VELO3R: MOTION-GUIDED UNIFIED ONLINE 4D HUMAN-SCENE RECONSTRUCTION

Anonymous authors
Paper under double-blind review

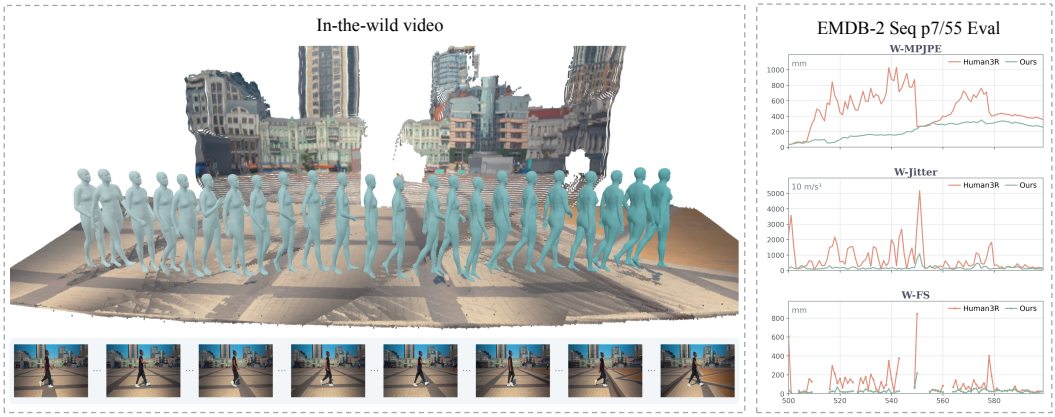

Figure 1: Overview of Velo3R on a walking sequence. Reconstructed human meshes and scene geometry are shown above the input frames; the right panels compare world-space MPJPE, jitter, and foot sliding (W-FS) with Human3R.

ABSTRACT

Online 4D human–scene reconstruction aims to recover human motion, camera trajectory, and scene geometry from a monocular video as it arrives, using only the current and preceding frames and expressing every estimate in a shared world coordinate system. Maintaining coherent motion is especially challenging online, as occlusions and ambiguities between body and camera movement must be handled without access to future frames. These ambiguities can lead to inconsistent frame-to-frame estimates, producing jitter and unreliable global trajectories. We seek an explicit link between successive body states to encourage temporal consistency while allowing new observations to correct the reconstruction. To this end, we introduce Velo3R, which uses explicit motion estimates to refine the latent human representations used for current-frame mesh prediction. Conditioned on temporal tokens and previously reconstructed body states, our conditional velocity field links joint positions, velocities, and accelerations through integration and differentiation. Inspired by physics-informed neural networks, we jointly supervise these quantities and feed field-derived motion features back into human representations before mesh regression. Experiments on EMDB-2 and RICH demonstrate improvements in global motion accuracy and temporal consistency. Compared with Human3R, Velo3R reduces W-MPJPE from 267.9 to 259.6 mm on EMDB-2 and from 184.9 to 170.6 mm on RICH. It also reduces world-coordinate jitter by 9.9% and 14.5%, respectively, with lower foot sliding on both datasets. These results highlight the value of explicit motion feedback for temporally coherent online 4D human–scene reconstruction. Our implementation will be made available upon acceptance.

1 INTRODUCTION

Online 4D human–scene reconstruction seeks to maintain a temporally coherent estimate of human motion, camera trajectory, and scene geometry as a monocular video stream unfolds. Such

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.

a representation describes both how people move and where their motion takes place, supporting applications in augmented reality and human–scene interaction. Recovering this representation is challenging: image motion combines body articulation, human translation, and camera movement, while occlusions leave parts of the body unobserved. An online human mesh recovery pipeline must interpret this incomplete evidence using only the current and preceding frames (Zhao et al., 2026). Uncertainty in these estimates can vary across frames, producing jitter and unreliable global trajectories. Maintaining coherent motion therefore requires reasoning about how the reconstructed body changes over time. Human3R (Chen et al., 2026) provides a unified online framework by combining human visual prompts with CUT3R's recurrent scene representation (Wang et al., 2025a). Its recurrent state carries evidence across frames, but the relation between successive body states remains implicit: *the representation does not specify the joint velocities and accelerations connecting them.* These observations motivate an explicit body-state transition whose motion estimates can guide the current reconstruction.

We therefore introduce **Velo3R**, which combines a conditional velocity field with motion feature feedback. For each tracked person, the field predicts velocities from a joint configuration, a temporal token, and time. The token gathers current visual evidence and recurrent context, while joints from the preceding reconstructed mesh anchor the transition. Integrating the field gives an updated joint state; differentiating the same field along its predicted motion gives acceleration. Inspired by physics-informed learning (Raissi et al., 2019), we jointly supervise these linked positions, velocities, and accelerations, with a soft bone-stretch penalty to discourage changes in skeletal segment lengths. Motion feature feedback connects this field to body reconstruction. We encode predicted velocities and integrated displacements, combine them with the temporal token, and use the resulting features to update the human representations before SMPL-X regression. This path lets the current mesh prediction use the estimated motion, while mesh reconstruction losses train the field through the same path. Joints from the final mesh then anchor the next transition, keeping the motion history tied to the reconstructed body.

Experiments show that Velo3R improves global joint accuracy over Human3R on EMDB-2 (Kaufmann et al., 2023) and RICH (Huang et al., 2022), while maintaining comparable local joint accuracy on 3DPW (von Marcard et al., 2018) and EMDB-1. Velo3R also improves temporal consistency, reducing world-coordinate jitter by 9.9% on EMDB-2 and 14.4% on RICH, together with lower acceleration error and foot sliding on both datasets. Ablations show the benefits of the conditional velocity field and motion feature feedback, and examine how joint-position and acceleration supervision affect the field.

Our contributions are:

- We formulate a conditional velocity field for online human motion modeling, explicitly representing body-state transitions and linking joint positions, velocities, and accelerations through integration, differentiation, and combined supervision.
- We introduce motion feature feedback that encodes predicted velocities, integrated displacements, and decoded temporal tokens to refine human representations before mesh regression, enabling current-frame reconstruction to use the estimated body-state transition.
- We demonstrate improved global motion accuracy and temporal consistency on EMDB-2 and RICH, and use progressive ablations to assess the contributions of velocity field supervision and motion feature feedback.

2 Related Work

Human Mesh and Motion Recovery. Parametric body models provide a common representation for human mesh recovery. SMPL (Loper et al., 2015) represents body shape and pose, SMPL-X (Pavlakos et al., 2019) adds articulated hands and facial expressions, and SKEL (Keller et al., 2023) combines the body surface with a biomechanical skeleton. HMR (Kanazawa et al., 2018), SPIN (Kolotouros et al., 2019), and HMR 2.0 (Goel et al., 2023) estimate SMPL parameters from images using direct regression, fitting-based supervision, and transformer architectures, respectively. Multi-HMR (Baradel et al., 2024) extends image-based recovery to multiple SMPL-X bodies, while HSMR (Xia et al., 2025) and SKEL-CF (Li et al., 2026a) recover SKEL parameters. Temporal methods reduce frame-level ambiguity by aggregating information across time. VIBE (Kocabas

et al., 2020) and TCMR (Choi et al., 2021) model video features, while TCMR uses bidirectional context from both past and future frames. Recovering global human motion from monocular video requires modeling the coupled effects of human and camera motion. GLAMR (Yuan et al., 2022) and SLAHMR (Ye et al., 2023) use motion priors and trajectory optimization to estimate global trajectories. WHAM (Shin et al., 2024) incorporates camera motion and contact-aware trajectory refinement, while TRAM (Wang et al., 2024) combines metric-scale camera estimation with local body motion. GVHMR (Shen et al., 2024) uses gravity-view coordinates to recover world-grounded motion.

