mmSimPrior: Learning Simulation Priors for Data-Efficient Real-World Generalizable Radar-Based Human Motion Reconstruction

Cheng Guo[1], **Qiming Cao**[2], **Shengkai Xu**[1],
Haoyu Xie[1], **Kaixiang Su**[1], **Pu Wang**[1], **Hongfei Xue**[1]
[1]University of North Carolina at Charlotte, Charlotte, NC, USA
[2]Purdue University, West Lafayette, IN, USA
{cguo3, sxu7, hxie3, ksu4, pwang13, hxue2}@charlotte.edu, cao393@purdue.edu

Abstract

Millimeter-wave (mmWave) radar offers privacy-preserving and lighting-robust sensing for human motion reconstruction, but learning models that generalize across real deployments requires diverse paired radar–motion data that are costly to collect. Simulation provides scalable supervision, yet models trained on clean synthetic signals transfer poorly because of multipath, clutter, response statistics, and resolution degradation. We present **mmSimPrior**, a simulation-pretrained framework that factorizes transferable knowledge into signal, motion, and radar-to-motion mapping priors. A multi-modal signal encoder is pretrained with a physics-informed domain-randomization curriculum that emulates propagation- and acquisition-level variations, while a joint-temporal tokenizer learns a discrete prior over plausible human motion. A shared mapping prior supports classification over a learned motion codebook for constrained zero-shot reconstruction and continuous regression for flexible limited-data adaptation. We further construct a 4.2M-frame, 31K-sequence dataset suite and introduce a No-Overlap Setting that excludes repeated complete subject–environment–location–motion configurations across adaptation and test. Experiments on mmSimPrior-Real and RT-Pose demonstrate consistent gains: with only 24 paired real sequences, mmSimPrior-Reg reduces MPJPE by 24.7–39.0% over the strongest baseline across the three environments, while mmSimPrior-Cls reduces zero-shot MPJPE by 8.5% without finetuning.

Project Page — https://ch3ngguo.github.io/mmsimprior/

1 Introduction

Human motion reconstruction supports applications in AR/VR, robotics, and health monitoring (Von Marcard et al. 2018; Goel et al. 2023; Chen et al. 2022; An and Ogras 2021). Although RGB and RGB-D methods have achieved substantial progress (Ionescu et al. 2013; Mehta et al. 2017; Kanazawa et al. 2018; Kolotouros et al. 2019; Kocabas, Athanasiou, and Black 2020; Goel et al. 2023), optical sensors degrade under poor illumination and occlusion and raise privacy concerns in sensitive indoor environments. Millimeter-wave (mmWave) radar instead provides range, Doppler, and angular measurements while remaining insensitive to lighting and preserving visual privacy (Richards 2014; Chen 2011; Zhao et al. 2018b,a, 2019).

Real-world radar-based human motion reconstruction remains challenging because radar responses are sparse, noisy, and jointly determined by body motion, subject geometry, sensing configuration, and the environment. Without diverse paired coverage, models can mistake deployment-specific reflection patterns for motion evidence. However, collecting radar–motion annotations requires participant recruitment, sensor calibration, synchronization, and repeated recordings across environments, leaving existing datasets limited in scale and models prone to collection-specific overfitting.

Simulation offers scalable paired supervision and broad motion coverage (Zhang, Li, and Zhang 2022; Xue et al. 2023; Chen and Zhang 2023; Deng et al. 2023; Joshi et al. 2024; Huang et al. 2025), but increasing synthetic scale alone does not ensure real-world transfer. Even physics-based simulators cannot fully reproduce multipath, clutter, hardware-dependent response statistics, and distance-dependent resolution degradation (Richards 2014; Chen 2011; Xue et al. 2023; Chen and Zhang 2023). The central question is therefore whether simulation can serve as the primary source of paired supervision for fine-grained radar-based human motion reconstruction, rather than merely augmenting a deployment-specific real training set.

We address this question with **mmSimPrior**, a simulation-pretrained transfer framework that factorizes simulation-derived knowledge across radar observation, human motion, and their correspondence. A multi-modal signal prior is pretrained under physics-informed domain randomization to improve robustness to propagation- and acquisition-level variations, while a joint-temporal motion prior captures diverse motion structure from mocap data. A dual-mode mapping prior learns radar–motion correspondence through codebook-based classification for stronger zero-shot constraints and continuous regression for flexible real-data adaptation. For the classification mode, the frozen motion tokenizer additionally provides target code distributions during limited-data adaptation, regularizing parameter-efficient updates toward the pretrained joint-temporal motion space. For the regression mode, the same prior can optionally refine direct predictions to improve temporal consistency.

Our contributions are summarized as follows:

- We formulate data-efficient radar-based human motion reconstruction as a **simulation-pretrained transfer problem** and introduce **mmSimPrior**, which supports zero-shot inference and minute-level adaptation to real-world deployments.

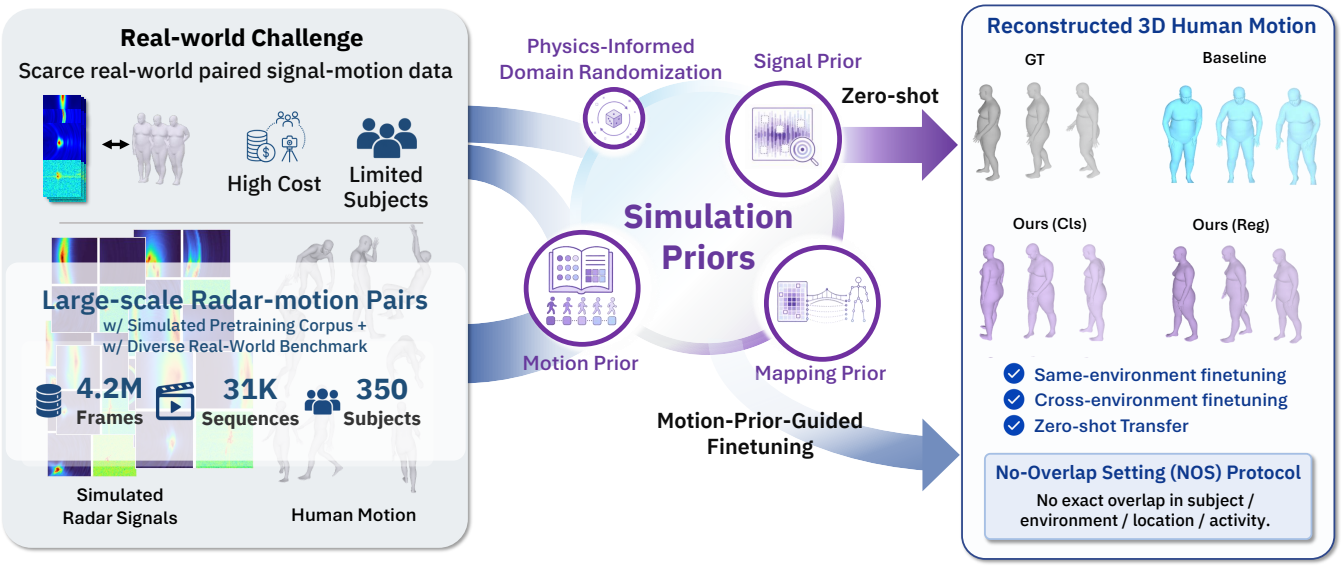


Figure 1: **Overview of mmSimPrior.** We construct large-scale simulated radar–motion pairs and learn transferable signal, motion, and mapping priors with physics-informed domain randomization. The learned priors support zero-shot transfer and motion-prior-guided real-world adaptation under a no-overlap evaluation protocol.

- We develop a unified architecture that factorizes simulation-derived knowledge into multi-modal signal, joint-temporal motion, and dual-mode mapping priors. Physics-informed domain randomization improves sim-to-real robustness, while classification and regression provide complementary trade-offs between constrained zero-shot reconstruction and flexible real-data adaptation.
- We further construct a 4.2M-frame, 31K-sequence dataset suite and introduce a No-Overlap Setting that excludes repeated complete subject–environment–location–motion configurations across adaptation and test.

2 Related Work

2.1 Radar-Based Human Motion Reconstruction and Generalization

Radar-based pose and mesh reconstruction has been studied using sparse point clouds, Range–Doppler or Range–Angle heatmaps, multi-view heatmaps, and high-dimensional radar tensors (Xue et al. 2021, 2022; Lee et al. 2023; Rahman et al. 2024; Ho et al. 2024; Kato et al. 2025; Fan et al. 2026). Prior methods employ direct mesh regression, diffusion, weak supervision, and radar-tensor modeling (Xue et al. 2021; Fan et al. 2024; Kato et al. 2025; Ho et al. 2024). Recent datasets and methods have also considered cross-subject, cross-environment, or unseen-activity evaluation (Xue et al. 2023; Rahman et al. 2024; Yang et al. 2023). Despite this progress, fine-grained radar-based human reconstruction still relies heavily on paired real recordings collected under specific subjects, environments, and sensing configurations. Performance therefore remains closely tied to the available real-data coverage and evaluation protocol. mmSimPrior instead studies simulation-pretrained zero-shot and limited-data transfer to real deployments, evaluated under a strict no-overlap protocol.

2.2 Simulation-Based Radar Supervision and Sim-to-Real Transfer

Prior work reduces radar-data collection costs through physics-based simulation, motion-capture-driven synthesis, video-conditioned generation, and generative models (Xue et al. 2023; Chen and Zhang 2023; Deng et al. 2023; Jiang et al. 2026; Huang et al. 2025). Synthetic observations have been used to augment real-world training, support RF sensing, and enable radar-based activity or language understanding (Xue et al. 2023; Chen and Zhang 2023; Lai et al. 2026; Yan et al. 2025). mmSimPrior instead treats simulation as the primary source of paired supervision rather than as supplementary augmentation. Unlike prior radar-based human motion reconstruction methods trained mainly on paired recordings from the deployment domain, we study whether simulation can replace most such supervision for fine-grained motion reconstruction, with either no data from the evaluation dataset or only minute-level real-world adaptation. To enable this shift, we factorize simulation-derived knowledge into signal, motion, and mapping priors and introduce physics-informed domain randomization for robust real-world transfer.

3 Radar Observation and Dataset Construction

This section first introduces the radar representations used throughout the paper and then describes the construction of the large-scale simulated corpus used for pretraining. Unlike RGB images, radar heatmaps are physically indexed response maps derived from complex radio-frequency measurements.

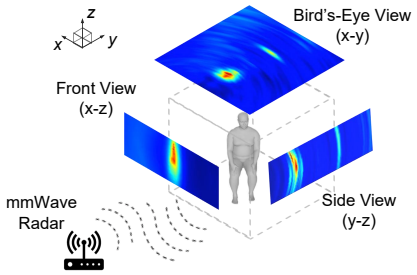


Figure 2: **Illustration of multi-view MVDR spatial representation.** We compute a 3D MVDR spatial heatmap around the human body and project it into three orthogonal 2D views.

To retain both radial kinematics and metric spatial structure, we represent each radar frame using an RD heatmap together with three orthogonal MVDR projections.

3.1 Factorized Radar Heatmap Representation

FMCW radar measurements. At frame t, an FMCW radar records a complex IF tensor $\mathbf{X}_t \in \mathbb{C}^{N_{\text{tx}} \times N_{\text{rx}} \times N_c \times N_s}$, indexed by transmitters, receivers, chirps, and ADC samples. The IF tensor contains propagation-delay, radial-motion, and inter-antenna phase information, but does not provide explicit correspondences between measurements and human body parts.

Range–Doppler heatmap. By applying Range-FFT along the fast-time dimension and a Doppler-FFT along the chirp dimension, the resulting Range–Doppler (RD) heatmap H_t^{RD} represents reflected response magnitude as a function of propagation distance and radial velocity. Therefore, unlike an RGB pixel indexed by image-plane position, each RD cell corresponds to a physically defined range–velocity bin. The RD representation preserves radial kinematic evidence that is useful for separating moving body responses from static or slowly varying background structures.

