Learning-rate schedule	Constant	Constant
Global batch size	24	24
Gradient clipping	1.0	1.0
Hardware	8×H100 80GB	8×H100 80GB

Here m_p restricts the loss-specific mask to part p, and $\mathcal{P}_m$ contains the parts with valid supervision. Counts cover the full logical batch across devices and accumulation steps; empty terms are zero. Window indices are omitted below unless needed.

Rotation and geometry. Root orientation is expressed in the camera frame and local rotations in their SMPL-X parent frames. Rotation supervision uses the geodesic angle in radians:

$$d_{\mathrm{SO}(3)}(\widehat{R}, R^{\mathrm{GT}}) = \arccos\left[\mathrm{clip}\left(\frac{\mathrm{tr}(\widehat{R}^{\top} R^{\mathrm{GT}}) - 1}{2}, -1, 1\right)\right], \qquad \mathcal{L}_{\mathrm{rot}} = \mathrm{PartMean}_{m^{\mathrm{rot}}}\left(d_{\mathrm{SO}(3)}(\widehat{R}_{t,j}, R^{\mathrm{GT}}_{t,j})\right). \tag{15}$$

The body group includes the root and 21 body rotations; each hand contributes 15 rotations. Predicted joints $\widehat{\mathbf{J}}_{t,j}$ are obtained from predicted pose and shared shape through differentiable SMPL-X. Relative coordinates are

$$\mathbf{r}_{t,j} = \mathbf{J}_{t,j} - \mathbf{J}_{t,c(j)}, \qquad c(j) = \begin{cases} \text{pelvis}, & j \in \text{body}, \\ \text{left wrist}, & j \in \text{left hand}, \\ \text{right wrist}, & j \in \text{right hand}. \end{cases} \tag{16}$$

Prediction and ground truth each use their own reference joint while retaining camera-frame axes. Relative joint and camera-frame pelvis losses use mean absolute coordinate errors in meters:

$$\mathcal{L}_{\mathrm{joint}} = \mathrm{PartMean}_{m^{\mathrm{joint}}}\left(\ell_1(\widehat{\mathbf{r}}_{t,j}, \mathbf{r}^{\mathrm{GT}}_{t,j})\right), \qquad \mathcal{L}_{\mathrm{root}} = \mathrm{Mean}_{m^{\mathrm{root}}}\left(\ell_1(\widehat{\mathbf{t}}_t, \mathbf{t}^{\mathrm{GT}}_t)\right). \tag{17}$$

Relative joint supervision covers 21 non-root body joints and 15 joints per hand, requiring both the joint and its reference to be valid. Shared shape contributes once per valid window w:

$$\mathcal{L}_{\mathrm{shape}} = \mathrm{Mean}_{m^{\mathrm{shape}}}\left(\frac{1}{10}\|\widehat{\boldsymbol{\beta}}_w - \boldsymbol{\beta}^{\mathrm{GT}}_w\|_2^2\right). \tag{18}$$

Joint localization. For a canvas of width W and height H, the ground-truth pixel coordinate $(u_{k,j}, v_{k,j})$ at endpoint k is normalized to canvas coordinates. The descriptor location loss is

$$\mathbf{x}^{\mathrm{GT}}_{k,j} = \left(\frac{2u_{k,j}}{W-1} - 1, \frac{2v_{k,j}}{H-1} - 1\right), \qquad \mathcal{L}_{\mathrm{2D}} = \mathrm{PartMean}_{m^{\mathrm{2D}}}\left(\ell_1(\widehat{\mathbf{x}}_{k,j}, \mathbf{x}^{\mathrm{GT}}_{k,j})\right). \tag{19}$$

This term supervises all 52 descriptor locations at the 21 endpoints, with the pelvis included in the body group. Available labels outside the observation region also contribute.