Joint Human–Scene Reconstruction. Scene geometry provides context for human scale, placement, and contact, motivating joint estimation of the body and its surroundings. PhySIC (Yalandur Muralidhar et al., 2025) uses depth and contact estimates to optimize human–scene configurations from a single image. JOSH (Liu et al., 2026) jointly optimizes human motion, camera trajectories, and dense geometry from monocular video using contact constraints. Other acquisition settings supply richer observations: two moving RGB-D cameras in EmbodMocap (Wang et al., 2026) and multiple views in CHROMM (Kim et al., 2026). Feed-forward methods learn the alignment between human and scene predictions. JOSH3R (Liu et al., 2026) learns pairwise human transformations from MASt3R (Leroy et al., 2024) features using supervision from JOSH's optimized reconstructions. UniSH (Li et al., 2026b) combines scale and body-placement prediction with geometry distillation, while SHOW (Shi et al., 2026) uses mask prompts and geometry-aware SMPL-X decoding. For online reconstruction, human and scene estimates need to update as frames arrive. OnlineHMR (Zhao et al., 2026) combines a causal human estimation branch with incremental SLAM and regularizes joint velocities and accelerations during training. Human3R (Chen et al., 2026) uses visual prompts to combine CUT3R's recurrent scene state (Wang et al., 2025a) with Multi-HMR's human priors (Baradel et al., 2024). UniCon3R (Sur et al., 2026) uses geometry and contact history to update human latents before mesh regression.

Physics-Informed Motion Modeling. Prior work uses motion priors and physical constraints to encourage temporal coherence and physical plausibility in human reconstruction. HuMoR (Rempe et al., 2021) learns body-state transitions as a prior for motion fitting, while HTD-Refine (Wei et al., 2026) predicts velocities and accelerations to guide sequence optimization. QuaMo (Le et al., 2026) formulates motion through quaternion differential equations and a meta-PD controller with acceleration enhancement, using integration to update the reconstructed motion. PhysPT (Zhang et al., 2024) incorporates differentiable body dynamics, contact forces, and Euler–Lagrange constraints into motion reconstruction. Simulation-based approaches include PhysHMR (Feng et al., 2025), which learns a humanoid control policy from visual features through distillation and reinforcement learning, and RePHO (Huang et al., 2026), which refines human–object interactions through simulated tracking and updates to reference trajectories. Physics-informed neural networks couple observations with governing equations, using automatic differentiation to evaluate their residuals (Raissi et al., 2019; Luo et al., 2025). Drawing on this principle, we formulate a conditional velocity field whose integration and differentiation couple joint positions, velocities, and accelerations. We supervise these quantities jointly and feed the estimated motion back into human representations to guide online body reconstruction.

3 METHOD

Given a monocular video $\{I_t\}_{t=1}^{T}$, Velo3R reconstructs human meshes, camera poses, and scene geometry online from observations available up to frame t. Figure 2 summarizes the pipeline. For each tracked person, a shared recurrent decoder outputs human and temporal tokens. The temporal token conditions a velocity field that integrates the joint state from the preceding reconstruction to estimate the current state. The decoded temporal token, predicted velocities, and integrated displacements form motion features that refine the human representation before SMPL-X regression. The resulting mesh supplies the joint state for the next transition.

3.1 PRELIMINARIES

Following Human3R (Chen et al., 2026), we fuse CUT3R scene features (Wang et al., 2025a) and Multi-HMR human-prior features (Baradel et al., 2024) with an MLP to form a human prompt

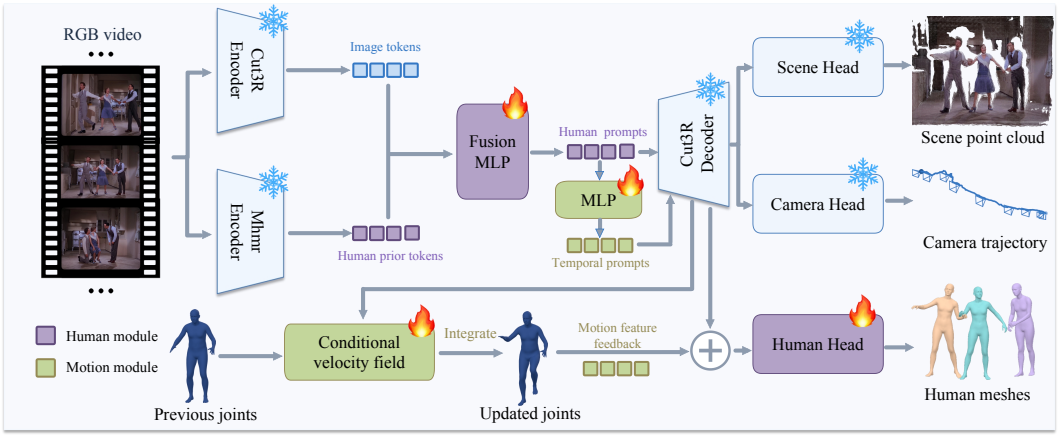

Figure 2: **Overview of Velo3R.** Frozen CUT3R and Multi-HMR features generate human prompts and detached temporal prompts. Within a shared frozen decoder, the human and temporal prompts interact with image tokens and recurrent scene memory. The decoder's temporal outputs condition a conditional velocity field for joint integration. The temporal tokens and the field outputs, including predicted velocities and integrated displacements, jointly form motion feedback that refines the human latents. The human head then uses the resulting refined features for SMPL-X regression.

$\mathbf{q}_t^{\text{human}}$ for each detected person. A temporal prompt is initialized as $\mathbf{q}_t^{\text{temp}} = \text{MLP}(\text{sg}(\mathbf{q}_t^{\text{human}}))$, where sg denotes stop-gradient. Both prompts share the same image location and interact with image tokens and persistent scene memory in the recurrent decoder, producing the human token $\mathbf{h}_t$ and temporal token $\mathbf{z}_t$. We denote the associated human-prior feature by $\mathbf{p}_t$.

For each tracked person, let $\mathbf{X}_t \in \mathbb{R}^{J\times 3}$ denote the camera-space positions of the SMPL-X body joints ($J = 22$) in the reconstructed body at frame t. The sequence $\{\mathbf{X}_t\}_{t=1}^{T}$ records the person's joint states over the video. For online inference, the motion state at frame t is initialized with a detached copy of the final-mesh joints reconstructed at the previous time step, $\mathbf{X}_{t-1}^{\text{history}} = \text{sg}(\mathbf{X}_{t-1})$. If τ_t denotes the timestamp of frame t, then a valid transition has $\Delta_t = \tau_t - \tau_{t-1} > 0$, measured in seconds. Joint states are in meters in each frame's camera coordinates, so their changes reflect both body motion and camera motion. Predicted camera poses transform reconstructed meshes into the shared world coordinate system.

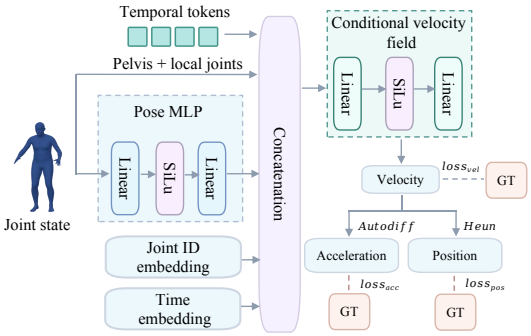

Figure 3: **Conditional velocity field.** Given a joint state and temporal features, the field predicts per-joint velocities. We integrate the velocity field to obtain joint positions and differentiate it along the joint trajectory to obtain accelerations.

3.2 Conditional Velocity Field

State and visual conditioning. We model each body-state transition with a conditional velocity field $f_\phi(\mathbf{X}, \mathbf{z}_t, s, \Delta_t) \in \mathbb{R}^{J\times 3}$. Fig. 3 summarizes its architecture and supervision. For joint j, the network predicts

$$\widetilde{\mathbf{v}}_j = F_\phi(\mathbf{z}_t \oplus \mathbf{e}_j \oplus \mathbf{x}_0 \oplus (\mathbf{x}_j - \mathbf{x}_0) \oplus E_{\text{pose}}(\text{Flatten}(\mathbf{X} - \mathbf{x}_0)) \oplus E_s(s) \oplus E_\Delta(\Delta_t)), \quad (1)$$

where $\oplus$ denotes concatenation along the channel axis, and $\mathbf{z}_t$ is the recurrent decoder's output for the temporal prompt, combining current visual evidence with recurrent context. The learned embedding $\mathbf{e}_j$ identifies joint j. The pelvis position $\mathbf{x}_0$ retains the body's camera-space location, and $\mathbf{x}_j - \mathbf{x}_0$ gives joint j's position relative to the pelvis. E_{pose} encodes the full pelvis-relative joint configuration $\mathbf{X} - \mathbf{x}_0$, where $\mathbf{X}$ contains the J joint positions input to the velocity field. The operator Flatten reshapes this $J \times 3$ matrix into a $3J$-dimensional vector. The resulting whole-body feature couples each joint's velocity to the overall configuration. E_s and E_Δ separately embed the

query time s and the elapsed time between frames Δ_t, both in seconds, with $s \in [0, \Delta_t]$. These three embedding functions are learned MLPs. The field predicts standardized velocities $\widetilde{\mathbf{v}}$, which we convert to physical velocities in meters per second as $f_\phi = \boldsymbol{\mu}_v + \boldsymbol{\sigma}_v \odot \widetilde{\mathbf{v}}$. We compute $\boldsymbol{\mu}_v$ and $\boldsymbol{\sigma}_v$ as the mean and standard deviation of ground-truth velocities for each joint and coordinate. These velocities are obtained by finite differences of annotated camera-space joint positions in the training data; the statistics remain fixed during training and inference.