Multi-view MVDR spatial heatmaps. The RD heatmap does not directly resolve the 3D direction of a reflection. We therefore use phase differences across the virtual antenna array to estimate spatial response power through Minimum Variance Distortionless Response (MVDR) beamforming. For a candidate location $\mathbf{p} = (x, y, z)$ on a discretized 3D grid, the MVDR spectrum is

$$P_t(\mathbf{p}) = \left| \frac{1}{\mathbf{a}(\mathbf{p})^H \mathbf{R}^{\dagger}_{t,r(\mathbf{p})} \mathbf{a}(\mathbf{p})} \right|, \tag{1}$$

where $\mathbf{a}(\mathbf{p})$ is the steering vector associated with the candidate position, $\mathbf{R}_{t,r(\mathbf{p})}$ is the antenna-domain covariance matrix at the corresponding range bin, and $\dagger$ denotes the Hermitian pseudoinverse. Intuitively, MVDR assigns high response power to locations whose expected inter-antenna phase pattern agrees with the measured signal while suppressing inconsistent interference.

Direct learning over the full 3D response volume is computationally expensive and susceptible to sparse volumetric artifacts. We instead factorize it into three orthogonal mean projections as shown in Fig. 2. The front view captures horizontal–vertical body structure, the bird's-eye view preserves ground-plane position and orientation, and the side view provides complementary depth–height evidence. Together, the factorized observation at frame t is $\mathbf{o}_t = \left\{H_t^{\text{RD}}, H_t^{\text{front}}, H_t^{\text{bird}}, H_t^{\text{side}}\right\}$. Given a radar heatmap sequence $\mathbf{o}_{1:T}$, the reconstruction task is to estimate the SMPL-X motion parameters $\Theta_{1:T} = \{\boldsymbol{\theta}_{1:T}, \mathbf{t}_{1:T}, \boldsymbol{\beta}\}$, where $\boldsymbol{\theta}_t$ contains continuous 6D body-joint rotations, $\mathbf{t}_t$ is the global root translation, and $\boldsymbol{\beta}$ denotes sequence-level body shape.

3.2 mmSimPrior-Sim: Physics-Based Radar–Motion Synthesis

We construct mmSimPrior-Sim by driving a physics-based FMCW radar simulator (Xue et al. 2023) with AMASS (Mahmood et al. 2019) motion sequences. Each motion sequence is represented by a temporally varying SMPL-X mesh and resampled to the radar frame rate. Mesh triangles are treated as surface reflectors characterized by their barycenters, surface normals, and area-dependent response strengths.

For a mesh triangle q, transmit antenna i, receive antenna j, and frequency sample n, its bistatic reflection can be expressed as

$$s_{qijn} = \frac{A_q G_{qij}}{d_{qi}^{\text{tx}} d_{qj}^{\text{rx}}} \exp\left(j2\pi(f_c + \Delta f_n)\tau_{qij}\right), \tag{2}$$

where A_q represents the surface-area contribution, d_{qi}^{tx} and d_{qj}^{rx} are the transmit- and receive-side path lengths, τ_{qij} is the corresponding propagation delay, and G_{qij} models the dependence of reflection strength on surface orientation. Geometrically invalid back-facing responses are removed, and the remaining triangle responses are accumulated to synthesize the complex IF signal.

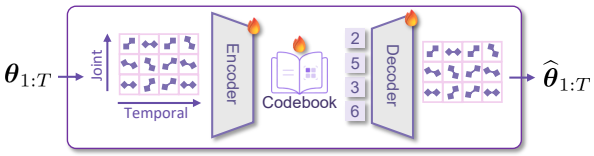
Each motion is rendered at multiple radar–subject distances to introduce distance-dependent variation in attenuation, angular resolution, and body-reflection patterns. This yields synchronized radar heatmaps paired with SMPL-X motion annotations. The resulting radar–motion pairs provide scalable supervision for learning the signal and mapping priors, while their underlying AMASS motions are used to pretrain the tokenized motion prior.

4 Methodology

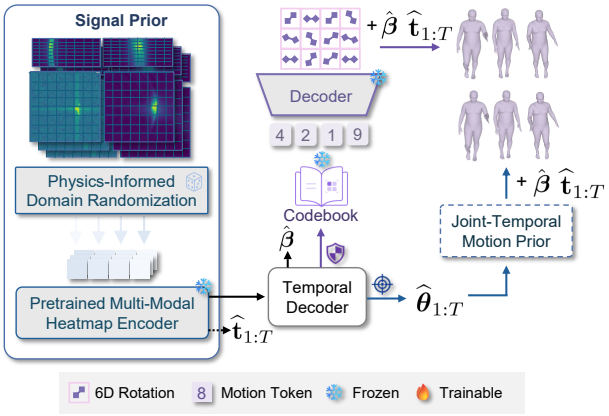
The goal of mmSimPrior is to recover SMPL-X pose, translation, and shape from factorized radar heatmap sequences. As shown in Fig. 3, a signal prior encodes synchronized RD and multi-view MVDR heatmaps, a joint-temporal motion prior models human motion structure, and a shared mapping prior connects the two. The mapping prior supports codebook-based classification for constrained zero-shot reconstruction and continuous regression for flexible real-world adaptation.

4.1 Physics-Informed Domain Randomization

Although simulated and real IF signals are processed through the same pipeline, their resulting heatmaps can still differ substantially. A clean simulator cannot fully capture the



(a) Joint-Temporal Motion Prior



(b) Mapping Prior with Dual Prediction Modes

Figure 3: **Overview of mmSimPrior.** (a) Joint-temporal tokenization learns the motion prior. (b) Domain-randomized radar heatmaps are encoded by the signal prior and decoded through classification- or regression-based mapping modes.

stochastic effects of real-world electromagnetic propagation and radar acquisition, including indirect multipath returns, clutter, and noise-floor statistics.

We therefore introduce physics-informed domain randomization in radar heatmap space. Rather than repeatedly simulating the complete propagation process, we construct efficient heatmap-space surrogates whose forms, spatial directions, and modality assignments are informed by the corresponding radar phenomena. As illustrated in Fig. 4, these transformations diversify low-level radar appearance while preserving the dominant human response and its paired motion annotation.

For a simulated heatmap sequence $\mathbf{o}_{1:T}$, we generate

$$\widetilde{\mathbf{o}}_{1:T} = \mathcal{A}_{\boldsymbol{\xi}}\left(\mathbf{o}_{1:T}\right), \qquad \boldsymbol{\xi} \sim q_s(\boldsymbol{\xi}), \tag{3}$$

where $\mathcal{A}_{\boldsymbol{\xi}}$ denotes a modality-conditioned composition of label-preserving transformations, and q_s controls their occurrence and severity at training step s. These label-preserving perturbations expose the model to propagation- and acquisition-level variations before deployment.

Multipath ghosting. Indirect propagation paths can generate displaced and attenuated replicas of strong human responses. We emulate this phenomenon by shifting high-energy regions along view-dependent spatial axes:

$$\tau_{\text{ghost}}(H) = H + \alpha(s)\mathcal{S}_{\Delta}\left(H \odot \mathbb{I}[H > \rho]\right), \tag{4}$$

where $\mathcal{S}_{\Delta}$ applies a spatial shift, ρ selects dominant responses, and $\alpha(s)$ controls the replica strength. The shift direction is conditioned on the physical axes represented by each MVDR view.

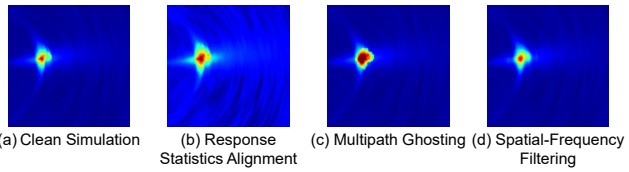


Figure 4: **Physics-informed domain randomization.** Modality-conditioned perturbations preserve the dominant human response while reducing the discrepancy between matched simulated and real MVDR heatmaps.

Resolution degradation. Finite angular resolution and increasing sensing distance can spread a localized reflection over neighboring spatial cells. We approximate this effect by applying scheduled Gaussian spreading to spatial responses$\tau_{\text{spread}}(H) = H + \beta(s)\left(\mathbb{I}[H > \rho] * G_{\sigma}\right)$, where G_{σ} is a Gaussian kernel. This transformation reduces reliance on simulation-specific sharp spatial responses.

Acquisition-level variations. We additionally perturb spatial-frequency content using $\tau_{\text{freq}}(H) = \mathcal{F}^{-1}(M_{\omega_c} \odot \mathcal{F}(H))$ and apply monotone intensity remapping $\tau_{\text{stat}}(H) = g_{\eta}(H)$ to capture acquisition-dependent variations in antenna response, gain, clutter, and noise-floor statistics. Both transformations alter low-level radar appearance while preserving the dominant spatial response structure.

Progressive nuisance curriculum. Applying severe randomization from the beginning can obscure the radar–motion correspondence. We therefore use a piecewise curriculum, $\boldsymbol{\phi}(s) = \boldsymbol{\phi}_i$ for $s \in [s_i, s_{i+1})$, to progressively increase selected nuisance strengths, including multipath replicas and peak spreading. This allows the model to first learn stable correspondences before being exposed to stronger deployment variations.

4.2 Multi-Modal Radar Signal Prior

The four radar modalities provide complementary cues but differ in resolution and aspect ratio. For modality k, a modality-specific adapter produces an equal token budget, $\mathbf{U}_t^{(k)} = f_{\text{patch}}^{(k)}(H_t^{(k)}) \in \mathbb{R}^{N \times d}$. The padded front and side views use anisotropic patches, whereas RD and bird's-eye use square patches.

We pretrain the adapters and a shared Transformer as a multi-modal masked autoencoder on domain-randomized simulation. We independently mask each modality and jointly encode all visible tokens with a shared Transformer, allowing attention to associate RD kinematics with MVDR spatial responses. Modality-specific reconstruction heads are used only during pretraining; the shared encoder is retained as the signal prior.

4.3 Joint-Temporal Motion Prior

To constrain this ambiguous radar-based motion reconstruction, we learn a motion prior from SMPL-X motion sequences using a VQ-VAE (van den Oord, Vinyals, and Kavukcuoglu 2017). A motion sequence is represented by continuous 6D rotations $\mathbf{R}_{1:T} \in \mathbb{R}^{T\times J\times 6}$, including the root orientation and body-joint rotations. The motion encoder maps the rotation sequence to a joint-temporal latent lattice $\mathbf{E} = E_{\mathrm{mot}}(\mathbf{R}_{1:T}) \in \mathbb{R}^{T'\times J'\times d_c}$, where T' and J' denote the reduced temporal and joint-axis resolutions, respectively. Each latent vector is assigned to its nearest entry in a learned codebook $\mathcal{C} = \{\mathbf{c}_m\}_{m=1}^{|\mathcal{C}|}$:

$$z_{t,j} = \arg\min_{m} \|\mathbf{E}_{t,j} - \mathbf{c}_m\|_2^2 . \tag{5}$$

A motion decoder reconstructs the continuous rotation sequence from the quantized joint-temporal representation $\widehat{\mathbf{R}}_{1:T} = D_{\mathrm{mot}}\left(\{\mathbf{c}_{z_{t,j}}\}\right)$. The tokenizer is trained with rotation-reconstruction, temporal, joint-geometry, and commitment losses, after which its codebook and decoder are frozen.

The motion prior serves asymmetric roles across the two modes. For Cls, its codebook and decoder form the primary prediction path, while its frozen tokenizer additionally provides target code indices during real-world adaptation. For Reg, the prior is used only as an optional refinement for temporal consistency.

4.4 Dual-Space Radar-Conditioned Motion Decoding

The mapping prior first temporally aligns frame-level heatmap features $\mathbf{U}_{1:T}$ as $\widetilde{\mathbf{U}}_{1:T'} = g_{\mathrm{align}}(\mathbf{U}_{1:T})$, and aggregates temporal context as $\mathbf{F}_{1:T'} = g_{\mathrm{temp}}(\widetilde{\mathbf{U}}_{1:T'})$. At each reduced time step, a shared temporal query cross-attends to the multi-modal context, $\mathbf{m}_t = g_{\mathrm{query}}(\mathbf{q}, \mathbf{F}_t)$. Both modes share this decoder and differ only in their output space.