Temporal velocity. For a position trajectory $\mathbf{y}_t$ and timestamps s_t in seconds, define the adjacent-frame velocity $D_t\mathbf{y} = (\mathbf{y}_{t+1} - \mathbf{y}_t)/(s_{t+1} - s_t)$. The root and relative joint components use mean absolute errors in meters per second:

$$\begin{aligned} \mathcal{L}_{\text{vel}}^{\text{root}} &= \text{Mean}_{m^{\text{vroot}}}\left(\ell_1(D_t\widehat{\mathbf{t}}, D_t\mathbf{t}^{\text{GT}})\right), \\ \mathcal{L}_{\text{vel}}^{\text{joint}} &= \text{PartMean}_{m^{\text{vjoint}}}\left(\ell_1(D_t\widehat{\mathbf{r}}_{\cdot,j}, D_t\mathbf{r}_{\cdot,j}^{\text{GT}})\right), \\ \mathcal{L}_{\text{vel}} &= \frac{\sum_{c\in\mathcal{V}} \mathcal{L}_{\text{vel}}^{c}}{\max(1, |\mathcal{V}|)}, \qquad \mathcal{V} \subseteq \{\text{root}, \text{joint}\}. \end{aligned} \tag{20}$$

Here $\mathcal{V}$ contains the components with valid supervision. Valid pairs are consecutive source frames of the same person with increasing timestamps, positive ground-truth presence at both ends, and valid position and reference labels. Geometry and presence use all 81 frames, giving up to 80 adjacent velocity pairs.

Presence and evidence support. For logit z and binary label y, define $\ell_{\text{BCE}}(z, y) = \log(1 + \exp z) - yz$. With presence logit z_t^{pres} and label a_t^{GT},

$$\mathcal{L}_{\text{pres}} = \text{Mean}_{m^{\text{pres}}}\left(\ell_{\text{BCE}}(z_t^{\text{pres}}, a_t^{\text{GT}})\right). \tag{21}$$

Support supervision balances the three groups defined in Section 3.3: supported joints inside the observation region, unsupported joints inside it, and joints outside it. Let $\mathcal{I}_g$ contain valid endpoint–joint annotations in group g, $\mathcal{G}_+$ contain the nonempty groups, and $q_{k,j} = \sigma(z_{k,j}^{\text{support}})$. Then

$$\mathcal{L}_{\text{support}} = \frac{1}{\max(1, |\mathcal{G}_+|)} \sum_{g\in\mathcal{G}_+} \frac{1}{|\mathcal{I}_g|} \sum_{(k,j)\in\mathcal{I}_g} \ell_{\text{BCE}}(z_{k,j}^{\text{support}}, q_{k,j}^{\text{GT}}). \tag{22}$$

Presence averages valid positive and negative frames; support averages nonempty groups over the 21 endpoints. Availability depends on annotation validity and identity, independently of predicted presence or evidence scores.

B Observation Construction and Evaluation

Training data. Table 8 summarizes the training mixture. Frame counts refer to retained person frames before observation augmentation. We sample source sequences equally within each dataset and then select an 81-frame window of the same person. Full, static, and dynamic observations share the camera, canvas, and person scale, using a fixed 848×480 canvas or its portrait counterpart.

Table 8: Training data and sampling proportions. Available annotations supervise the corresponding output fields.

Dataset	Person frames	Sampling proportion
BEDLAM	0.82M	18%
BEDLAM2	1.07M	22%
EgoBody	342K	25%
ARCTIC	287K	15%
InterCap	121K	10%
IDEA400	89K	10%

Observation regions. A reference box (c_x, c_y, w, h) is computed once per source sequence from the median complete SMPL-X projection box over reliable frames. Eight static templates comprise two horizontal regions of size $1.25w \times 0.55h$, two vertical regions of size $0.55w \times 1.25h$, and four local rectangles of size $0.70w \times 0.70h$. Their centers are displaced from the reference center by $0.25h$ vertically or $0.25w$ horizontally. Horizontal regions use the up/down offsets, vertical regions use left/right offsets, and local rectangles use all four directions. Eight dynamic templates move a local rectangle horizontally or vertically from these four initial positions, with a maximum displacement of $0.35w$ or $0.35h$ over the window. Regions are intersected with the physical canvas; pixels outside are set to neutral gray, $[0.5, 0.5, 0.5]$. Figure 5 illustrates the construction.