Integrating the body state. For each valid transition, we initialize the joint state $\mathbf{X}_0 = \mathbf{X}_{t-1}^{\text{history}}$ with the detached joint positions reconstructed in the preceding frame. Let $\mathbf{X}(s)$ denote the joint positions obtained by integrating the conditional velocity field from $\mathbf{X}_0$ for an elapsed time s, where s ranges from 0 to the frame interval Δ_t. The evolution of these positions is governed by the following differential equation and initial condition:

$$\frac{d\mathbf{X}(s)}{ds} = f_\phi(\mathbf{X}(s), \mathbf{z}_t, s, \Delta_t), \qquad \mathbf{X}(0) = \mathbf{X}_0. \tag{2}$$

The conditional velocity field f_ϕ predicts the instantaneous joint velocities at the intermediate joint positions $\mathbf{X}(s)$. During integration, the conditioning token $\mathbf{z}_t$ and the total interval Δ_t remain fixed, while the joint positions and elapsed time vary. To estimate the joint positions at the current frame, we integrate the conditional velocity field over the frame interval Δ_t using a single step of Heun's method (Hairer et al., 1993):

$$\begin{aligned} \mathbf{v}_0 &= f_\phi(\mathbf{X}_0, \mathbf{z}_t, 0, \Delta_t), \qquad \mathbf{X}_{\mathrm{E}} = \mathbf{X}_0 + \Delta_t \mathbf{v}_0, \\ \mathbf{v}_1 &= f_\phi(\mathbf{X}_{\mathrm{E}}, \mathbf{z}_t, \Delta_t, \Delta_t), \qquad \mathbf{v}_t = \tfrac{1}{2}(\mathbf{v}_0 + \mathbf{v}_1), \\ \mathbf{X}_t^{\text{field}} &= \mathbf{X}_0 + \Delta_t \mathbf{v}_t. \end{aligned} \tag{3}$$

The same velocity-field network predicts $\mathbf{v}_0$ at the initial joint positions and $\mathbf{v}_1$ at the provisional joint positions $\mathbf{X}_{\mathrm{E}}$. Multiplying their average velocity $\mathbf{v}_t$ by the frame interval gives the joint displacement. Adding this displacement to the initial joint positions $\mathbf{X}_0$ yields the estimated current-frame joint positions $\mathbf{X}_t^{\text{field}}$. We supervise these positions using ground-truth joints and encode the displacement as part of the motion features for feedback.

Deriving acceleration from the velocity field. To compute joint accelerations by differentiating the velocity field, we use SiLU activations in the velocity-field network and its embedding MLPs to ensure smooth derivatives with respect to joint positions and time. Along the joint trajectory in Eq. 2, acceleration is the rate of change of the predicted velocity:

$$\mathbf{v}(s) = f_\phi(\mathbf{X}(s), \mathbf{z}_t, s, \Delta_t), \qquad \mathbf{a}(s) = \frac{d\mathbf{v}(s)}{ds}. \tag{4}$$

Here $\mathbf{v}(s)$ is the instantaneous joint velocity and $\mathbf{a}(s)$ is the corresponding acceleration. The total derivative accounts for both the direct effect of time on the predicted velocities and the changes in joint positions along the trajectory, while $\mathbf{z}_t$ and Δ_t remain fixed. We compute the current-frame acceleration $\mathbf{a}_t^{\text{field}}$ at $s = \Delta_t$, using the estimated joint positions $\mathbf{X}_t^{\text{field}}$ from Eq. 3. We compute these accelerations using PyTorch's automatic differentiation.

3.3 Motion Feature Feedback

The field estimates a joint-state transition, while the body regressor produces the final parametric mesh. Feature feedback connects these predictions: it allows the estimated transition to refine body features and allows mesh losses to supervise the field through this path. We use an MLP to encode the two standardized velocity estimates, the joint displacement, and an MLP embedding of the frame interval. Joint arrays are flattened before concatenation. The resulting motion feature $\mathbf{r}_t$ updates the temporal token $\mathbf{z}_t$ as follows:

$$\begin{aligned} \mathbf{r}_t &= \mathrm{MLP}\big(\widetilde{\mathbf{v}}_0 \oplus \widetilde{\mathbf{v}}_1 \oplus (\mathbf{X}_t^{\text{field}} - \mathbf{X}_0) \oplus \mathrm{MLP}(\Delta_t)\big), \\ \mathbf{z}_t' &= \mathbf{z}_t + m_t \tanh(\mathbf{r}_t). \end{aligned} \tag{5}$$

Here m_t is a binary mask that controls whether motion feedback is applied. It equals 1 when the current person has a valid identity match in the preceding frame and $\Delta_t > 0$, and 0 otherwise.

Another MLP takes the updated temporal token and the human-prior feature as input and predicts additive updates to the human token and human-prior feature:

$$\mathbf{h}'_t \oplus \mathbf{p}'_t = (\mathbf{h}_t \oplus \mathbf{p}_t) + \mathrm{MLP}(\mathbf{z}'_t \oplus \mathbf{p}_t). \tag{6}$$

The human head regresses SMPL-X (Pavlakos et al., 2019) pose, shape, and expression from $\mathbf{h}'_t \oplus \mathbf{p}'_t$, and camera-space translation from $\mathbf{h}'_t$. Let $\mathcal{J}$ extract the first 22 body joints from the decoded mesh. We update the motion history as

$$\mathbf{X}_t^{\mathrm{history}} = \mathrm{sg}(\mathcal{J}(\mathcal{M}_t)), \tag{7}$$

where $\mathcal{M}_t$ is the final camera-space mesh. We store the reconstructed joints as the starting positions for the next frame and treat them as fixed inputs. Within each frame, the human reconstruction loss also trains the velocity field through motion feature feedback.

3.4 Training Strategy

Following the unified online method Human3R (Chen et al., 2026), we train Velo3R on the synthetic BEDLAM dataset (Black et al., 2023) using scene, camera, and SMPL-X annotations. We retain RGB reconstruction, confidence-aware pointmap regression, and camera-pose supervision in $\mathcal{L}_{\mathrm{recon}}$. Human reconstruction is supervised by $\mathcal{L}_{\mathrm{human}}$, which applies L1 losses to SMPL-X parameters, 3D joints and vertices, and their image reprojections. Our temporal loss $\mathcal{L}_{\mathrm{temporal}}$ supervises the velocity field through integrated joint positions, interval-average velocities, and endpoint accelerations, with a bone-stretch penalty to discourage changes in skeletal segment lengths. We jointly train the human prompt and reconstruction head together with the temporal prompt, velocity field, and feedback networks, while keeping the pretrained scene and human-prior modules frozen. The full objective, supervision targets, loss weights, and optimization details are provided in Section C of Sup.Mat.

4 Experiments

We evaluate Velo3R in terms of global human motion accuracy, local mesh reconstruction, and temporal consistency. Quantitative comparisons measure reconstruction accuracy and motion quality, while qualitative results visualize the recovered humans and scenes in a shared coordinate system. Controlled ablations further examine the contributions of the conditional velocity field and motion feature feedback, followed by an analysis of the field supervision objectives.

4.1 Datasets and Evaluation Metrics

We evaluate global motion and temporal consistency on EMDB-2 (25 sequences; Kaufmann et al., 2023) and RICH (191 static-camera sequences; Huang et al., 2022). We report W-MPJPE, WA-MPJPE, and RTE for global motion, and W-Jitter (predicted jerk), W-Accel (acceleration error), and W-FS (foot sliding) for temporal consistency in world coordinates. Local mesh reconstruction is evaluated on 3DPW (von Marcard et al., 2018) and EMDB-1 (Kaufmann et al., 2023) using pelvis-relative MPJPE and PVE, and Procrustes-aligned PA-MPJPE. Lower is better for all metrics. Further evaluation details are provided in Section C of Sup.Mat.