Classification-based mode. The Cls head predicts codebook logits $\boldsymbol{\ell}_{t,j} = h_{\mathrm{Cls}}(\mathbf{m}_t)_j$. Following previous work (Dwivedi et al. 2024; Saleem et al. 2025), we obtain code probabilities $\boldsymbol{\pi}_{t,j} = \mathrm{softmax}(\boldsymbol{\ell}_{t,j})$ and perform differentiable soft code retrieval:

$$\begin{aligned} \widetilde{\mathbf{c}}_{t,j} &= \sum_{m=1}^{|\mathcal{C}|} \boldsymbol{\pi}_{t,j}[m]\mathbf{c}_m, \\ \widehat{\mathbf{R}}^{\mathrm{Cls}}_{1:T} &= D_{\mathrm{mot}}\left(\{\widetilde{\mathbf{c}}_{t,j}\}_{t,j}\right). \end{aligned} \tag{6}$$

The frozen codebook and decoder constrain predictions toward learned joint-temporal motion patterns.

Regression-based mode. The Reg head directly predicts continuous body rotations, $\widehat{\mathbf{R}}^{\mathrm{Reg}}_{1:T} = h_{\mathrm{Reg}}(\mathbf{m}_{1:T'})$, retaining greater flexibility for real-domain correction. Optionally, the Reg output can be projected through the frozen motion prior to improve temporal smoothness.

Both modes are trained using SMPL-X supervision:

$$\begin{aligned} \mathcal{L}_{\mathrm{map}} = \lambda_\theta\mathcal{L}_\theta + \lambda_{\mathrm{joint}}\mathcal{L}_{\mathrm{joint}} + \lambda_{\mathrm{trans}}\mathcal{L}_{\mathrm{trans}} + \lambda_\beta\mathcal{L}_\beta \\ + \lambda_{\mathrm{dyn}}\mathcal{L}_{\mathrm{dyn}}, \end{aligned} \tag{7}$$

where the losses supervise body rotations, root-relative joints, global translation, body shape, and translation dynamics.

4.5 Motion-Prior-Guided Real-World Adaptation

Given a small paired real set $\mathcal{D}_{\mathrm{ft}}$, we freeze the pretrained signal-encoder weights, insert LoRA adapters into its Transformer blocks (Hu et al. 2022), and jointly optimize the adapters with the temporal mapping modules and prediction heads.

For Cls, the frozen tokenizer converts ground-truth motion into target indices $z^*_{t,j}$, and we optimize $\mathcal{L}^{\mathrm{Cls}}_{\mathrm{ft}} = \mathcal{L}_{\mathrm{map}} + \lambda_{\mathrm{tok}}\mathcal{L}_{\mathrm{tok}}$, where $\mathcal{L}_{\mathrm{tok}} = \frac{1}{T'J'}\sum_{t,j}\mathrm{CE}(\boldsymbol{\ell}_{t,j}, z^*_{t,j})$. We linearly warm λ_{tok} and refer to this token supervision as motion-prior-guided (MPG) adaptation. Reg follows the same strategy without $\mathcal{L}_{\mathrm{tok}}$.

5 Experiments

5.1 Experimental Setup

Benchmarks and inputs. We use **mmSimPrior-Sim** as the AMASS-driven simulated radar–motion corpus for pretraining, and evaluate on two real-world benchmarks. **mmSimPrior-Real** denotes our collected real-world benchmark with three indoor environments. We further evaluate on the public **RT-Pose** benchmark (Ho et al. 2024) as an external validation dataset. Because RT-Pose provides only 3D keypoints, we derive pseudo-SMPL-X annotations by fitting RGB-based human mesh recovery estimates (Wang et al. 2025; Yang et al. 2026) to the provided keypoints. Unless otherwise specified, all methods are pretrained on mmSimPrior-Sim and evaluated on real radar heatmaps under zero-shot or limited-data finetuning protocols.

mmSimPrior-Real annotations. mmSimPrior-Real contains recordings from three indoor environments (Hall, Lab, and Studio), covering six volunteers, 40 motion categories, and three sensing-distance ranges. A synchronized stereo-camera system is used only to derive pseudo-SMPL-X annotations; radar heatmaps are the sole model input.

No-overlap protocol and metrics. On mmSimPrior-Real, we evaluate using a No-Overlap Setting (NOS): no finetuning-test sequence pair shares the same complete setting across subject identity, environment, sensing location, and motion category. Unless otherwise specified, limited finetuning uses 24 real sequences; Zero-shot denotes evaluation without any annotation, paired supervision, or model selection from the evaluated benchmark. We report MPJPE and PA-MPJPE, W-MPJPE when evaluating world-space localization, and MPVPE for mesh-level evaluation. Avg. denotes a sequence-weighted average.

Baselines. We compare against adapted radar-based human motion reconstruction baselines, including mmDiff†(Fan et al. 2024), RAPTR† (Kato et al. 2025), and mmGPE† (Xue et al. 2023). For fairness, all adapted baselines use the same four-modality input, simulated pretraining corpus, finetuning budget, evaluation splits, and metrics as mmSimPrior. Baseline adaptation details, metric definitions, averaging formula, and split construction are provided in the supplementary material.

Method	Hall			Lab			Studio		
	MPJPE↓	PA-MPJPE↓	W-MPJPE↓	MPJPE↓	PA-MPJPE↓	W-MPJPE↓	MPJPE↓	PA-MPJPE↓	W-MPJPE↓
mmDiff†	104.0	74.7	173.4	102.5	78.6	160.2	104.8	80.4	167.4
RAPTR†	115.9	87.1	174.8	107.8	85.1	171.0	118.3	93.3	172.9
mmGPE†	106.4	75.5	147.0	104.7	76.0	144.9	101.9	74.7	135.6
mmSimPrior-Reg	**75.9**	**54.7**	**116.0**	**77.2**	**57.0**	**116.0**	**62.2**	**49.5**	**95.0**
mmSimPrior-Cls	89.5	65.5	125.6	92.4	68.5	127.5	86.2	65.8	114.2

Table 1: **Same-environment 24-sequence NOS finetuning on mmSimPrior-Real.** All methods use the same four-modality radar heatmap input, mmSimPrior-Sim pretraining corpus, finetuning budget, and NOS split. The best / second best results are in **boldface**, and underlined, respectively. Metrics are in millimeters; lower is better. † denotes adapted baselines pretrained on mmSimPrior-Sim.

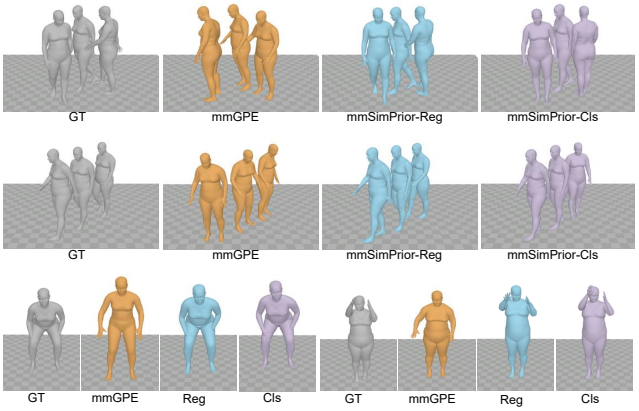


Figure 5: **Qualitative comparison on real-world radar sequences.** We compare mmGPE, mmSimPrior-Reg, and mmSimPrior-Cls with the SMPL-X ground truth.

5.2 Overall Result

Performance on mmSimPrior-Real. Table 1 evaluates 24-sequence NOS finetuning within each environment. Although adaptation and testing occur in the same environment, the support set contains only about one minute of paired data and shares no exact subject–location–motion configuration with the test set. mmSimPrior-Reg achieves the best result across all three metrics and environments, while mmSimPrior-Cls consistently ranks second. Relative to the strongest adapted MPJPE baseline, Reg reduces error by 27.0%, 24.7%, and 39.0% in Hall, Lab, and Studio, respectively. The consistent improvements in MPJPE, PA-MPJPE, and W-MPJPE show that the learned priors improve both articulated pose recovery and world-space localization under minute-level adaptation. The stronger Reg results further indicate that continuous decoding can effectively absorb target-domain corrections from limited real supervision.

Figure 5 shows that both mmSimPrior variants recover motion progression and local articulation more faithfully than mmGPE. Qualitatively, they also maintain more stable foot–ground contact with substantially less foot sliding. Beyond frame-level reconstruction accuracy, the optional motion-prior refinement acts as a targeted temporal regularizer for Reg. It reduces joint-velocity, acceleration, and jerk errors by 7.7%, 18.0%, and 36.7%, respectively, while changing all spatial reconstruction metrics by at most 1.3% (Table 8 in the supplementary material).

Method	MPJPE↓	PA-MPJPE↓	MPVPE↓
mmDiff†	126.0	81.6	165.8
RAPTR†	131.0	89.3	171.0
mmGPE†	126.1	82.5	173.0
mmSimPrior-Reg	**116.9**	**78.1**	**153.2**
mmSimPrior-Cls	119.9	80.8	156.2

Table 2: **Minute-level external validation on RT-Pose (Ho et al. 2024).** The best / second best results are in **boldface**, and underlined, respectively. Metrics are in millimeters; lower is better.

External validation on RT-Pose. Table 2 uses the same minute-level support budget for all methods. mmSimPrior-Reg achieves the best MPJPE, PA-MPJPE, and MPVPE, improving over the strongest baseline by 9.1, 3.5, and 12.6 mm, respectively, while mmSimPrior-Cls ranks second. Because RT-Pose uses a different radar platform, these gains also support cross-device transfer under limited adaptation. Figure 6 further shows improved action phase and foot placement; Reg favors fine-grained accuracy, whereas Cls suppresses implausible intermediate poses through the learned motion codebook.

Figure 6 shows that these advantages persist under the RT-Pose domain shift. Both mmSimPrior variants better preserve the temporal action phase and maintain more stable foot placement, resulting in visibly less foot sliding. Reg more closely matches fine-grained limb configurations, while Cls avoids implausible intermediate poses by decoding through the learned motion codebook. Together with Table 2, these results show that the learned priors improve both reconstruction accuracy and temporal plausibility.

5.3 Generalization Across Unseen Environments

We next study robustness to environment shift by evaluating transfer across previously unseen real-world environments.

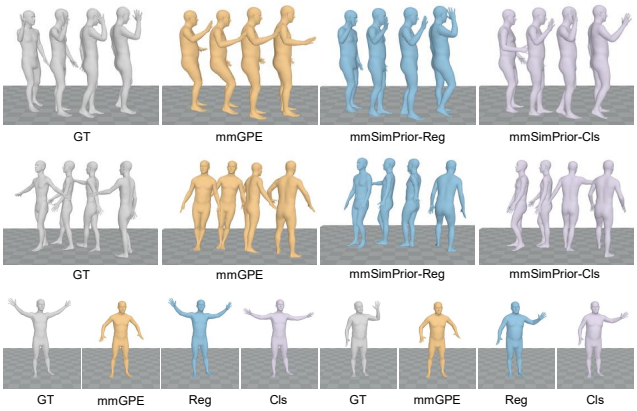


Figure 6: **Qualitative comparison on RT-Pose.** We compare mmGPE, mmSimPrior-Reg, and mmSimPrior-Cls with pseudo-SMPL-X Ground Truth.

Method	Zero-shot Avg.		24-seq finetune	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
mmDiff†	130.0	97.2	117.1	87.1
RAPTR†	170.2	104.0	119.0	90.4
mmGPE†	141.3	97.0	111.8	81.0
mmSimPrior-Reg	119.1	93.1	**96.7**	**73.3**
mmSimPrior-Cls	**118.9**	**89.7**	100.0	75.2

Table 3: **Unseen-environment transfer on mmSimPrior-Real.** Zero-shot reports sequence-weighted averages over all splits without finetuning. The 24-seq columns report Hall→Lab cross-environment finetuning, where models are finetuned on Hall and evaluated on the no-overlap Lab split. The best / second best results are in **boldface**, and underlined, respectively. Metrics are in millimeters; lower is better.