Figure 5: **Observation construction.** Two RICH frames (top to bottom) under full, static, and dynamic observations. Static regions stay fixed; dynamic regions move horizontally. Pixels outside the region are neutral gray. All panels share the display crop and scale.

Tracking and supervision. A frozen HumanArt YOLOX-X detector with ByteTrack provides the target tracks from each observed RGB sequence. The presence label is positive when the observation region contains at least 32 visible target-person pixels. A joint has positive target-pixel support when its ground-truth projection lies in both the observation region and the target's visible instance mask. Support supervision uses frames with reliable, aligned visible-instance masks; other samples contribute their available geometric and 2D labels. Complete body-and-hand geometry remains supervised wherever the corresponding annotations are available.

Evaluation sets. The occlusion evaluation and ablation holdouts each contain 90 BEDLAM and 30 EgoBody source sequences. Each sequence supplies two source windows; each window yields one full, eight static, and eight dynamic observations, giving 4,080 windows per holdout. The two holdouts are disjoint from each other, training, and validation, with identities, recording sessions, source motions, and scenes separated across splits. EgoBody examples use third-person videos with body-and-hand annotations. Evaluation on widely used benchmarks uses the official 3DPW test, EMDB-1, and RICH test sets with their original RGB inputs.

Shared inputs and output alignment. All compared methods receive the same observed RGB and target tracks. Video methods process complete windows; image methods process per-frame crops from the shared tracks. We use released inference models without test-time flipping or additional temporal smoothing. SMPLest-X uses the official H40 weights, and 3DB-DINOv3 includes its released hand decoder. Public body evaluation uses 14/24/24 joints for 3DPW/EMDB-1/RICH, respectively, with common joint regressors and frozen topology mappings for SMPL-X, MHR, and SMPL-X/MANO predictions. MPJPE and PVE use pelvis alignment; PA-MPJPE additionally fits a similarity transform. Under occlusion, body MPJPE uses 21 non-root joints relative to the pelvis; hand MPJPE uses 15 joints per hand relative to the wrist, with equal left/right weighting.

Sequence prediction and aggregation. For long videos, DreamHMR centers its 81-frame windows at frames $0, 40, 80, \ldots$ and the final frame, replicating boundary frames where necessary. Each

real frame takes the prediction from its nearest window center, with ties assigned to the earlier center. Public benchmark results average sequences equally. For occlusion evaluation, errors are averaged within each source sequence, then equally across sequences within each dataset; BEDLAM and EgoBody contribute 75% and 25%, respectively, and static/dynamic conditions contribute equally. The temporal-context diagnostic evaluates only the center frame: repeated inputs duplicate its RGB and observation mask while retaining the original tracks and timestamps.

C Rendering Details

This appendix expands the renderer summarized by Equation 2. For every annotated frame, it converts SMPL-X pose and shape parameters into one RGB image of a complete, camera-aligned human. The same procedure is applied to all frames to form the supervision video $I^{\mathrm{rend}}_{1:T}$. Canonical coloring assigns fixed template-coordinate colors to the posed mesh, and shading determines brightness from its current normals. Rasterization resolves visibility and evaluates their product at covered pixels; compositing with the source RGB produces the target frame. All rendering choices are fixed before training and shared across frames, subjects, and supervision variants.

Notation and camera-aligned mesh. The frame index is $t \in \{1, \dots, T\}$ and the SMPL-X mesh has N vertices with a fixed vertex ordering and triangle topology. We write $\boldsymbol{\phi}_t$ for the root orientation, $\boldsymbol{\theta}_t$ for the local body-and-hand pose, and $\boldsymbol{\beta}$ for the shape. Let $\mathrm{SMPLX}(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\beta}; \mathbf{s})$ return the mesh produced by SMPL-X when its native translation is $\mathbf{s}$, and let $\mathbf{J}_{0,t}$ be the pelvis returned by the same pose and shape with native translation set to zero. The annotated pelvis position in the camera frame is $\mathbf{t}_t$. Passing $\mathbf{t}_t - \mathbf{J}_{0,t}$ as the native translation therefore places the pelvis at $\mathbf{t}_t$:

$$\mathbf{V}_t = \left[\mathbf{V}_{t,1}, \dots, \mathbf{V}_{t,N}\right] = \mathrm{SMPLX}(\boldsymbol{\phi}_t, \boldsymbol{\theta}_t, \boldsymbol{\beta};\, \mathbf{t}_t - \mathbf{J}_{0,t})\,. \tag{23}$$