4.2 Comparison with Existing Methods

Global human motion. Velo3R achieves W-MPJPEs of 259.6 mm on EMDB-2 and 170.6 mm on RICH, outperforming Human3R by 3.1% and 7.7%, respectively (Table 1). WA-MPJPE decreases to 107.4 and 104.6 mm, while RTE reaches 2.1% and 3.1%, respectively, improving over Human3R on both datasets. Among online methods, Velo3R achieves the lowest RTE on both benchmarks and leads across all three metrics on RICH, demonstrating accurate world-space joint localization and trajectory recovery.

Local human mesh reconstruction. On 3DPW, Velo3R improves all three local metrics over Human3R, with PA-MPJPE and MPJPE of 43.9 and 70.5 mm, respectively. On EMDB-1, Velo3R also attains the lowest PVE among online methods, at 85.8 mm (Table 2). Together, these results complement the global-motion gains with competitive local mesh accuracy.

Table 1: Global human motion on EMDB-2 and RICH.

Method	EMDB-2 (Kaufmann et al., 2023)			RICH (Huang et al., 2022)		
	WA-MPJPE ↓	W-MPJPE ↓	RTE ↓	WA-MPJPE ↓	W-MPJPE ↓	RTE ↓
Offline						
GVHMR (Shen et al., 2024)	111.0	276.5	2.0	**78.8**	**126.3**	2.4
TRAM (Wang et al., 2024)	76.4	222.4	1.4	127.8	238.0	6.0
JOSH$_3$ (Liu et al., 2026)	**68.9**	**174.7**	**1.3**	–	–	–
UniSH (Li et al., 2026b)	118.5	270.1	5.8	118.1	183.2	4.8
SHOW (Shi et al., 2026)	109.1	262.3	2.1	107.3	172.7	**2.2**
CHROMM (Kim et al., 2026)	102.6	255.0	1.7	87.5	138.3	3.3
EmbodMocap (Wang et al., 2026)	82.2	220.7	1.7	–	–	–
Online						
WHAM (Shin et al., 2024)	135.6	354.8	6.0	108.4	196.1	4.5
JOSH3R (Liu et al., 2026)	220.0	661.7	13.1	–	–	–
Human3R (Chen et al., 2026)	112.2	267.9	2.2	110.0	184.9	3.3
OnlineHMR (Zhao et al., 2026)	**93.5**	310.4	2.2	–	–	–
UniCon3R (Sur et al., 2026)	113.7	285.6	2.3	–	–	–
Velo3R	107.4	**259.6**	**2.1**	**104.6**	**170.6**	**3.1**

Table 2: Local mesh accuracy of global human reconstruction methods on 3DPW and EMDB-1.

Method	3DPW (von Marcard et al., 2018)			EMDB-1 (Kaufmann et al., 2023)		
	PA-MPJPE ↓	MPJPE ↓	PVE ↓	PA-MPJPE ↓	MPJPE ↓	PVE ↓
Offline						
GVHMR (Shen et al., 2024)	37.0	**56.6**	68.7	44.5	74.2	85.9
TRAM (Wang et al., 2024)	35.6	59.3	69.6	45.7	74.4	86.6
PromptHMR-Vid (Wang et al., 2025b)	**35.5**	56.9	**67.3**	**40.1**	**68.1**	**79.2**
UniSH (Li et al., 2026b)	48.8	75.6	88.8	86.6	112.9	140.0
SHOW (Shi et al., 2026)	41.0	67.7	78.7	45.8	75.7	88.8
Online						
Human3R (Chen et al., 2026)	44.1	71.2	84.9	48.5	**73.9**	86.0
OnlineHMR (Zhao et al., 2026)	**43.7**	**69.9**	**83.7**	**46.0**	74.0	86.1
UniCon3R (Sur et al., 2026)	–	72.5	86.8	–	–	–
Velo3R	43.9	70.5	84.4	51.4	76.4	**85.8**

Temporal consistency. Velo3R improves all three world-space temporal metrics over Human3R on both datasets, with reductions summarized in Table 3: on EMDB-2, W-Jitter, W-Accel, and W-FS decrease by 9.9%, 8.5%, and 6.6%, respectively; on RICH, they decrease by 14.5%, 14.0%, and 10.8%. Lower jitter and acceleration error indicate smoother motion that better matches the annotated dynamics, while lower W-FS reflects reduced foot sliding without explicit foot-contact supervision.

Table 3: Temporal consistency.

Method	W-Jitter	W-Accel	W-FS
EMDB-2 (Kaufmann et al., 2023)			
Human3R	225.1	46.9	31.6
Velo3R	**202.8**	**42.9**	**29.5**
RICH (Huang et al., 2022)			
Human3R	170.9	32.1	26.0
Velo3R	**146.2**	**27.6**	**23.2**

Qualitative human–scene reconstruction. Compared with Human3R, Velo3R more closely matches the ground-truth body positions and orientations on EMDB-2 (Fig. 4). Fig. 5 shows Velo3R's qualitative human–scene reconstructions, including human meshes, scene point clouds, and camera poses in a shared coordinate system. The examples span walking, bending, raised-leg poses, and stair traversal, illustrating articulated human motion within its scene context. These visual results are consistent with the improvements in global joint accuracy and temporal consistency reported in Tables 1 and 3. For more visual examples, see Section B of Sup.Mat.

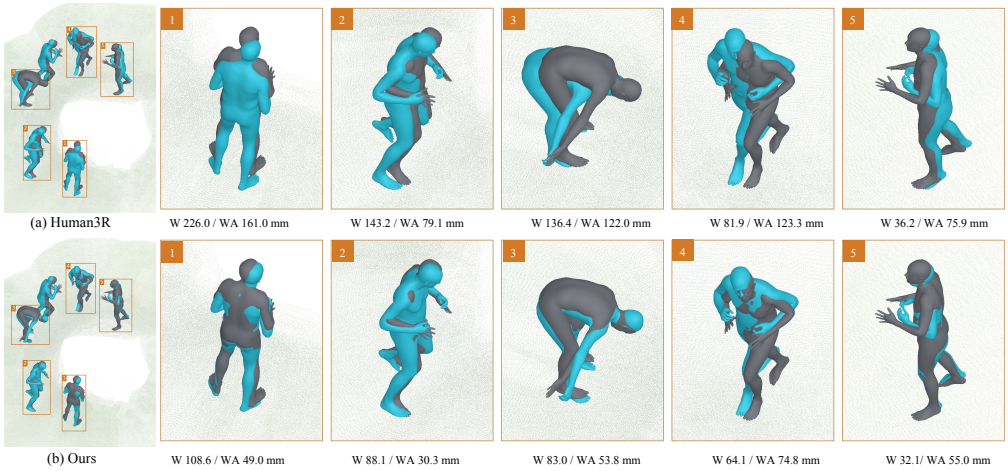

Figure 4: **World-space reconstruction on EMDB-2.** Human3R (top) and Velo3R (bottom) predictions are shown in cyan against gray ground truth. We report the W (W-MPJPE) and WA (WA-MPJPE) errors for each of the five examples.

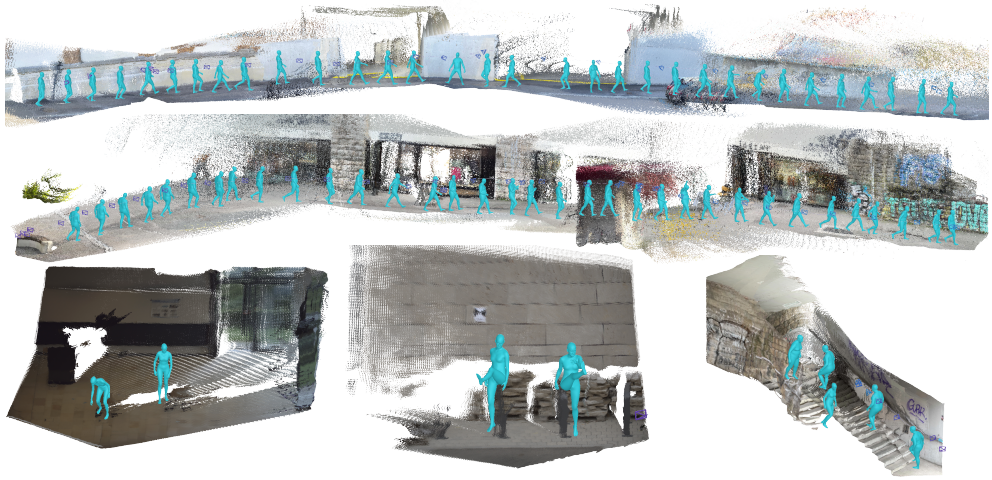

Figure 5: **Qualitative 4D human–scene reconstruction.** Velo3R jointly reconstructs human motion, scene geometry, and camera poses online from monocular video.