Compared with same-environment evaluation, this setting is more challenging because the test environment can introduce unseen clutter, multipath patterns, and reflective structures.

Unseen-environment performance. Table 3 evaluates two deployment regimes: direct zero-shot transfer and Hall→Lab adaptation without using any Lab sequence for finetuning. In the zero-shot setting, mmSimPrior-Cls achieves the best results, reducing MPJPE / PA-MPJPE over the strongest baseline by 8.5% / 7.5%. Its frozen codebook and motion decoder constrain uncertain radar features to plausible joint-temporal patterns when target-domain statistics are unavailable. After finetuning on 24 Hall sequences, mmSimPrior-Reg becomes the strongest mode and reduces the two errors by 13.5% / 9.5%. Continuous decoding is less constrained by the discrete motion space and therefore better exploits limited real supervision for domain-specific pose correction. The switch between the two modes validates their complementary roles: Cls is preferable without target-benchmark adaptation, whereas Reg provides higher accuracy when a small real support set is available.

Mode	Sim.	DR	MPG	Zero-shot MPJPE	Zero-shot PA	24-seq finetune MPJPE	24-seq finetune PA
Reg	×	×	–	–	–	130.2	90.9
	✓	×	–	140.6	90.5	126.6	85.4
	✓	✓	–	**129.4**	**85.4**	**75.9**	**54.7**
Cls	×	×	×	–	–	120.8	88.7
	✓	×	×	159.5	113.1	101.7	77.1
	✓	✓	×	**123.9**	**84.4**	94.9	66.6
	✓	✓	✓	**123.9**	**84.4**	**89.5**	**65.5**

Table 4: **Ablation on the Hall split of mmSimPrior-Real.** Sim. denotes pretraining on mmSimPrior-Sim, DR denotes physics-informed domain randomization during simulated pretraining, and MPG denotes motion-prior-guided adaptation. The best / second best results are in **boldface**, and underlined, respectively. Metrics are in millimeters; lower is better.

5.4 Ablation Studies

Table 4 isolates the effects of simulation pretraining, physics-informed domain randomization (DR), and Motion-Prior-Guided (MPG) adaptation. Simulation pretraining and DR are complementary: the former supplies broad paired motion coverage, whereas the latter prevents the learned signal and mapping priors from overfitting simulation-specific response statistics. Clean simulation pretraining improves finetuned MPJPE / PA-MPJPE by 2.8% / 6.1% for Reg and 15.8% / 13.1% for Cls, while enabling zero-shot inference. Adding DR produces the dominant gains: 8.0% / 5.6% zero-shot and 40.0% / 35.9% finetuned improvements for Reg, and 22.3% / 25.4% zero-shot improvements for Cls. Thus, synthetic scale alone is insufficient; DR makes the learned simulation priors transferable. Finally, MPG improves finetuned Cls by a further 5.7% / 1.7%, confirming the benefit of frozen-tokenizer supervision.

6 Conclusion

We presented **mmSimPrior**, a sim-to-real framework for data-efficient radar-based human motion reconstruction. By shifting paired supervision from scarce real recordings to large-scale simulated radar–motion data, mmSimPrior factorizes simulation-derived knowledge into transferable signal, motion, and mapping priors. Physics-informed domain randomization improves their robustness to real-world propagation and acquisition shifts, while the classification- and regression-based mapping modes provide complementary trade-offs between constrained zero-shot reconstruction and flexible real-data adaptation. The motion prior further guides limited-data Cls adaptation through frozen-tokenizer supervision. Experiments under the No-Overlap Setting, together with external validation on RT-Pose, demonstrate consistent transfer across same-environment, unseen-environment, and cross-dataset settings using only minute-level real data. Future work will extend this framework to more complex human scenes and broader real-world sensing conditions.

References

An, S.; Li, Y.; and Ogras, U. 2022. mri: Multi-modal 3d human pose estimation dataset using mmwave, rgb-d, and inertial sensors. *Advances in neural information processing systems*, 35: 27414–27426.

An, S.; and Ogras, U. Y. 2021. Mars: mmwave-based assistive rehabilitation system for smart healthcare. *ACM Transactions on Embedded Computing Systems (TECS)*, 20(5s): 1–22.

Chen, A.; Wang, X.; Zhu, S.; Li, Y.; Chen, J.; and Ye, Q. 2022. mmbody benchmark: 3d body reconstruction dataset and analysis for millimeter wave radar. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3501–3510.

Chen, V. C. 2011. *The Micro-Doppler Effect in Radar*. Artech House.

Chen, X.; and Zhang, X. 2023. Rf genesis: Zero-shot generalization of mmwave sensing through simulation-based data synthesis and generative diffusion models. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*, 28–42.

Deng, K.; Zhao, D.; Han, Q.; Zhang, Z.; Wang, S.; Zhou, Y.; and Ma, Y. 2023. Midas: Generating mmWave Radar Data from Videos for Training Pervasive and Privacy-preserving Human Sensing Tasks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(1): 1–26.

Dwivedi, S. K.; Sun, Y.; Patel, P.; Feng, Y.; and Black, M. J. 2024. Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1323–1333.

Fan, J.; Yang, J.; Xu, Y.; and Xie, L. 2024. Diffusion model is a good pose estimator from 3d rf-vision. In *European Conference on Computer Vision*, 1–18. Springer.

Fan, J.; Zhou, Y.; Yang, Y.; Cui, X.; Zhang, J.; Xie, L.; Yang, J.; Lu, C. X.; and Ding, F. 2026. M4human: A large-scale multimodal mmwave radar benchmark for human mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 42836–42846.

Goel, S.; Pavlakos, G.; Rajasegaran, J.; Kanazawa, A.; and Malik, J. 2023. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14783–14794.

Ho, Y.-H.; Cheng, J.-H.; Kuan, S. Y.; Jiang, Z.; Chai, W.; Huang, H.-W.; Lin, C.-L.; and Hwang, J.-N. 2024. Rt-pose: A 4d radar tensor-based 3d human pose estimation and localization benchmark. In *European Conference on Computer Vision*, 107–125. Springer.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huang, T.; Ding, H.; Sun, W.; Zhao, C.; Wang, G.; Wang, F.; Zhao, K.; Wang, Z.; and Xi, W. 2025. One Snapshot is All You Need: A Generalized Method for mmWave Signal Generation. In *IEEE INFOCOM 2025-IEEE Conference on Computer Communications*, 1–10. IEEE.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7): 1325–1339.

Jiang, Z.; Wu, Y.; Zhou, Z.; Wang, W.; and Han, J. 2026. PODM-mmHPE: Exploiting Physical Optics based mmWave Data Augmentation via Diffusion Model for Generalizable Human Pose Estimation. *IEEE Internet of Things Journal*.

Joshi, V.; Xu, S.; Cao, Q.; Zhu, Y.; Wang, P.; and Xue, H. 2024. Towards Robust mmWave-based Human Activity Recognition using Large Simulated Dataset for Model Pretraining. In *2024 IEEE International Conference on Big Data (BigData)*, 7332–7337. IEEE.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-End Recovery of Human Shape and Pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Kato, S.; Yataka, R.; Wang, P. P.; Miraldo, P.; Fujihashi, T.; and Boufounos, P. 2025. RAPTR: Radar-based 3D Pose Estimation using Transformer. *arXiv preprint arXiv:2511.08387*.

Kocabas, M.; Athanasiou, N.; and Black, M. J. 2020. VIBE: Video Inference for Human Body Pose and Shape Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5253–5263.

Kolotouros, N.; Pavlakos, G.; Black, M. J.; and Daniilidis, K. 2019. 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*, 2252–2261.

Lai, Z.; Yang, J.; Xia, S.; Lin, L.; Sun, L.; Wang, R.; Liu, J.; Wu, Q.; and Pei, L. 2026. Radarllm: Empowering large language models to understand human motion from millimeter-wave point cloud sequence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 5791–5799.

Lee, S.-P.; Kini, N. P.; Peng, W.-H.; Ma, C.-W.; and Hwang, J.-N. 2023. Hupr: A benchmark for human pose estimation using millimeter wave radar. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5715–5724.

Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of Motion Capture as Surface Shapes. In *International Conference on Computer Vision*, 5442–5451.

Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; and Theobalt, C. 2017. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, 506–516. IEEE.

Rahman, M. M.; Yataka, R.; Kato, S.; Wang, P.; Li, P.; Cardace, A.; and Boufounos, P. 2024. MMVR: Millimeter-wave multi-view radar dataset and benchmark for indoor perception. In *European Conference on Computer Vision*, 306–322. Springer.

Richards, M. A. 2014. *Fundamentals of Radar Signal Processing*. McGraw-Hill Education, 2 edition.

Saleem, M. U.; Pinyoanuntapong, E.; Wang, P.; Xue, H.; Das, S.; and Chen, C. 2025. Genhmr: Generative human mesh recovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6749–6757.

Sengupta, A.; Jin, F.; Zhang, R.; and Cao, S. 2020. mm-Pose: Real-Time Human Skeletal Posture Estimation Using mmWave Radars and CNNs. *IEEE Sensors Journal*, 20(17): 10032–10044.

Singh, A. D.; Sandha, S. S.; Garcia, L.; and Srivastava, M. 2019. Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar. In *Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems*, 51–56.

Tesch, J.; Becherini, G.; Achar, P.; Yiannakidis, A.; Kocabas, M.; Patel, P.; and Black, M. J. 2025. BEDLAM2. 0: Synthetic humans and cameras in motion. *arXiv preprint arXiv:2511.14394*.

van den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems*, volume 30.

Von Marcard, T.; Henschel, R.; Black, M. J.; Rosenhahn, B.; and Pons-Moll, G. 2018. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, 601–617.

Wang, Y.; Sun, Y.; Patel, P.; Daniilidis, K.; Black, M. J.; and Kocabas, M. 2025. Prompthmr: Promptable human mesh recovery. In *Proceedings of the computer vision and pattern recognition conference*, 1148–1159.

Wu, Z.; Zhang, D.; Xie, C.; Yu, C.; Chen, J.; Hu, Y.; and Chen, Y. 2022. RFMask: A simple baseline for human silhouette segmentation with radio signals. *IEEE Transactions on Multimedia*, 25: 4730–4741.

Xue, H.; Cao, Q.; Ju, Y.; Hu, H.; Wang, H.; Zhang, A.; and Su, L. 2022. M4esh: mmWave-based 3D Human Mesh Construction for Multiple Subjects. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, 391–406.

Xue, H.; Cao, Q.; Miao, C.; Ju, Y.; Hu, H.; Zhang, A.; and Su, L. 2023. Towards generalized mmwave-based human pose estimation through signal augmentation. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, 1–15.

Xue, H.; Ju, Y.; Miao, C.; Wang, Y.; Wang, S.; Zhang, A.; and Su, L. 2021. mmMesh: Towards 3D real-time dynamic human mesh construction using millimeter-wave. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, 269–282.

Yan, Y.; Yang, S.; Guo, X.; Wang, X.; Chow, W.; Shu, Y.; and He, S. 2025. mmExpert: Integrating Large Language Models for Comprehensive mmWave Data Synthesis and Understanding. In *Proceedings of the Twenty-sixth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, 1–10.

Yang, J.; Huang, H.; Zhou, Y.; Chen, X.; Xu, Y.; Yuan, S.; Zou, H.; Lu, C. X.; and Xie, L. 2023. Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing. *Advances in Neural Information Processing Systems*, 36: 18756–18768.

Yang, X.; Kukreja, D.; Pinkus, D.; Fan, T.; Park, J.; Shin, S.; Cao, J.; Liu, J.-W.; Ugrinovic, N.; Sagar, A.; et al. 2026. Sam 3d body: Robust full-body human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7209–7219.

Zhang, X.; Li, Z.; and Zhang, J. 2022. Synthesized millimeter-waves for human motion sensing. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, 377–390.