Here $\mathbf{V}_{t,i} \in \mathbb{R}^3$ is the camera-frame position of vertex i at frame t. The camera frame is centered at the optical center, with x pointing right, y pointing down, and z pointing forward; positions are measured in meters.

The rendered mesh and the source RGB frame use the same resized-and-padded canvas. If the source intrinsics are (f_x, f_y, c_x, c_y), a resize by s and a pixel offset (o_x, o_y) give canvas intrinsics

$$\bar{f}_x = s f_x, \qquad \bar{f}_y = s f_y, \qquad \bar{c}_x = s c_x + o_x, \qquad \bar{c}_y = s c_y + o_y. \tag{24}$$

A camera-frame point $\mathbf{X} = (X, Y, Z)$ with $Z > 0$ projects to

$$\pi_t(\mathbf{X}) = \left(\bar{f}_x X/Z + \bar{c}_x,\ \bar{f}_y Y/Z + \bar{c}_y\right). \tag{25}$$

The mesh and RGB input share the same canvas and camera calibration.

Canonical surface map. The canonical surface map assigns every template vertex an RGB value that encodes its canonical surface coordinates. Let $\mathbf{T}_i \in \mathbb{R}^3$ be vertex i of one fixed neutral SMPL-X template in its canonical pose. Define the componentwise template extrema

$$\mathbf{T}_{\min} = \min_i \mathbf{T}_i, \qquad \mathbf{T}_{\max} = \max_i \mathbf{T}_i,$$

where the minimum and maximum are taken independently along the template x, y, and z axes. We encode these three canonical surface coordinates in the RGB channels:

$$\mathbf{c}_i^{\mathrm{can}} = \frac{\mathbf{T}_i - \mathbf{T}_{\min}}{\mathbf{T}_{\max} - \mathbf{T}_{\min}} \in [0, 1]^3. \tag{26}$$

The denominator and all $\mathbf{c}_i^{\mathrm{can}}$ values are computed once from the fixed template. Every posed mesh keeps the template's vertex ordering and topology, so vertex i inherits the same canonical surface coordinates as pose and shape change its 3D position. The map is therefore shared across frames and identifies the same surface location in every pose.

Rasterization and visibility. For each covered pixel $\mathbf{u}$, the renderer computes screen-space barycentric weights $\lambda_{t,k}(\mathbf{u})$ for the projected triangles and selects the face with minimum interpolated camera depth, with $z_{t,i} = [\mathbf{V}_{t,i}]_z$,

$$d_{t,f}(\mathbf{u}) = \sum_{k=0}^{2} \lambda_{t,k}(\mathbf{u}) z_{t,i_k}. \tag{27}$$

Let Ω_t denote mesh coverage and (i_0, i_1, i_2) the selected face. The pixel's canonical color is

$$\mathbf{c}_t(\mathbf{u}) = \sum_{k=0}^{2} \lambda^{\star}_{t,k}(\mathbf{u}) \mathbf{c}^{\mathrm{can}}_{i_k}, \qquad \mathbf{u} \in \Omega_t. \tag{28}$$

Coverage is evaluated over the full canvas independently of the observation mask M_t.