4.3 Ablation Studies and Analysis

We conduct two controlled ablation studies on EMDB-2 to evaluate the two main components of Velo3R: the conditional velocity field and motion feature feedback. In each study, we vary only the component of interest while keeping the remaining model components and training settings fixed. For motion modeling, Table 4 compares a baseline without explicit motion modeling, direct motion regression, and our conditional velocity field. For motion feedback, it compares no feedback, loss-based feedback, and our motion feature feedback while retaining the same conditional field and field supervision. Table 5 further analyzes the supervision applied to the conditional velocity field. Further ablation analyses and results are provided in Section A of Sup.Mat.

Effect of the conditional velocity field. For the motion-modeling comparison in Table 4, we use Human3R as the baseline and evaluate its temporal dynamics using the released checkpoints. Without explicit motion modeling, it obtains a W-MPJPE of 267.9 mm. Direct regression lowers W-MPJPE to 262.9 mm, confirming the benefit of introducing explicit motion cues. Our conditional velocity field further reduces this error to 259.6 mm, improving upon direct regression by 3.3 mm. We attribute this additional gain to its state-conditioned and dynamically coupled representation. Unlike direct regression, which estimates velocity and acceleration independently from the temporal token,

Table 4: Conditional-velocity-field and motion-feature-feedback ablations on EMDB-2.

Variant	WA-MPJPE	W-MPJPE	W-Jitter	W-Accel	W-FS
Motion modeling					
No explicit motion modeling	112.2	267.9	225.1	46.9	31.6
Direct motion regression	108.4	262.9	209.3	44.1	30.2
Conditional velocity field (ours)	**107.4**	**259.6**	**202.8**	**42.9**	**29.5**
Motion feedback					
No feedback	111.3	266.2	213.3	44.8	30.7
Loss-based feedback	111.0	265.5	209.9	44.2	30.3
Motion feature feedback (ours)	**107.4**	**259.6**	**202.8**	**42.9**	**29.5**

the field predicts instantaneous velocity as a function of the evolving joint configuration, elapsed time, and frame interval. Its integrated joint positions and differentiated accelerations therefore describe the same body-state transition, allowing position, velocity, and acceleration supervision to jointly constrain a shared motion function. This structured coupling discourages mutually inconsistent motion estimates and provides more reliable motion cues for mesh reconstruction, thereby reducing global joint errors and improving temporal consistency.

Effect of motion feature feedback. Across the three feedback variants, W-MPJPE decreases from 266.2 mm with no feedback to 265.5 mm with loss-based feedback, and further to 259.6 mm with Motion feature feedback (Table 4). The no-feedback variant leaves the field estimates unused by the reconstruction branch, whereas loss-based feedback adds an output-level consistency loss between field-integrated and reconstructed joint positions. Motion feature feedback instead encodes the estimated body-state transition in temporal tokens and uses them to update the human latents before mesh regression. This comparison suggests that feeding motion-aware temporal information into the human latents is more effective for mesh reconstruction than aligning the two predictions through a consistency loss alone.

Table 5: Position and acceleration supervision on EMDB-2.

Pos. sup.	Acc. sup.	WA-MPJPE	W-MPJPE	W-Jitter	W-Accel	W-FS
✗	✗	108.0	259.8	205.3	43.4	29.6
✓	✗	**107.4**	259.7	203.2	43.0	**29.4**
✗	✓	109.0	264.2	205.8	43.5	29.6
✓	✓	**107.4**	**259.6**	**202.8**	**42.9**	29.5

Effect of velocity field supervision. With velocity supervision alone, Velo3R already achieves a W-MPJPE of 259.8 mm (Table 5). Adding acceleration supervision without joint-position supervision slightly increases W-Jitter from 205.3 to 205.8 and W-Accel from 43.4 to 43.5, while W-FS remains at 29.6. This pattern may reflect that acceleration supervision constrains changes in velocity without directly anchoring joint positions; thus, position estimates may still drift over time. When joint-position supervision is included, acceleration supervision becomes complementary: it further reduces W-Jitter and W-Accel while keeping W-FS comparable.

5 Conclusion

We present Velo3R, a unified online framework for 4D human–scene reconstruction from monocular video. Its conditional velocity field couples joint positions, velocities, and accelerations through integration and differentiation, allowing these quantities to be supervised together. Motion feature feedback uses predicted velocities and integrated displacements to refine human features before mesh regression, while joints from the reconstructed mesh anchor the next transition. Velo3R improves global motion accuracy and temporal consistency on EMDB-2 and RICH, while achieving competitive local reconstruction results on 3DPW and EMDB-1. Ablation studies further support the effectiveness of the conditional velocity field and motion feature feedback.

AI USE STATEMENT

We wrote the initial draft of this manuscript and used large language models to improve its clarity and readability. These tools helped improve the clarity and presentation of the manuscript. We reviewed all revisions and take full responsibility for the final content.

REFERENCES

Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-HMR: Multi-person whole-body human mesh recovery in a single shot. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 202–218, 2024. doi: 10.1007/978-3-031-73337-6_12.

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 (CVPR)*, pp. 8726–8737, June 2023. doi: 10.1109/CVPR52729.2023.00843.

Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3R: Everyone everywhere all at once. In *The Fourteenth International Conference on Learning Representations*, 2026.

Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3D human pose and shape from a video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1964–1973, 2021.

Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J. Black. TokenHMR: Advancing human mesh recovery with a tokenized pose representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1323–1333, 2024.

Qiao Feng, Yiming Huang, Yufu Wang, Jiatao Gu, and Lingjie Liu. PhysHMR: Learning humanoid control policies from vision for physically plausible human motion reconstruction. In *SIGGRAPH Asia 2025 Conference Papers*. Association for Computing Machinery, 2025. doi: 10.1145/3757377.3763951.

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 (ICCV)*, pp. 14783–14794, 2023.

Ernst Hairer, Syvert P. Nørsett, and Gerhard Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*, volume 8 of *Springer Series in Computational Mathematics*. Springer, 2 edition, 1993. doi: 10.1007/978-3-540-78862-1.

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 (CVPR)*, pp. 13274–13285, 2022.

Dingbang Huang, Etienne Vouga, Qixing Huang, and Georgios Pavlakos. Recovering physically plausible human-object interactions from monocular videos. 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7122–7131, 2018.

Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. EMDB: The electromagnetic database of global 3D human pose and shape in the wild. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14586–14597, 2023.

Marilyn Keller, Keenon Werling, Soyong Shin, Scott Delp, Sergi Pujades, C. Karen Liu, and Michael J. Black. From skin to skeleton: Towards biomechanically accurate 3D digital humans. *ACM Transactions on Graphics*, 42(6):253:1–253:15, 2023. doi: 10.1145/3618381.

Sangmin Kim, Minhyuk Hwang, Geonho Cha, Dongyoon Wee, and Jaesik Park. Coherent human-scene reconstruction from multi-person multi-view video in a single pass. *arXiv preprint arXiv:2603.12789*, 2026.

Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5252–5262, 2020.

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 (ICCV)*, pp. 2252–2261, 2019.

Cuong Le, Pavlo Melnyk, Urs Waldmann, Mårten Wadenbäck, and Bastian Wandt. QuaMo: Quaternion motions for vision-based 3D human kinematics capture. In *International Conference on Learning Representations (ICLR)*, 2026.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 71–91, 2024. doi: 10.1007/978-3-031-73220-1_5.

Da Li, Jiping Jin, Xuanlong Yu, Wei Liu, Xiaodong Cun, Kai Chen, Rui Fan, Jiangang Kong, and Xi Shen. SKEL-CF: Coarse-to-fine biomechanical skeleton and surface mesh recovery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 151–169, 2026a.

Mengfei Li, Peng Li, Zheng Zhang, Jiahao Lu, Chengfeng Zhao, Wei Xue, Qifeng Liu, Sida Peng, Wenxiao Zhang, Wenhan Luo, Yuan Liu, and Yike Guo. UniSH: Unifying scene and human reconstruction in a feed-forward pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14038–14049, 2026b.

Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 590–606, 2022. doi: 10.1007/978-3-031-20065-6_34.

Zhizheng Liu, Joe Lin, Wayne Wu, and Bolei Zhou. Joint optimization for 4D human-scene reconstruction in the wild. In *International Conference on Learning Representations (ICLR)*, 2026.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248:1–248:16, 2015. doi: 10.1145/2816795.2818013.