Zhao, M.; Li, T.; Abu Alsheikh, M.; Tian, Y.; Zhao, H.; Torralba, A.; and Katabi, D. 2018a. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7356–7365.

Zhao, M.; Liu, Y.; Raghu, A.; Li, T.; Zhao, H.; Torralba, A.; and Katabi, D. 2019. Through-wall human mesh recovery using radio signals. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10113–10122.

Zhao, M.; Tian, Y.; Zhao, H.; Alsheikh, M. A.; Li, T.; Hristov, R.; Kabelac, Z.; Katabi, D.; and Torralba, A. 2018b. RF-based 3D skeletons. In *Proceedings of the 2018 conference of the ACM special interest group on data communication*, 267–281.

A Implementation Details

mmSimPrior operates on 32-frame clips at 10 Hz and receives one Range–Doppler (RD) heatmap together with front, bird's-eye, and side MVDR heatmaps. Training comprises signal-prior pretraining, motion-prior pretraining, radar-conditioned mapping pretraining, and limited-data real-world adaptation.

Multi-modal signal prior. RD and bird's-eye inputs have resolution 128×128; front and side inputs have resolution 128×50 and are symmetrically padded to 128×64. Patch sizes are 16×16 for RD and bird's-eye views and 16×8 for front and side views, yielding 64 tokens per modality and 256 tokens per frame. The masked autoencoder uses a 12-layer shared Transformer encoder (width 768, 12 heads) and four-layer modality-specific decoders (width 512, eight heads). We independently mask 75% of patches in each modality and optimize the mean masked-patch reconstruction loss for 200K iterations with AdamW, batch size 256, learning rate 2×10^{-5}, and weight decay 0.05.

Joint-temporal motion prior. The motion tokenizer receives 32 frames of 6D rotations for 22 SMPL-X body joints and encodes them into an 8×40 joint-temporal latent lattice containing 320 motion tokens. Each token is quantized using an EMA-updated codebook with 2,048 entries of dimension 256. The encoder and decoder use width 512 and two residual blocks. The tokenizer objective is $20\mathcal{L}_{\text{rot}} + 10\mathcal{L}_{\text{vel}} + 100\mathcal{L}_{\text{joint}} + \mathcal{L}_{\text{com}}$, where the terms supervise 6D rotations, temporal differences, SMPL-X joint geometry, and codebook commitment, respectively. We train the tokenizer for 200K iterations using AdamW with batch size 256 and learning rate 2×10^{-4}. The entire motion tokenizer is frozen in all subsequent stages.

Radar-conditioned mapping prior. Two stride-two temporal reductions compress the 32-frame signal features into eight temporal groups. A four-layer temporal Transformer and a shared six-layer cross-attention decoder aggregate the radar context. Cls predicts 40 codebook distributions per temporal group (320 in total) and decodes their probability-weighted embeddings through the frozen motion decoder, whereas Reg directly predicts 32-frame 6D rotations. Both modes also predict framewise root translation and one sequence-level ten-dimensional shape vector. Mapping losses supervise rotation, root-relative joints, translation, shape, translation velocity, and translation acceleration with weights 1, 50, 1, 0.5, 2, and 0.5, respectively. We train for 100K iterations using AdamW with batch size 32,learning rate 10^{-4} for the mapping prior, and 10^{-5} for the signal encoder.

Real-world adaptation and inference. The pretrained signal-encoder weights remain frozen except for LoRA modules inserted into its attention and feed-forward layers (rank $r = 16$, LoRA alpha $\alpha = 32$, dropout 0.1). The mapping modules and prediction heads remain trainable, while the motion tokenizer is frozen. AdamW uses learning rates 10^{-4} for the mapping modules and 3×10^{-5} for LoRA, weight decay 10^{-4}, and effective batch size eight. For Cls with motion-prior-guided adaptation, the frozen tokenizer converts reference rotations into code targets, and the token cross-entropy weight is linearly warmed to 0.05 over the first 500 steps.

mmSimPrior-Real adaptation uses 24 support sequences, a fixed 4K-iteration horizon, and seed 2026. RT-Pose uses the same support budget and effective batch size, with a common 8K horizon for all methods. All reported checkpoints are evaluated over all 32 target frames without target-test checkpoint selection. On mmSimPrior-Real, reported Reg results include evaluation-time motion-prior refinement (MPR) unless direct and refined outputs are explicitly compared.

Physics-informed domain randomization. Propagation-level perturbations comprise multipath ghosting and spatial response spreading, while acquisition-level perturbations comprise response-statistics alignment and spatial-frequency filtering. Ghosting is applied to the bird's-eye and side views with probability 0.5. Gaussian spreading is applied to the front view with probability 0.5. The perturbation strength is increased progressively during pretraining.

Response alignment performs per-frame histogram matching to fixed per-modality reference histograms estimated once from an unlabeled Hall recording that does not appear in any evaluation manifest. The histograms contain radar intensities only; no RT-Pose data or pose, mesh, or motion-category annotations are used. Alignment is applied independently to each modality with probability 0.5. Spatial-frequency filtering uses a radial low-pass mask with cutoff 0.15 cycles/pixel and probabilities 0.3 during signal-prior pretraining and 0.5 during mapping pretraining. Random strengths are shared across frames of a sequence to preserve temporal consistency. All perturbations are disabled during real-world adaptation and evaluation, and the same configuration is fixed across benchmarks.

All models were implemented in PyTorch and trained on NVIDIA A6000 GPUs.

B Radar Simulation and Signal Processing

B.1 IF Signal Simulation from Dynamic Human Meshes

We generate mmSimPrior-Sim by driving a physics-based FMCW radar simulator with AMASS motion sequences. Each sequence is represented as a temporally varying SMPL-X mesh and resampled to 10 FPS. At every radar frame, mesh facets are characterized by their barycenters, surface normals, and areas. Back-facing and occluded facets are removed using viewpoint-conditioned visibility filtering, including Hidden Point Removal over the facet locations.

For each retained facet and transmit–receive antenna pair, the simulator computes a bistatic propagation path over the sampled frequencies. Reflection strength depends on the transmit–facet and facet–receive path lengths, phase delay, geometric attenuation, facet area, and surface orientation. The facet responses are coherently accumulated to synthesize the complex IF signal according to the bistatic reflection model in the main paper.

Each AMASS motion is rendered at three radar–subject distances, introducing variation in attenuation, angular resolution, and the spatial distribution of human reflections. The resulting complex IF sequences are paired with aligned

SMPL-X pose, root translation, and body-shape annotations. The simulated radar observations supervise the signal and mapping priors, while the corresponding AMASS trajectories are also used to train the motion prior.

B.2 Heatmap Construction from IF Signals

The simulated and real IF signals are converted into a common four-modality radar representation consisting of one Range–Doppler heatmap and three orthogonal MVDR spatial heatmaps. Applying the same signal-processing pipeline to simulation, mmSimPrior-Real, and RT-Pose prevents representation-specific differences from confounding the transfer evaluation.

Range–Doppler heatmap. Let $\mathbf{X}_t \in \mathbb{C}^{N_{\mathrm{tx}} \times N_{\mathrm{rx}} \times N_{\mathrm{c}} \times N_{\mathrm{s}}}$ denote the IF tensor at radar frame t, where N_{tx} and N_{rx} are the numbers of transmit and receive antennas, N_{c} is the number of chirps, and N_{s} is the number of ADC samples per chirp. After arranging the transmit–receive pairs into $N_{\mathrm{ant}} = N_{\mathrm{tx}} N_{\mathrm{rx}}$ virtual antenna channels, we apply a windowed Range-FFT along the fast-time dimension and a Doppler-FFT along the slow-time dimension. The resulting Range–Doppler response is

$$\mathbf{H}_t^{\mathrm{RD}}[k, f] = \left| \frac{1}{N_{\mathrm{ant}}} \sum_{a=1}^{N_{\mathrm{ant}}} \sum_{l=0}^{N_{\mathrm{c}}-1} \mathbf{Z}_t[a, l, k] e^{-j2\pi f l / N_{\mathrm{c}}} \right|, \tag{S1}$$

where $\mathbf{Z}_t$ is the range-domain response and k and f index the range and Doppler bins. The RD modality preserves radial-velocity evidence complementary to the spatial MVDR views.

Multi-view MVDR spatial heatmaps. For each range bin, we estimate the antenna-domain covariance and evaluate the MVDR spectrum $P_t(\mathbf{p})$ over a discretized 3D spatial grid $\mathbf{p} = (x, y, z)$. Rather than directly processing the full response volume, we compute three orthogonal mean projections:

$$\begin{aligned} \mathbf{H}_t^{\mathrm{front}}(x, z) &= \mathbb{E}_y[P_t(x, y, z)] \,, \\ \mathbf{H}_t^{\mathrm{bird}}(x, y) &= \mathbb{E}_z[P_t(x, y, z)] \,, \\ \mathbf{H}_t^{\mathrm{side}}(y, z) &= \mathbb{E}_x[P_t(x, y, z)] \,. \end{aligned} \tag{S2}$$

The front and side views have spatial resolution 128×50, while the bird's-eye and RD heatmaps have resolution 128×128. All modalities are log-compressed and independently normalized per frame. Together, the RD heatmap and three MVDR projections form the common four-modality observation used for simulated pretraining, real-world adaptation, and evaluation.

C Dataset and Evaluation Protocol

mmSimPrior dataset suite. The mmSimPrior dataset suite consists of **mmSimPrior-Sim**, a large-scale simulated pretraining corpus, and **mmSimPrior-Real**, a paired real-world benchmark for evaluating zero-shot and limited-data transfer. mmSimPrior-Sim is generated from AMASS motions spanning 344 mocap subjects and contains approximately 4.0M frames from 30K radar–motion sequences. mmSimPrior-Real contains 1,126 synchronized radar–motion recordings collected from six participants across three indoor environments, 40 motion categories, and three sensing-distance ranges. Together, the two components contain 4.2M frames, 31K sequences, and 350 subjects.

mmSimPrior-Real collection. As illustrated in Fig. S1, we collect synchronized mmWave radar and stereo-camera observations in three indoor environments: Hall, Lab, and Studio. The environments differ in room geometry, background clutter, reflective surfaces, and radar placement, resulting in distinct multipath and acquisition conditions. The radar and stereo camera are rigidly mounted on the same tripod and temporally synchronized. We use a Texas Instruments AWR1843BOOST FMCW radar configured with three transmit and four receive channels. Raw complex ADC data are captured through a DCA1000EVM at 10 FPS, using 128 chirps per frame and 256 ADC samples per chirp, and are processed into one RD and three orthogonal MVDR heatmaps following Sec. B.2 The stereo observations are used only to construct reference SMPL-X annotations; all reconstruction models receive radar heatmaps alone during adaptation and evaluation.

The complete benchmark contains 1,126 unique recordings. For each environment, 24 recordings form a fixed support set for limited-data adaptation, while the remaining 1,054 recordings form the evaluation population. The support and evaluation recordings are sequence-disjoint.

All data-collection procedures involving human participants were reviewed and approved by our Institutional Review Board. Each participant provided informed consent and was informed of the sensing devices, collection procedure, and intended research use.

C.1 Evaluation Protocol Details

Metrics. We report MPJPE over the 22 SMPL-X body joints and PA-MPJPE after framewise similarity alignment. W-MPJPE evaluates absolute body placement in the world coordinate system, while MPVPE measures the mean error over the SMPL-X body vertices. Because the real-world mesh references are obtained through camera-based reconstruction and fitting, MPVPE is interpreted as a mesh-level reconstruction metric rather than an isolated measure of body-shape accuracy. All errors are reported in millimeters.

Averaging. Unless otherwise specified, results are averaged uniformly over valid test recordings. Each recording therefore contributes equally, irrespective of its number of frames. When results are grouped by environment, sensing distance, or motion category, the reported aggregate is weighted by the number of valid recordings in each group.