Pose-dependent shading. The canonical surface map identifies a visible surface location, while shading describes its current orientation. The normal is recomputed from the posed mesh at every frame; for face f it is the normalized cross product of the two current edge vectors, and $\hat{\mathbf{n}}_t(\mathbf{u})$ denotes the visible-point normal. A fixed camera-frame light has direction $\hat{\mathbf{l}} \propto (0, -0.4, -1)$, and the scalar Lambertian factor is

$$\ell_t(\mathbf{u}) = a + d \max\Big(0, \hat{\mathbf{n}}_t(\mathbf{u}) \cdot \hat{\mathbf{l}}\Big), \qquad a = 0.4, \quad d = 0.6. \tag{29}$$

The max term removes back-facing light. Multiplying the rasterized canonical surface map by this scalar gives the shaded mesh RGB:

$$\mathbf{m}_t(\mathbf{u}) = \ell_t(\mathbf{u})\, \mathbf{c}_t(\mathbf{u}), \qquad \mathbf{u} \in \Omega_t. \tag{30}$$

The light is fixed in the camera frame, whereas $\hat{\mathbf{n}}_t(\mathbf{u})$ follows the posed geometry; brightness changes with pose while the canonical surface coordinates stay tied to the surface location.

Compositing the complete target. Let $I^{\mathrm{src}}_t(\mathbf{u})$ be the source RGB frame after the full, static, or dynamic observation construction. The rendered frame replaces source pixels only where the posed mesh is visible:

$$I^{\mathrm{rend}}_t(\mathbf{u}) = \begin{cases} \mathbf{m}_t(\mathbf{u}), & \mathbf{u} \in \Omega_t, \\ I^{\mathrm{src}}_t(\mathbf{u}), & \mathbf{u} \notin \Omega_t. \end{cases} \tag{31}$$

Mesh coverage is independent of M_t, so projected human pixels outside the observed region remain in the target whenever they lie on the physical canvas. Ordinary visibility and self-occlusion still apply. Repeating Equation 31 for $t = 1, \ldots, T$ produces $I^{\mathrm{rend}}_{1:T}$.

Supervision variants. The final target rasterizes the canonical surface coordinates from Equation 26 and applies the shading in Equation 29. The canonical coloring variant uses $\mathbf{c}_t(\mathbf{u})$ without the shading multiplier. The shading variant replaces every $\mathbf{c}^{\mathrm{can}}_i$ by the fixed color $(0.32, 0.70, 0.84)$ before shading. The shading + skeleton variant uses that same shaded mesh and overlays the projected 52-joint body-and-hand skeleton with fixed colors, line widths, and drawing settings. All variants share the same SMPL-X geometry, camera, canvas, source background, visibility resolution, and rasterization procedure.

D Additional Results

D.1 Effect of Backbone Initialization

Table 9: **Backbone initialization.** Body/hand MPJPE (mm) over all 81 frames on the ablation holdout after 10,000 task-training steps. Paired Wan runs vary DiT initialization; DINOv3-H+ is the image self-supervised reference.

		Full		Occlusion	
Backbone	Initialization	Body ↓	Hands ↓	Body ↓	Hands ↓
DINOv3-H+	Image SSL	53.92	33.40	116.47	65.36
Wan-1.3B	Random	56.44	34.20	126.34	70.13
Wan-1.3B	Video pretrained	52.96	31.21	112.81	61.85
Wan-5B	Random	55.78	33.65	124.22	68.46
Wan-5B	Video pretrained	52.11	30.68	108.87	55.49

Video generation pretraining provides a useful initialization for reconstruction under a fixed task-training budget. Table 9 compares pretrained and randomly initialized DiTs within the same Wan architecture after 10,000 task-training steps, with all 81 frames evaluated. The paired runs share task data, the structured readout, evidence gating, and optimization settings; the corresponding pretrained VAE and text encoder are retained in both runs. Pretrained initialization reduces body/hand MPJPE under occlusion by 13.53/8.28 mm at 1.3B and 15.35/12.97 mm at 5B.

DINOv3-H+ (Simeoni et al., 2025) provides a complementary image self-supervised representation reference under the same task-training settings. The pretrained Wan-1.3B configuration reduces body/hand MPJPE under occlusion by 3.66/3.51 mm relative to this reference. This comparison evaluates the two representation choices, which differ in architecture, pretraining, and temporal encoding.