Kuang Luo, Jingshang Zhao, Yingping Wang, Jiayao Li, Junjie Wen, Jiong Liang, Henry Soekmadji, and Shaolin Liao. Physics-informed neural networks for PDE problems: A comprehensive review. *Artificial Intelligence Review*, 58:323, 2025. doi: 10.1007/s10462-025-11322-7.

Priyanka Patel and Michael J. Black. CameraHMR: Aligning people with perspective. In *2025 International Conference on 3D Vision (3DV)*, pp. 1562–1571, 2025. doi: 10.1109/3DV66043.2025.00146.

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 (CVPR)*, pp. 10975–10985, 2019.

Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045.

Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. HuMoR: 3D human motion model for robust pose estimation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11468–11479, 2021. doi: 10.1109/ICCV48922.2021.01129.

István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3D human pose and shape estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 140032–140065, 2024. doi: 10.52202/079017-4444.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 144:1–144:11, 2024. doi: 10.1145/3680528.3687565.

Boao Shi, Qiao Feng, Yiming Huang, and Lingjie Liu. Scene and human in one world: Reconstruction in a feedforward pass. *arXiv preprint arXiv:2606.27720*, 2026.

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2070–2080, 2024.

Chentao Song, He Zhang, Haolei Yuan, Haozhe Lin, Jianhua Tao, Hongwen Zhang, and Tao Yu. MetricHMSR: Metric human mesh and scene recovery from monocular images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J. Black. Putting people in their place: Monocular regression of 3D people in depth. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13233–13242, 2022. doi: 10.1109/CVPR52688.2022.01289.

Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J. Black. TRACE: 5D temporal regression of avatars with dynamic cameras in 3D environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8856–8866, 2023.

Tanuj Sur, Shashank Tripathi, Nikos Athanasiou, Ha Linh Nguyen, Kai Xu, Michael J. Black, and Angela Yao. UniCon3R: Contact-aware 3D human-scene reconstruction from monocular video, 2026.

Timo von Marcard, Roberto Henschel, Michael J. 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)*, pp. 614–631. Springer, 2018. doi: 10.1007/978-3-030-01249-6_37.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522, 2025a.

Wenjia Wang, Liang Pan, Huaijin Pi, Yuke Lou, Xuqian Ren, Yifan Wu, Zhouyingcheng Liao, Lei Yang, Rishabh Dabral, Christian Theobalt, and Taku Komura. EmbodMocap: In-the-wild 4D human-scene reconstruction for embodied agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. TRAM: Global trajectory and motion of 3D humans from in-the-wild videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 467–487, 2024. doi: 10.1007/978-3-031-73247-8_27.

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 (CVPR)*, pp. 1148–1159, 2025b. doi: 10.1109/CVPR52734.2025.00115.

Dingkun Wei, Zehong Shen, Yan Xia, Georgios Pavlakos, Yujun Shen, and Xiaowei Zhou. Natural human motion recovery by aligning high-order temporal dynamics from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Yan Xia, Xiaowei Zhou, Etienne Vouga, Qixing Huang, and Georgios Pavlakos. Reconstructing humans with a biomechanically accurate skeleton. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5355–5365, 2025. doi: 10.1109/CVPR52734.2025.00504.

Pradyumna Yalandur Muralidhar, Yuxuan Xue, Xianghui Xie, Margaret Kostyrko, and Gerard Pons-Moll. PhySIC: Physically plausible 3D human-scene interaction and contact from a single image. In *SIGGRAPH Asia 2025 Conference Papers*, 2025. doi: 10.1145/3757377.3763862.

Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21222–21232, 2023. doi: 10.1109/CVPR52729.2023.02033.

Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11038–11049, 2022.

Yufei Zhang, Jeffrey O. Kephart, Zijun Cui, and Qiang Ji. PhysPT: Physics-aware pretrained transformer for estimating human dynamics from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2305–2317, 2024.

Yiwen Zhao, Ce Zheng, Yufu Wang, Hsueh-Han Daniel Yang, Liting Wen, and László A. Jeni. OnlineHMR: Video-based online world-grounded human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13951–13961, 2026.

Appendix

Table of Contents

A Additional Experimental Results **14**
A.1 Additional Analysis of Velocity Field Supervision 14
A.2 Effect of the Joint-Position Loss Function 15
A.3 Training Context Length 15
A.4 Local Mesh Reconstruction 15
A.5 Inference Efficiency 16
B More Visualizations **17**
B.1 Qualitative Results 17
B.2 Cross-Attention Visualization 18
B.3 Long Sequences 19
C Implementation and Training Details **19**
C.1 Network Architecture and Initialization 19
C.2 Training Objectives and Loss Weights 19
C.3 Training Settings 21
C.4 Evaluation Protocols and Metrics 21

Limitations and future work. While the results are encouraging on the evaluated benchmarks, the current study does not cover the full range of scenes and capture conditions encountered in practice. Further evaluation on a broader range of real-world sequences would help characterize the method's generalization and robustness. Extending this evaluation and exploring improvements for challenging cases remain promising directions for future work.

A Additional Experimental Results

A.1 Additional Analysis of Velocity Field Supervision

Table A1: Effect of velocity field supervision on EMDB-2. Variants A–D add joint-position supervision, acceleration supervision, and bone-stretch regularization in turn, all with motion feature feedback disabled; the final model enables feedback.

	Components				Global Motion		Temporal Consistency		
Model variant	Pos.	Acc.	Bone	Feed.	WA-MPJPE	W-MPJPE	W-Jitter	W-Accel	W-FS
Human3R	–	–	–	–	112.2	267.9	225.1	46.9	31.6
A: Field + velocity sup.	✗	✗	✗	✗	111.9	266.9	215.9	45.3	31.2
B: + Joint-position sup.	✓	✗	✗	✗	112.0	267.7	216.3	45.4	31.1
C: + Acceleration sup.	✓	✓	✗	✗	<u>110.8</u>	<u>265.9</u>	<u>210.7</u>	<u>44.3</u>	<u>30.5</u>
D: + Bone stretch reg.	✓	✓	✓	✗	111.3	266.2	213.3	44.8	30.7
Velo3R	✓	✓	✓	✓	**107.4**	**259.6**	**202.8**	**42.9**	**29.5**

Table A1 complements the supervision analysis in Table 5 with a cumulative ablation in which motion feature feedback is disabled. With joint-position supervision held fixed, adding acceleration supervision (B→C) improves all five reported metrics, reducing W-MPJPE from 267.7 to 265.9 mm, W-Jitter from 216.3 to 210.7, and W-Accel from 45.4 to 44.3. This result echoes the main-paper finding that combining joint-position and acceleration supervision improves temporal consistency, and shows the same pattern with motion feature feedback disabled.

Choice of camera-space motion supervision. We supervise the field with camera-space joint positions, velocities, and accelerations. Consider a person standing nearly still while the camera moves: the person's world-space joints change little, but their coordinates in successive camera frames change with the viewpoint. Camera-space targets encode the combined effects of human motion and camera movement on the observed trajectories. They therefore provide temporal cues about viewpoint changes in addition to body motion. Within the shared decoder, temporal prompts interact with the token used for camera-pose prediction, providing a route for temporal information to affect the camera estimate. Motion feature feedback separately passes the field's cues to the human regressor to refine body pose and camera-space translation. By informing both body and camera estimates, these temporal cues can improve world-space reconstruction, consistent with the gains reported in Table 1.

A.2 Effect of the Joint-Position Loss Function

Table A2: Joint-position objective comparison on EMDB-2 and RICH.

Dataset	Joint-position loss	WA-MPJPE	W-MPJPE	W-Jitter	W-Accel	W-FS
EMDB-2	Smooth-L1	107.8	**259.3**	**202.6**	**42.91**	**29.27**
EMDB-2	MSE	**107.4**	259.6	202.8	42.95	29.46
RICH	Smooth-L1	105.4	173.8	149.1	28.11	23.44
RICH	MSE	**104.6**	**170.6**	**146.2**	**27.63**	**23.17**

This ablation compares MSE and Smooth-L1 for the joint-position loss used to supervise integrated positions on EMDB-2 and RICH. We focus on this objective because integrated positions directly constrain the predicted body configuration at the current frame, and their displacement from the preceding state is encoded by motion feature feedback before SMPL-X regression. We keep the normalized MSE objectives for velocity and acceleration fixed in this ablation; their supervision is evaluated separately in Tables 5 and A1. On EMDB-2, MSE and Smooth-L1 perform comparably, with differences below 0.4 mm in WA-MPJPE and W-MPJPE. On RICH (Huang et al., 2022), MSE improves all five metrics, reducing W-MPJPE by 3.2 mm and W-Jitter by 2.9 (Table A2). Because our synthetic joint-position targets are accurate, robustness to annotation outliers is less critical. MSE retains a quadratic penalty across residual magnitudes, whereas Smooth-L1 transitions to a linear penalty for larger residuals. Based on these results, we use MSE for integrated joint-position supervision, while retaining normalized MSE for velocity and acceleration.