NOS split construction. For each real-world recording, we define its sensing configuration by four factors: subject identity, environment, sensing location, and motion category. The No-Overlap Setting (NOS) requires that no adaptation–test pair share the same complete four-factor configuration. Individual factors may overlap, but every test recording must

Task	Dataset	Data	#Subj	#Seq	#Frame	Sim	Public
HMR	MPI-INF-3DHP (Mehta et al. 2017)	RGB	8	16	1.3M	×	✓
	Human3.6M (Ionescu et al. 2013)	RGB	11	210	3.6M	×	✓
	BEDLAM2.0 (Tesch et al. 2025)	RGB	>4K	>27K	>8M	✓	✓
HPE	RadHAR (Singh et al. 2019)	PC	2	16	167K	×	✓
	mm-Pose (Sengupta et al. 2020)	PC	2	8	40K	×	×
	MARS (An and Ogras 2021)	PC	4	80	40K	×	✓
	mRI (An, Li, and Ogras 2022)	PC	20	300	160K	×	×
	MM-Fi (Yang et al. 2023)	PC	40	1K	320K	×	✓
	RF-Pose (Zhao et al. 2018a)	HM	100	–	–	×	×
	HuPR (Lee et al. 2023)	HM	6	235	141K	×	✓
HMR	mmMesh (Xue et al. 2021)	PC	20	160	480K	×	×
	mmBody (Chen et al. 2022)	PC	20	48	<70K	×	✓
	HIBER (Wu et al. 2022)	HM	–	–	179K	×	Partial
	MMVR (Rahman et al. 2024)	HM	20	395	345K	×	✓
	RT-Pose (Ho et al. 2024)	HM	–	240	72K	×	✓
	RF-Genesis (Chen and Zhang 2023)	PC	–	–	–	✓	×
	mmGPE (Xue et al. 2023)	HM	4	34	276K	✓	×
	mmSimPrior-Sim (Ours)	**HM**	**344**	**30K**	**4.0M**	✓	✓
	mmSimPrior-Real (Ours)	**HM**	**6**	**1,126**	**162K**	×	✓

Table S1: **Comparison of RGB- and radar-based human motion datasets by task, sensing representation, and scale.** HMR denotes human motion reconstruction, PC denotes mmWave radar point clouds, and HM denotes radar heatmaps. Sequence counts follow the definitions of the original datasets and are therefore not directly comparable across datasets. **mmSimPrior-Sim** provides large-scale paired radar–motion supervision generated from AMASS, whereas **mmSimPrior-Real** provides a real-world benchmark for zero-shot and minute-level adaptation. Together, they form the mmSimPrior dataset suite with 4.2M frames and 31K sequences.

differ from every adaptation recording in at least one factor. This prevents evaluation on another temporal segment of an identical subject–environment–location–motion configuration.

For same-environment adaptation, we use 24 recordings from the target environment and evaluate on the remaining recordings satisfying NOS. For cross-environment adaptation, the adaptation and evaluation environments differ, and the source–target direction is specified in the corresponding table. Zero-shot denotes evaluation without pose annotations, paired radar–motion supervision, adaptation, or model selection from the evaluated population. The response-alignment histograms are estimated once from a separate unlabeled Hall recording excluded from the evaluation manifest and are fixed for all experiments.

Baseline adaptation. The compared methods were originally developed with different radar representations and temporal protocols. For a controlled comparison, we adapt each baseline to the common four-modality observation consisting of one Range–Doppler heatmap and three orthogonal MVDR views, while preserving its defining temporal modeling and pose-prediction mechanism. All methods use the same simulation corpus, 24-sequence real-data budget, fixed adaptation horizon, NOS split, and evaluation metrics.

Evaluation covers the same 32-frame target window at 10 FPS. Methods with a shorter native temporal context process the window in contiguous segments, and metrics are aggregated over all 32 target frames. The † marker denotes baseline variants pretrained under this unified simulated-pretraining and evaluation protocol.

C.2 External Validation on RT-Pose

We additionally evaluate on the public RT-Pose benchmark (Ho et al. 2024), which provides synchronized radar ADC measurements, radar tensors, stereo RGB observations, LiDAR point clouds, and 3D body-joint annotations across diverse indoor and outdoor scenes. We focus on its single-person recordings. Starting from the released ADC signals, we apply the same signal-processing pipeline used for mmSimPrior-Real to generate one Range–Doppler heatmap and three orthogonal MVDR views. All compared methods therefore receive the same four-modality radar input and are evaluated over 32-frame windows at 10 FPS.

Pseudo-annotation consistency audit. We assess the pseudo-SMPL-X references using 12 common limb landmarks: the bilateral hips, knees, ankles, shoulders, elbows, and wrists. Because the released 3D keypoints and left-view 2D detections participate in fitting, these diagnostics measure fitting and image-level consistency rather than independent annotation accuracy.

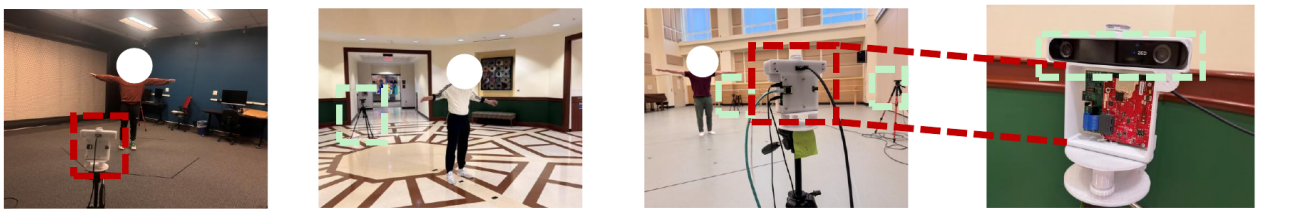


Figure S1: **Real-world data collection setup.** We collect synchronized mmWave radar and stereo-camera observations in three indoor environments: (a) Lab, (b) Hall, and (c) Studio. The environments differ in spatial layout, clutter, and reflective structures, producing distinct multipath and sensing conditions. (d) The radar and stereo camera are rigidly mounted on the same tripod. Stereo images are used only to construct reference SMPL-X annotations; all evaluated models use radar heatmaps alone.

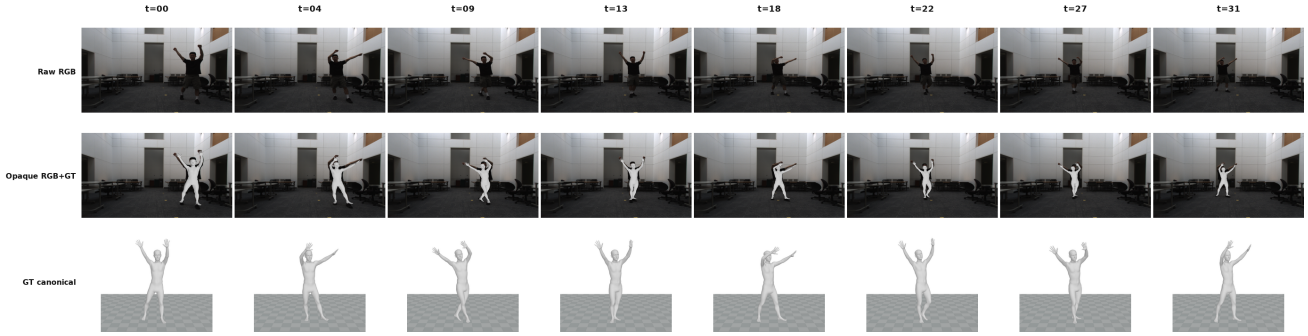


Figure S2: **Pseudo-SMPL-X annotation examples on RT-Pose.** Rows show the raw RGB observation, the fitted SMPL-X overlay, and the canonicalized reference mesh across the sequence.

Lower-body joints differ from the official 3D keypoints by 39.6 mm on average, with hip PCK@50 above 80%. Median reprojection error is 17.0 pixels in the fitting view and 24.1 pixels in the held-out right view. Upper-body residuals are larger (140.8 mm), primarily because RT-Pose uses surface-level shoulder landmarks whereas SMPL-X represents kinematic shoulder centers; the resulting systematic offset propagates along the arm chain. Despite this convention-sensitive absolute residual, the fitted trajectories remain image-consistent, and direct evaluation against the official keypoints preserves the method ranking and nearly the same Reg margin (Sec. D.8). We therefore use these references for shared comparative mesh evaluation, rather than interpreting their absolute errors as metric ground-truth accuracy.

Temporal regularization produces consistent SMPL-X pose, shape, and root-translation trajectories. We exclude sequences with severe occlusion, low illumination, or fitting divergence based solely on the quality of the RGB annotation pipeline, before evaluating any radar model; 111 sequences pass this screening. RGB is used only for offline reference construction, while all adaptation and evaluation use radar heatmaps alone.

D Additional Experimental Results

We complement the main evaluation with analyses of data efficiency, held-out factors, mesh and temporal quality, and tokenizer capacity. Unless otherwise specified, errors are reported in millimeters and lower values are better. The best and second-best results are shown in **boldface** and underlined, respectively. All experiments follow Sec. C.1. On mmSimPrior-Real, mmSimPrior-Reg includes evaluation-time motion-prior refinement (MPR) unless direct and refined outputs are explicitly compared.

D.1 Data-Efficient Real-World Adaptation

We evaluate mmSimPrior-Reg and mmSimPrior-Cls with 0, 1, 4, 8, and 24 real-world adaptation sequences; 0 denotes zero-shot evaluation. The same-environment setting adapts and evaluates in Hall under NOS, whereas the cross-environment setting adapts on Hall and evaluates on Lab. For each nonzero budget, all three optimization seeds use the same support subset.

Fig. S3 shows complementary behavior. Cls provides the stronger Hall-to-Lab zero-shot initialization, whereas Reg improves faster once several labeled sequences are available and achieves the lowest final errors. The mildly non-monotonic one-sequence Cls result indicates instability only in the most data-scarce regime; the trend becomes consistent from four sequences.

Adaptation stability. Because limited-data adaptation uses only 24 support sequences, we verify that the reported results are not an artifact of a particular adaptation run. Table S2 repeats the 24-sequence adaptation three times per environment. Every resulting checkpoint differs, but the errors do not: the largest standard deviation across all environments and metrics is 0.36 mm, two orders of magnitude below the

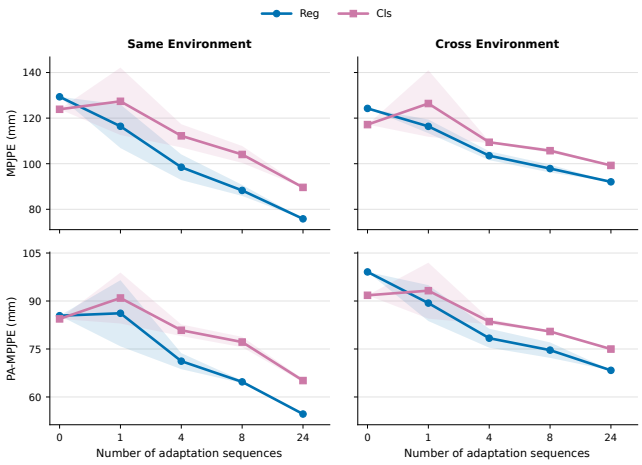


Figure S3: **Data-efficient adaptation of Reg and Cls.** Left panels report same-environment Hall adaptation; right panels report Hall-to-Lab transfer. Shading denotes mean ± standard deviation over three optimization seeds.

Table S2: **Stability of limited-data adaptation.** mmSimPrior-Reg under three adaptation runs with the simulation checkpoint, 4K horizon, and test manifest held fixed. Lab and Studio vary only the optimization seed; Hall additionally draws a different 24-sequence support subset per run. Values are mean ± sample standard deviation in millimeters.

Metric	Hall*	Lab	Studio
MPJPE↓	75.8 ± 0.1	77.2 ± 0.0	62.1 ± 0.1
PA-MPJPE↓	54.7 ± 0.0	57.0 ± 0.0	49.5 ± 0.0
W-MPJPE↓	115.7 ± 0.4	116.1 ± 0.1	94.6 ± 0.3
MPVPE↓	101.8 ± 0.2	103.5 ± 0.0	85.3 ± 0.0

*Support subset also varies across runs.