D.2 Evidence Gating with Static and Dynamic Observations

Table 10: **Evidence gating across observation patterns.** Wan-1.3B body/hand MPJPE (mm) on the ablation holdout. Ungated uses unit local weights; Gated uses supervised evidence scores. Δ is Gated minus Ungated, computed from unrounded means.

	Body			Hands		
Observation	Ungated	Gated	Δ	Ungated	Gated	Δ
Static	115.7	**103.5**	−12.2	57.7	**56.1**	−1.6
Dynamic	118.1	**109.1**	−9.0	63.3	**59.5**	−3.8

Evidence gating improves reconstruction under both observation patterns (Table 10). On the Wan-1.3B ablation holdout, it reduces body MPJPE by 12.2 mm with static regions and 9.0 mm with dynamic regions. Hand MPJPE improves by 1.6 and 3.8 mm, respectively. These results complement the pooled ablations by showing the benefit of weighting local updates when the observation region is fixed or moves across the person.

D.3 Additional Qualitative Comparisons

Figure 6 follows a two-second bench-dip motion. DreamHMR keeps the hands gripping the bench edge and both feet inside the shoes, while both baselines leave the shoes uncovered.

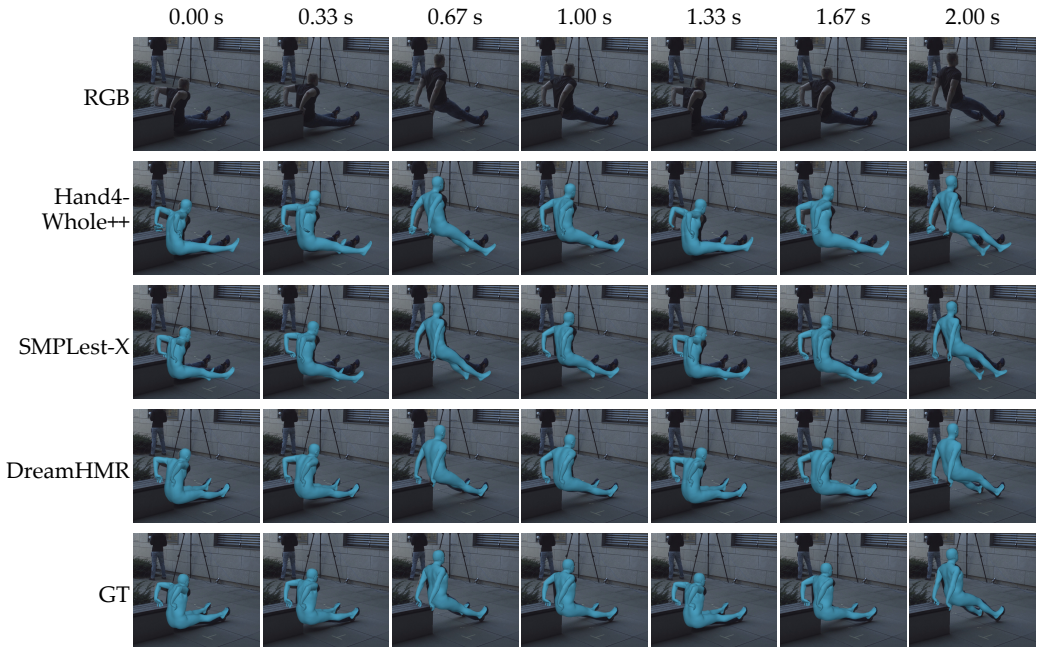

Figure 6: **Reconstruction over time.** Wan-1.3B on a RICH bench-dip sequence: seven frames at 0.33-second intervals span two seconds. All methods share a fixed crop and scale throughout the sequence. GT denotes ground-truth meshes.

Figure 7 compares all four DreamHMR configurations with Hand4Whole++ and SMPLest-X across eight videos, including bench support, walking, reaching, and foreground occlusions.



Figure 7: **Qualitative comparisons across scenes.** DreamHMR with four video backbones, compared with Hand4Whole++ and SMPLest-X. RGB and ground truth are shown for reference.