A.3 Training Context Length

Table A3: Effect of training context length on EMDB-2. The two runs use the same training recipe and differ in frames per training clip.

Training clip	WA-MPJPE	W-MPJPE	W-Jitter	W-Accel	W-FS	RTE
4 frames	107.4	**259.6**	202.8	42.9	29.5	**2.1**
8 frames	**106.1**	260.0	**201.2**	**42.7**	**29.2**	**2.1**

This ablation evaluates the effect of clip length on accuracy and training cost. The 8-frame setting improves WA-MPJPE by 1.3 mm, W-Jitter by 1.6, W-Accel by 0.23, W-FS by 0.31 mm, and RTE by 0.018 percentage points, while W-MPJPE is 0.4 mm higher (Table A3). Both settings use the same number of clips and optimizer updates, so doubling the clip length approximately doubles training time and computation. Given these modest gains relative to the added cost, we use 4-frame clips for the final model.

A.4 Local Mesh Reconstruction

Table A4: Broader local human mesh reconstruction comparison on 3DPW and EMDB-1.

	3DPW			EMDB-1		
Method	PA-MPJPE ↓	MPJPE ↓	PVE ↓	PA-MPJPE ↓	MPJPE ↓	PVE ↓
Multi-stage						
CLIFF (Li et al., 2022)	43.0	69.0	81.2	68.3	103.3	123.7
HMR 2.0a (Goel et al., 2023)	44.4	69.8	82.2	61.5	97.8	120.0
TokenHMR (Dwivedi et al., 2024)	44.3	71.0	84.6	55.6	91.7	109.4
CameraHMR (Patel & Black, 2025)	38.5	62.1	72.9	43.7	73.0	85.4
NLF (Sárándi & Pons-Moll, 2024)	37.3	60.3	71.4	41.2	69.6	82.4
PromptHMR-Vid (Wang et al., 2025b)	**35.5**	**56.9**	**67.3**	**40.1**	**68.1**	**79.2**
One-stage						
BEV (Sun et al., 2022)	46.9	78.5	92.3	70.9	112.2	133.4
Multi-HMR (Baradel et al., 2024)	45.9	73.1	87.1	50.1	81.6	95.7
Human3R (Chen et al., 2026)	44.1	71.2	84.9	**48.5**	**73.9**	86.0
Velo3R	**43.9**	**70.5**	**84.4**	51.4	76.4	**85.8**

Table A4 compares Velo3R with image-based HMR methods, from multi-stage pipelines that detect and crop each person before single-person estimation to one-stage methods that handle both steps in a single network. Among one-stage methods, Velo3R achieves the lowest errors on all three 3DPW metrics. On EMDB-1, Velo3R achieves joint-position errors comparable to Human3R, with a slightly lower PVE.

A.5 Inference Efficiency

Table A5: Inference efficiency and EMDB-2 W-MPJPE$_{100}$. Delay is in seconds; error is in mm.

Method	FPS ↑	Avg. delay ↓	W-MPJPE$_{100}$ ↓
Offline			
GLAMR (Yuan et al., 2022)	17.2	14.16	726.6
SLAHMR (Ye et al., 2023)	0.1	2435.00	776.1
WHAM (w/ DPVO) (Shin et al., 2024)	9.3	26.18	354.8
TRAM (Wang et al., 2024)	2.1	115.95	222.4
GVHMR (w/ DPVO) (Shen et al., 2024)	12.1	20.12	276.5
Online			
TRACE (Sun et al., 2023)	24.6	0.04	1702.3
Human3R (Chen et al., 2026)	4.8	0.21	267.9
OnlineHMR (Zhao et al., 2026)	3.3	0.30	310.4
Velo3R-Lightweight	10.8	0.09	–
Velo3R-Final	4.6	0.22	259.6

Table A5 compares throughput, average delay, and EMDB-2 W-MPJPE$_{100}$ for offline and online world-coordinate HMR methods. Both Velo3R variants use only the current and preceding frames for inference. The lightweight model uses a Multi-HMR ViT-S human-prior encoder and reaches 10.8 FPS with an average delay of 0.09 s. The final model uses Multi-HMR ViT-L/896 and reaches 4.6 FPS with an average delay of 0.22 s, while achieving the lowest W-MPJPE$_{100}$ among the online methods (259.6 mm). The final model's measured inference cost is 5.734×10^{12} FLOPs. We fine-tune 73.23M parameters (6.24% of the full model), and training takes approximately one day on a single RTX 4090 GPU.

B More Visualizations

B.1 Qualitative Results

Our method remains robust under severe occlusion and truncation, as illustrated in Fig. A1. Building on Multi-HMR's one-stage multi-person reconstruction, our framework directly reasons about multiple people from full-frame images without requiring separate detection and cropping. It further strengthens this capability by exploiting cross-frame motion cues to maintain consistent human reconstruction even when parts of the body are not fully visible.



Figure A1: Qualitative 4D human–scene reconstruction under occlusion and truncation. Each pair shows an input image and the corresponding reconstruction; our method remains robust to partially visible and truncated people.

B.2 Cross-Attention Visualization

Figure A2 compares the spatial responses induced by the temporal tokens and the human-prompt tokens over five examples. The temporal-token responses are generally more concentrated on visible people, whereas the human-prompt responses can remain diffuse over nearby vegetation, buildings, railings, or other scene content. This pattern indicates that the temporal prompt can clean part of the background response inherited from the human prompt while retaining human features relevant to motion reasoning.



Figure A2: Cross-attention visualization for five examples. Temporal-token responses are more spatially concentrated on visible human regions, while human-prompt responses retain more diffuse activations on surrounding scene content. This qualitative pattern suggests that the temporal prompt filters part of the background response while preserving human features useful for motion reasoning.

B.3 Long Sequences

Figure A3 shows the reconstruction of a long sequence in a shared world coordinate system. Human meshes from different frames are overlaid on the scene point cloud, providing an overview of the recovered motion and its spatial relationship to the surrounding geometry.

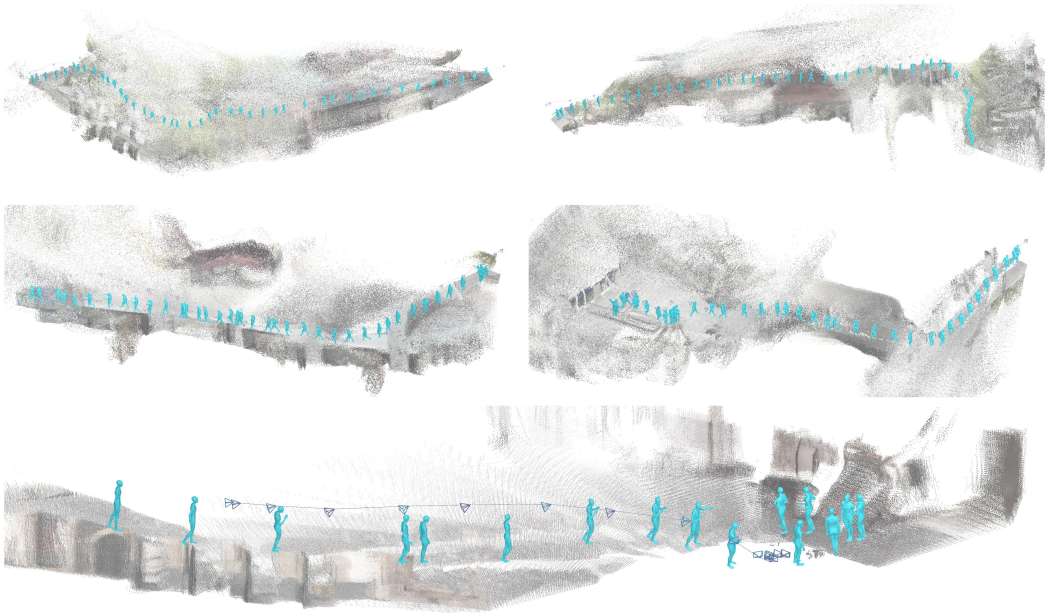

Figure A3: **Long-sequence world reconstruction.** Reconstructed human meshes, scene point clouds, and the camera trajectory, shown from multiple viewpoints in a shared world coordinate system.