24.7–39.0% margins to the strongest adapted baseline. In Hall the support subset varies across runs as well, so the same conclusion holds for support selection and not only for optimization noise. mmSimPrior-Cls is comparably stable, with a maximum standard deviation of 0.4 mm across Lab and Studio. The main tables report the seed-2026 run, and all baselines use the same fixed seed.

D.2 Generalization to Held-Out Factors

We isolate three sources of real-world shift: subject identity, motion category, and environment. The tested factor is absent from adaptation, or the entire target benchmark is unseen in the zero-shot setting. Avg. is computed from unrounded sequence-level errors and is weighted by the number of valid recordings.

Held-out subjects. Table S3 reports nine-fold leave-one-subject-out adaptation. Four Lab folds and five Studio folds reserve one participant for testing while adapting on 24 sequences from the remaining participants. Hall is excluded because it contains only one participant.

Table S3: **Nine-fold leave-one-subject-out adaptation.** Results are sequence-weighted over 1,013 held-out-subject recordings.

Method	MPJPE↓	PA-MPJPE↓	W-MPJPE↓	MPVPE↓
mmDiff†	105.6	81.6	172.4	150.9
RAPTR†	117.3	92.7	176.8	168.6
mmGPE†	106.5	77.2	142.3	148.4
mmSimPrior-Reg	**72.6**	**54.6**	**108.0**	**98.0**
mmSimPrior-Cls	92.4	69.4	124.4	126.4

Table S4: **MPJPE across sensing distances after 24-sequence adaptation.** Near, Mid, and Far contain 303, 379, and 372 recordings, respectively. Avg. is sequence-weighted.

Method	Near	Mid	Far	Avg.
mmDiff† (Fan et al. 2024)	95.1	104.9	109.8	103.8
RAPTR† (Kato et al. 2025)	107.5	115.3	117.6	113.9
mmGPE† (Xue et al. 2023)	97.5	102.0	109.7	103.4
mmSimPrior-Reg	**65.7**	**68.3**	**73.4**	**69.4**
mmSimPrior-Cls	84.3	87.1	94.7	89.0

Across the nine folds, Reg and Cls obtain unweighted fold-level MPJPEs of 72.7 ± 8.6 and 92.4 ± 4.9 mm. The sequence-weighted results confirm that both modes transfer to unseen body geometry and motion style, with Reg providing the strongest accuracy.

Held-out motions and environments. Table S11 removes motion categories from the adaptation set and separately evaluates each environment without any target- benchmark adaptation or model selection.

Reg is strongest in every environment when limited adaptation is allowed. In zero-shot transfer, Cls is strongest in Hall and Lab and gives the best weighted average, supporting the discrete prior when target supervision is unavailable.

D.3 Generalization to Distance and Motion Regime

We further examine whether the adapted models remain robust to variation in sensing distance and motion regime. Because zero-shot behavior is already analyzed in Fig. S3 and Table S11, we focus here on MPJPE after 24-sequence real-world adaptation.

Tables S4 and S5 provide complementary diagnostics of the adapted models. Across sensing distances, all methods exhibit some degradation as the subject moves farther from the radar. Nevertheless, mmSimPrior-Reg remains strongest in every range, with MPJPE increasing moderately from 65.7 mm in the Near group to 73.4 mm in the Far group. mmSimPrior-Cls consistently ranks second, indicating that both prediction modes retain their advantage under reduced signal strength and spatial resolution.

Motion regime produces a larger performance difference. For mmSimPrior-Reg, MPJPE increases from 66.5 mm

Table S5: **MPJPE across motion regimes after 24-sequence adaptation.** The evaluation set contains 872 in-place and 182 locomotion recordings. Avg. is sequence-weighted.

Method	In-place	Locomotion	Avg.
mmDiff† (Fan et al. 2024)	95.0	146.1	103.8
RAPTR† (Kato et al. 2025)	105.6	153.2	113.9
mmGPE† (Xue et al. 2023)	96.8	135.3	103.4
mmSimPrior-Reg	**66.5**	**83.1**	**69.4**
mmSimPrior-Cls	83.9	113.4	89.0

Table S6: **Mesh-level reconstruction.** Sequence-weighted MPVPE over the common 1,054-sequence population.

Method	Zero-shot	24-sequence
mmDiff† (Fan et al. 2024)	189.7	148.5
RAPTR† (Kato et al. 2025)	244.1	163.6
mmGPE† (Xue et al. 2023)	207.2	145.9
mmSimPrior-Reg	173.4	**94.0**
mmSimPrior-Cls	**172.4**	121.2

for in-place motions to 83.1 mm for locomotion, while mmSimPrior-Cls increases from 83.9 mm to 113.4 mm. Despite this more challenging setting, both variants maintain clear margins over all adapted baselines. These results suggest that the learned simulation priors remain robust to sensing-range variation, while locomotion remains the more difficult real-world operating regime.

D.4 Mesh Accuracy and Temporal Refinement

We evaluate mesh-level reconstruction and isolate the effect of projecting direct Reg predictions through the frozen motion prior.

Cls gives the best zero-shot MPVPE, whereas Reg is substantially stronger after adaptation. MPR changes frame-level spatial errors by at most 1.3%, but reduces velocity, acceleration, and jerk errors by 7.7%, 18.0%, and 36.7%, respectively. It therefore acts primarily as a temporal regularizer.

Table S7: **Effect of motion-prior refinement.** Relative change is measured from Direct Reg; negative is better.

Metric	Direct	+MPR	Rel. change
MPJPE↓	68.6	69.4	+1.1%
PA-MPJPE↓	52.7	53.0	+0.5%
W-MPJPE↓	104.7	105.2	+0.4%
MPVPE↓	92.8	94.0	+1.3%
Joint Vel. Err.↓	41.7	38.5	**-7.7%**
Joint Accel. Err.↓	49.9	40.9	**-18.0%**
Pred. Jerk↓	63.1	40.0	**-36.7%**

D.5 Motion-Tokenizer Capacity

Finally, we test whether the frozen tokenizer imposes a reconstruction ceiling by varying its codebook and number of joint-temporal tokens.

Table S8: **Motion-tokenizer capacity.** Each entry is MPJPE / PA-MPJPE / MPVPE. The codebook block fixes 320 tokens; the token-count block fixes a 2048×256 codebook.

	Setting	AMASS	Lab
Codebook	1024×256	10.45 / 6.28 / 13.90	14.24 / 9.55 / 19.43
	2048×128	10.11 / 5.42 / 13.30	13.60 / 8.75 / 18.18
	2048×256	**9.15 / 5.39 / 12.26**	**10.38 / 7.95 / 14.54**
Tokens	80	20.47 / 12.79 / 28.13	30.83 / 21.44 / 43.65
	160	14.95 / 9.06 / 20.19	21.99 / 14.84 / 30.07
	320	**9.15 / 5.39 / 12.26**	**10.38 / 7.95 / 14.54**

At the selected 2048×256 codebook with 320 tokens, tokenizer reconstruction error is far below the radar-conditioned errors above, indicating that tokenizer capacity is not the dominant bottleneck.

D.6 Domain Randomization on a Strong Baseline

To test whether domain randomization alone explains the gains of mmSimPrior, we apply the same full DR pipeline to mmGPE, the closest simulation-based baseline. The clean and DR variants use the same architecture, simulated corpus, training horizon, seed, and real-world adaptation protocol; DR is the only intended difference. We additionally report the final mmSimPrior variants for reference.

Applying full DR to mmGPE substantially improves zero-shot transfer, reducing Hall MPJPE / PA-MPJPE from 156.0 / 100.8 mm to 141.3 / 87.0 mm. However, it does not improve the 24-sequence adapted results in either the same- or cross-environment setting. This indicates that DR improves robustness of the synthetic initialization, but its benefit after real-world adaptation depends on the representation and adaptation mechanism. The consistently lower errors of mmSimPrior therefore arise from the joint design of its factorized signal, motion, and mapping priors with DR, rather than from applying stronger augmentation alone.

D.7 Effect of the Unified Input Representation

The main comparison adapts every baseline to the unified four-heatmap observation. To verify that this adaptation does not disadvantage the baselines, we additionally pretrain and adapt mmGPE using its native sensing inputs—a top-view MVDR heatmap and the range–Doppler map—under an otherwise identical protocol. Table S10 shows that the unified representation strengthens rather than weakens the baseline: zero-shot MPJPE improves by 42.3% and PA-MPJPE by 31.7% over the native inputs, and after 24-sequence adaptation the unified variant remains at least comparable across all four metrics. The four-view representation strengthens rather than disadvantages mmGPE, especially in zero-shot transfer. Thus, the gains of mmSimPrior cannot be attributed to an unfavorable conversion of the baseline input.

Table S9: **Domain randomization on mmGPE.** The two mmGPE rows form a strictly paired clean/full-DR comparison; the mmSimPrior rows report the final full models for reference. Adapted results use 24 Hall sequences, 4K updates, terminal adaptation checkpoints, and no test-set model selection. Each entry is MPJPE / PA-MPJPE.

Method	DR	Hall ZS	Hall FT	H→L FT
mmGPE†	×	156.0 / 100.8	106.4 / 75.5	111.8 / 81.0
mmGPE†	✓	141.3 / 87.0	108.1 / 77.4	115.1 / 82.9
mmSimPrior-Reg	✓	129.4 / 85.4	**75.9 / 54.7**	**96.7 / 73.3**
mmSimPrior-Cls	✓	**123.9 / 84.4**	89.5 / 65.5	100.0 / 75.2

Table S10: **Native-input versus unified-input control for mmGPE on Hall.** The native variant uses mmGPE's original sensing inputs (top-view MVDR and range–Doppler); the unified variant uses the four heatmaps provided to all methods in our comparison. The model, simulated corpus, 80K pretraining schedule, seed, terminal-checkpoint policy, and 24-sequence 4K adaptation protocol are otherwise identical. Metrics are in millimeters on the fixed 89-recording Hall test set.

Setting	Metric	Native (2 maps)	Unified (4 maps)
Zero-shot	MPJPE↓	270.3	**156.0**
	PA-MPJPE↓	147.5	**100.8**
	W-MPJPE↓	552.2	**469.3**
	MPVPE↓	380.7	**226.0**
24-seq finetune	MPJPE↓	108.0	**106.4**
	PA-MPJPE↓	75.7	**75.5**
	W-MPJPE↓	152.3	**147.0**
	MPVPE↓	152.1	**150.8**

D.8 Evaluation Against Official RT-Pose Keypoints

To test whether the RT-Pose conclusions depend on our pseudo-SMPL-X references, we additionally evaluate all locked models directly against 12 common limb landmarks in the official annotations: the bilateral hips, knees, ankles, shoulders, elbows, and wrists. Pelvis, thorax, and head are excluded because they do not have unambiguous SMPL-X correspondences.

Table S12 preserves the ranking under both metrics. mmSimPrior-Reg improves over the strongest baseline by 8.3 mm MPJPE (paired-bootstrap 95% CI $[3.1, 13.4]$; sequence win rate 59.5%) and 3.0 mm PA-MPJPE, close to the 9.1 and 3.5 mm margins obtained with the pseudo-SMPL-X references. Thus, the comparative gain is robust to the reference-construction pipeline, although absolute errors remain affected by landmark-convention differences. mmSimPrior-Cls remains second, but we do not claim statistical significance for its margin under this protocol.

E Qualitative Results

We provide additional qualitative comparisons under zero-shot transfer and three held-out-factor protocols. The RGB views and fitted pseudo-SMPL-X overlays are included only to visualize the reference annotations; all evaluated methods receive radar heatmaps alone.

Zero-shot transfer. Figure S4 examines the most challenging setting, in which the entire target benchmark is unseen during model adaptation and selection. mmGPE often recovers a coarse human configuration but collapses ambiguous limb evidence toward generic arm poses, missing action-specific configurations such as raised, crossed, or head-touching arms. Both mmSimPrior variants reduce these severe articulation errors. In particular, mmSimPrior-Cls more consistently preserves plausible action-specific joint configurations by decoding through the frozen motion codebook, while mmSimPrior-Reg also provides substantially closer body orientation and local articulation than the baseline. These observations are consistent with the stronger zero-shot constraint provided by the learned joint-temporal motion prior.