Figure 7 (continued)

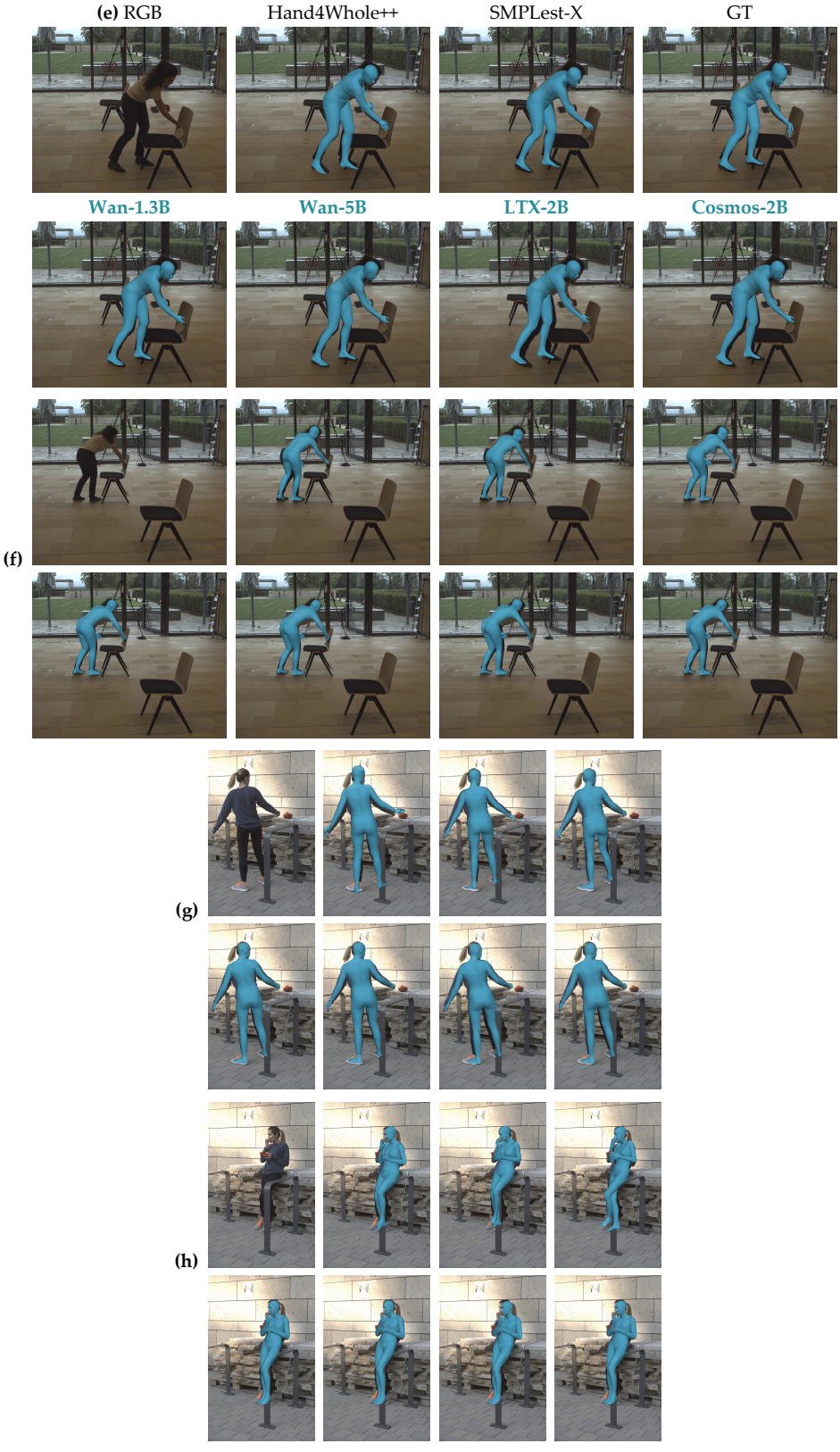

Figure 7 (continued)