C Implementation and Training Details

C.1 Network Architecture and Initialization

Human and temporal tokens are 768-dimensional, while the human-prior feature is 1,024-dimensional. At each frame, a $768 \to 768 \to 768$ MLP with GELU activations generates the temporal prompt from a detached copy of the current human prompt. This reuses the prompt features while keeping the projection independently trainable. The conditional velocity field combines a 128-dimensional whole-pose encoding, 32-dimensional joint embeddings, and separate 16-dimensional embeddings of query time and frame interval. Its MLP has three linear layers, hidden width 512, and SiLU activations. The three MLPs in Eqs. 5–6 have independent parameters. The outer MLP in Eq. 5 and the MLP in Eq. 6 each have two linear layers, hidden width 768, and GELU activations. We initialize both output layers with zero-mean normal weights of standard deviation 10^{-4} and zero biases. Finally, the SMPL-X decoder uses head-centered translation: the predicted translation specifies the head position in camera coordinates.

C.2 Training Objectives and Loss Weights

Overall objective. The training objective is the sum of scene reconstruction, human reconstruction, and temporal field supervision:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{human}} + \mathcal{L}_{\text{temporal}}. \tag{8}$$

Scene reconstruction. The scene objective is

$$\mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{rgb}} + \mathcal{L}_{\text{point}} + \mathcal{L}_{\text{cam}}. \tag{9}$$

The RGB term is a squared-error loss; the pointmap term combines confidence-weighted Euclidean residuals with a log-confidence penalty (coefficient 0.2), for pointmaps in the current and first-frame

camera coordinates. The camera term sums Euclidean errors in normalized translation and rotation-quaternion representations.

Human reconstruction. The human objective includes detection, mask, body-parameter, 3D, and 2D supervision:

$$\mathcal{L}_{\text{human}} = \lambda_{\text{det}}\mathcal{L}_{\text{det}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{smpl}} + \mathcal{L}_{\text{3D}} + \lambda_{\text{2D}}\mathcal{L}_{\text{2D}}. \tag{10}$$

Here $\mathcal{L}_{\text{det}}$ is the human-detection focal loss and $\mathcal{L}_{\text{mask}}$ is the human-mask binary cross-entropy. The SMPL-X parameter loss is

$$\mathcal{L}_{\text{smpl}} = \mathcal{L}_{\text{pose}} + 0.1\mathcal{L}_{\text{shape}} + \mathcal{L}_{\text{trans}}, \tag{11}$$

where each term is an L1 error on pose rotation matrices, shape coefficients, and head-centered translation, respectively. The 3D term sums L1 errors for joints and vertices in pelvis-relative and camera coordinates; the 2D term sums L1 errors on their image reprojections. $\lambda_{\text{det}} = 0.1$, $\lambda_{\text{mask}} = 1$, the outer weights on $\mathcal{L}_{\text{smpl}}$ and $\mathcal{L}_{\text{3D}}$ are 1, and $\lambda_{\text{2D}} = 0.01$.

Temporal supervision. The field is supervised using its integrated joint positions, interval-average velocity, and acceleration at time t. We use mean-reduced MSE for all three terms:

$$\mathcal{L}_{\text{temporal}} = \mathcal{L}_{\text{pos}} + \mathcal{L}_{\text{vel}} + \mathcal{L}_{\text{acc}} + 0.01\mathcal{L}_{\text{bone}}. \tag{12}$$

Let $\hat{\mathbf{X}}_t$, $\hat{\mathbf{v}}_t$, and $\hat{\mathbf{a}}_t$ denote the ground-truth camera-space joint positions in meters, velocities, and accelerations. The position loss compares the integrated joint positions directly in meters. For velocity and acceleration, we apply the fixed per-joint, per-coordinate training normalization to both prediction and target, with $\tilde{\mathbf{u}} = (\mathbf{u} - \boldsymbol{\mu}_u)/\mathbf{s}_u$, where $\boldsymbol{\mu}_u$ and $\mathbf{s}_u$ are the training mean and standard deviation. With MSE denoting the mean squared error over the coordinates of one person's joint array, the field terms are

$$\mathcal{L}_{\text{pos}} = \text{MSE}(\mathbf{X}_t^{\text{field}}, \hat{\mathbf{X}}_t), \quad \mathcal{L}_{\text{vel}} = \text{MSE}(\tilde{\mathbf{v}}_t, \tilde{\hat{\mathbf{v}}}_t), \quad \mathcal{L}_{\text{acc}} = \text{MSE}(\tilde{\mathbf{a}}_t^{\text{field}}, \tilde{\hat{\mathbf{a}}}_t). \tag{13}$$

The acceleration prediction is the JVP of the velocity field at the interval endpoint, as in Eq. 4.

Temporal targets. The velocity target in Eq. 13 matches the sampled transition interval:

$$\hat{\mathbf{v}}_t = (\hat{\mathbf{X}}_t - \hat{\mathbf{X}}_{t-1})/\Delta_t, \tag{14}$$

with identities aligned at the two sampled endpoints. The acceleration target $\hat{\mathbf{a}}_t$ uses centered finite differences on the original source timeline and accounts for nonuniform time intervals. These neighboring annotations provide training labels; inference remains causal. For a joint array with J joints, the per-person MSE is

$$\text{MSE}(\mathbf{u}, \hat{\mathbf{u}}) = \frac{1}{3J}\left\|\mathbf{u} - \hat{\mathbf{u}}\right\|_2^2. \tag{15}$$

All field terms require valid motion history, a valid image and human annotation, and a detected person; velocity and acceleration also require their respective valid target annotations. We average over valid people and joint coordinates within each frame, then over clip frames. Frames with no valid samples contribute zero.

Bone-stretch regularization. The regularizer penalizes the instantaneous rate of change of skeletal bone lengths at the start of each transition:

$$\mathcal{L}_{\text{bone}} = \frac{1}{|\mathcal{E}|}\sum_{(i,j)\in\mathcal{E}}\left[\mathbf{b}_{ij}^{\mathsf{T}}(\mathbf{v}_{0,j} - \mathbf{v}_{0,i})\right]^2, \tag{16}$$

Here $\mathbf{b}_{ij} = \mathbf{X}_{0,j} - \mathbf{X}_{0,i}$ is the start-state bone vector, $\mathbf{v}_0$ is the field velocity at the start, and $\mathcal{E}$ contains the 21 edges of the 22-joint skeleton. The inner product is the instantaneous change of half the squared bone length. We average its square over edges and valid people per frame, then over clip frames.

C.3 Training Settings

We jointly train the human prompt, human head, temporal prompt, velocity field, and feedback networks, while keeping the pretrained scene backbone, human-prior encoder, detector, and scene and camera heads frozen. Training samples 4,000 four-frame clips per epoch from the filtered BEDLAM pool, with a maximum frame-index interval of four. A batch size of eight gives 500 updates per epoch and 20,000 updates over 40 epochs. We use AdamW with $\beta = (0.9, 0.95)$, weight decay 0.05, and gradient-norm clipping at 1. The learning rate warms up for five epochs to 5×10^{-5}, then follows cosine decay toward 10^{-6}. The training seed is 0, and the base seed for per-epoch clip sampling is 777. We use pretrained CUT3R and Multi-HMR weights for the corresponding components and initialize the new motion modules randomly.

C.4 Evaluation Protocols and Metrics

EMDB-2 evaluation covers 25 sequences and 43,129 frames; RICH evaluation covers the 191 static-camera sequences. We resize inputs to a long side of 512 pixels and use every frame. The recurrent state resets every 50 frames on EMDB-2. On RICH, it is not reset within each sequence. W-MPJPE and WA-MPJPE measure global joint errors after alignment using the first two frames and the full evaluation segment, respectively; RTE measures root-trajectory error normalized by the ground-truth path length. W-Jitter is the mean norm of the third finite difference of predicted world-space joints, multiplied by the cubed frame rate and divided by 10. W-FS averages predicted displacement of four foot vertices when the corresponding ground-truth displacement is below $10\,\text{mm}$. W-Accel compares predicted and ground-truth second derivatives in world coordinates, using a common orientation from the first evaluated camera without scale or pelvis alignment. It uses valid consecutive samples of the same identity and is aggregated by the number of valid derivative windows.