Generalization to held-out factors. Figure S5 separates the effects of motion, subject, and environment shifts. For held-out motions, mmGPE captures the broad action trajectory but often loses action-specific arm configurations and lower-body motion phase. mmSimPrior-Reg more closely follows fine-grained articulation, while mmSimPrior-Cls maintains a plausible temporal progression under the discrete motion constraint. For held-out subjects, changes in body geometry and motion style lead to larger orientation and limb-configuration errors for the baseline; both mmSimPrior variants preserve the body configuration and walking progression more consistently. Under the unseen-environment shift, mmGPE exhibits pronounced torso- and arm-orientation errors, whereas both variants remain closer to the reference global orientation and limb arrangement. Across the adapted settings, Reg generally provides the closest local pose match, while Cls trades some continuous precision for stronger structural plausibility. Together, these examples show that the learned simulation priors reduce not only average joint error but also severe geometric and temporal failures under multiple sources of real-world shift.

(a) Motion categories unseen during real-world adaptation

Method	Hall	Lab	Studio	Avg.
mmDiff† (Fan et al. 2024)	109.6 / 73.4	107.5 / 73.9	109.4 / 73.7	108.7 / 73.7
RAPTR† (Kato et al. 2025)	121.2 / 85.3	112.2 / 81.5	123.4 / 89.8	118.8 / 86.0
mmGPE† (Xue et al. 2023)	112.1 / 73.4	110.2 / 72.9	106.9 / 70.2	108.9 / 71.7
mmSimPrior-Reg	**85.8 / 58.4**	**87.4 / 59.8**	**68.4 / 51.8**	**78.0 / 55.8**
mmSimPrior-Cls	100.0 / 68.6	100.2 / 69.0	93.0 / 64.5	96.7 / 66.7

(b) Per-environment zero-shot transfer

Method	Hall	Lab	Studio	Avg.
mmDiff† (Fan et al. 2024)	139.2 / 107.0	133.9 / 104.1	125.4 / 90.2	130.0 / 97.2
RAPTR† (Kato et al. 2025)	187.7 / 103.2	175.5 / 114.7	163.2 / 95.8	170.2 / 104.0
mmGPE† (Xue et al. 2023)	156.0 / 100.8	142.8 / 99.7	137.7 / 94.2	141.3 / 97.0
mmSimPrior-Reg	129.4 / 85.4	124.3 / 99.1	**113.3** / 89.7	119.1 / 93.1
mmSimPrior-Cls	**123.9 / 84.4**	**117.2 / 91.8**	119.4 / **88.9**	**118.9 / 89.7**

Table S11: **Generalization across held-out motions and environments.** Panel (a) evaluates motion categories absent from the 24-sequence adaptation set on 59 Hall, 175 Lab, and 220 Studio recordings. Panel (b) reports zero-shot performance over 89 Hall, 424 Lab, and 541 Studio recordings. Each entry is MPJPE / PA-MPJPE; Avg. is sequence-weighted.

Table S12: **Evaluation against the official RT-Pose keypoints.** Errors over the 12 exactly corresponding landmarks on all 111 evaluated sequences.

Method	official MPJPE↓	PA-MPJPE↓
mmDiff† (Fan et al. 2024)	187.7	132.6
RAPTR† (Kato et al. 2025)	194.5	139.8
mmGPE† (Xue et al. 2023)	187.9	134.0
mmSimPrior-Reg	**179.4**	**129.6**
mmSimPrior-Cls	185.4	131.6

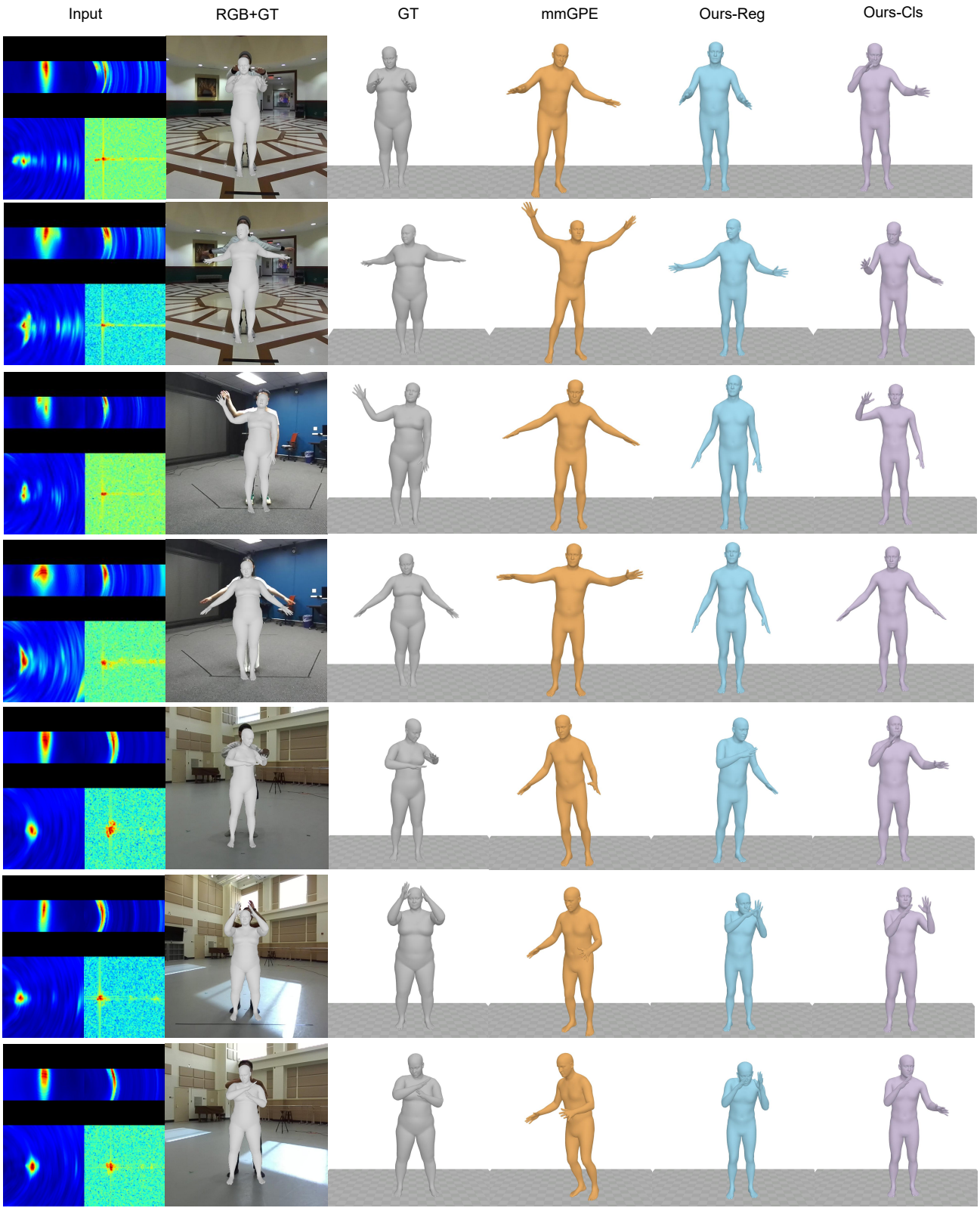


Figure S4: **Zero-shot qualitative results on mmSimPrior-Real.** Rows show the radar input, the stereo reference with its fitted pseudo-SMPL-X overlay, the canonical ground truth, and predictions from mmGPE, mmSimPrior-Reg, and mmSimPrior-Cls. The RGB view is shown only for reference; no mmSimPrior-Real sequence is used for adaptation or model selection.

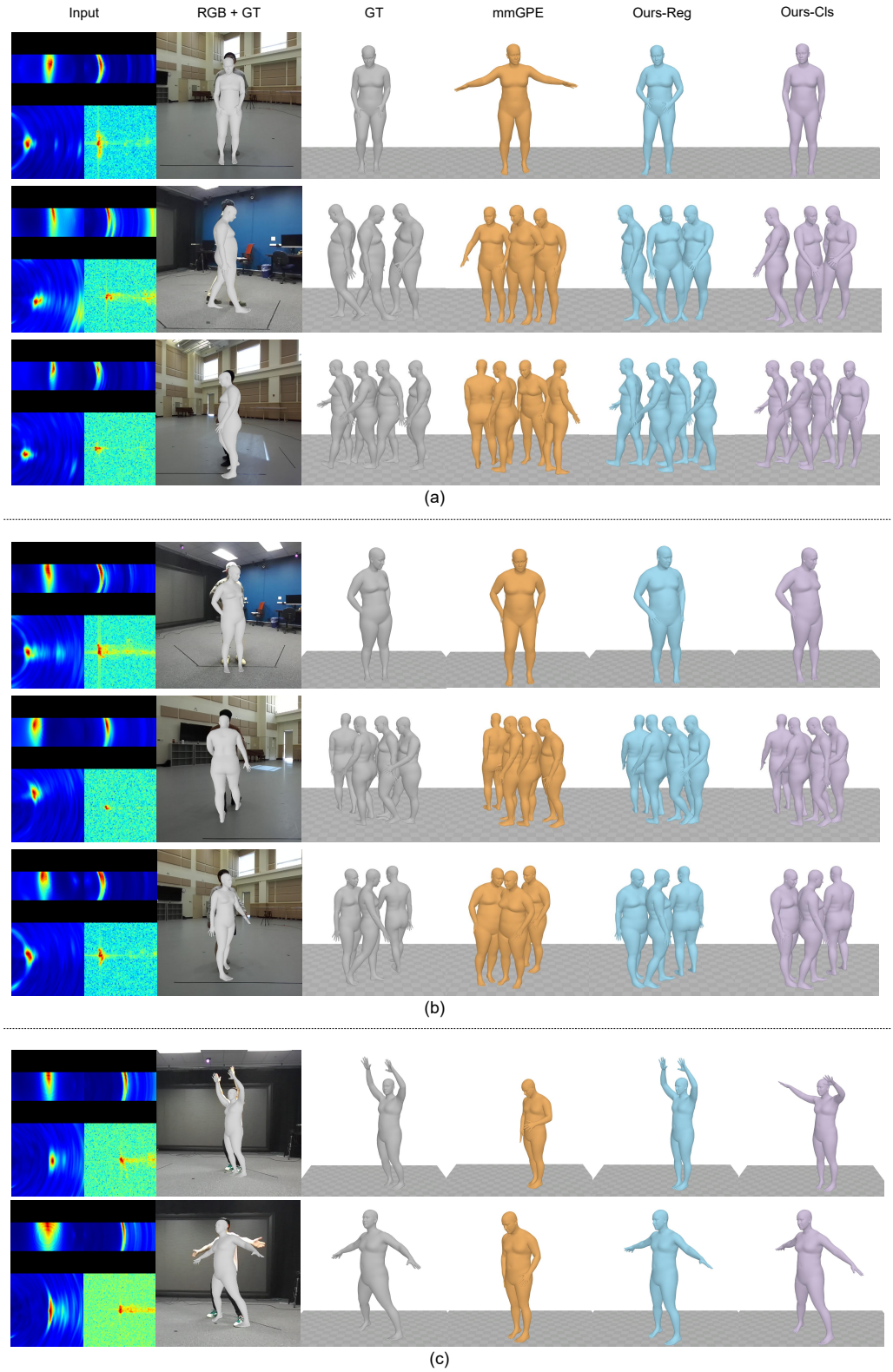


Figure S5: **Qualitative generalization to held-out factors.** We evaluate (a) motion categories excluded from the 24-sequence adaptation set, (b) leave-one-subject-out transfer, and (c) transfer to an environment not used for adaptation. We compare mmGPE, mmSimPrior-Reg, and mmSimPrior-Cls against the pseudo-SMPL-X ground truth. In sequence examples, reconstructed frames progress from left to right.