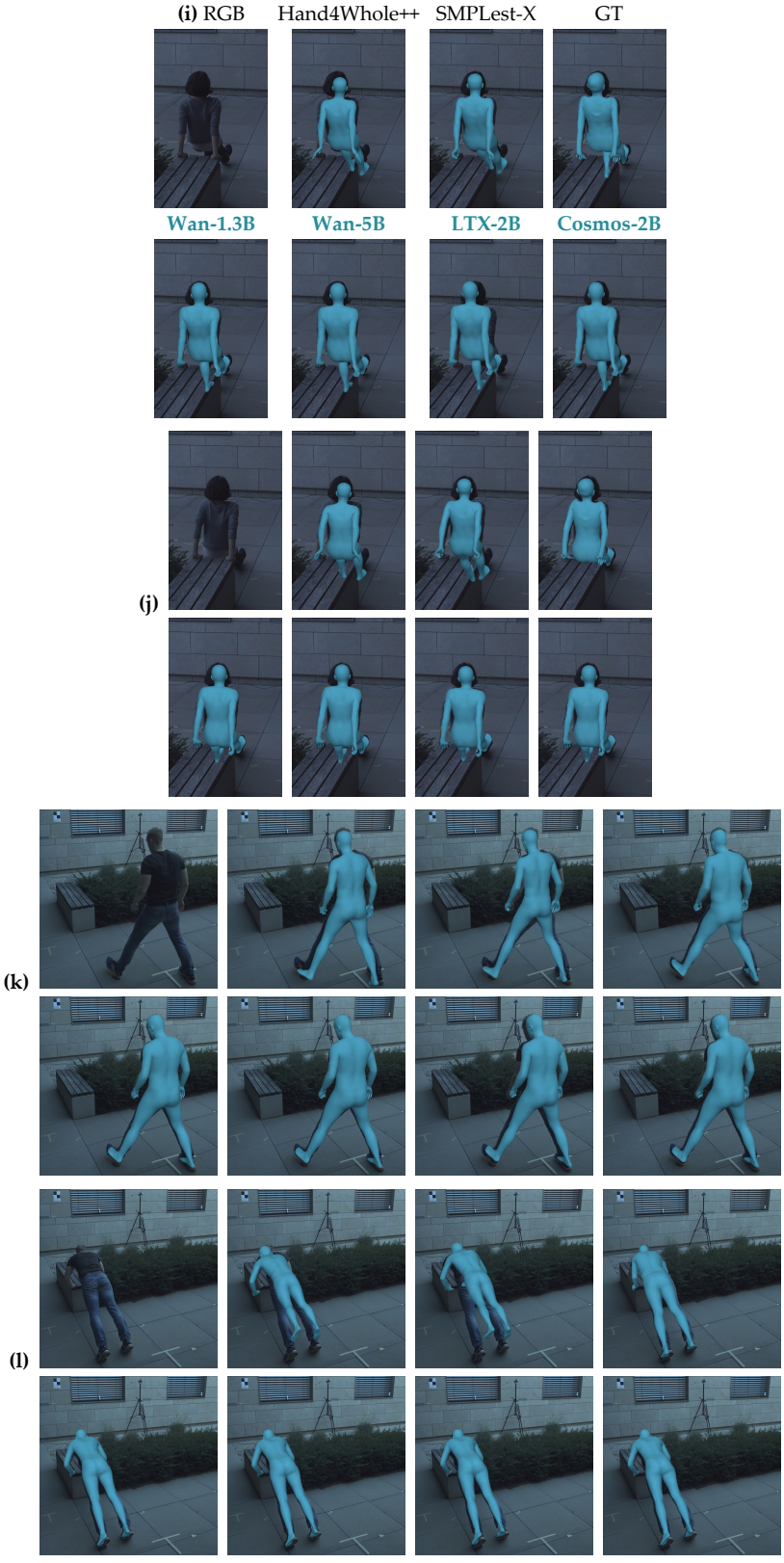

Figure 7 (continued)

