STANCE: SCENE-CONSISTENT REAL-TIME MOTION ESTIMATION FROM SPARSE WEARABLES VIA POSTERIOR-CORRECTED RESIDUAL DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Wearable tracking has evolved from driving game avatars to supporting embodied intelligence, where demonstrations from sparse body-worn sensors train agents in simulation. Configurations range from three 6-DoF signals on a headset and two controllers to six with hip and foot trackers. Existing methods for arbitrary sensor sets assign each tracker a fixed influence, masking absent channels with hand-crafted weights that encode presence rather than reliability. Motion captured in a physical room and replayed in a virtual scene also penetrates the geometry it should contact. We present STANCE, a predict-then-correct framework. A configuration-agnostic predictor produces a temporally coherent prior from fixed core trackers, making continuity independent of the active tracker set. A single-frame residual denoiser corrects only tracker-inconsistent motion, avoiding an accuracy-latency tradeoff. Per-joint trust, measured by the gap a partial inverse-kinematics solve leaves against its tracker targets, gates how much the denoiser may rewrite, letting one model serve every configuration. A training-free posterior correction enforces scene constraints at every diffusion step, removing penetration without retraining or a learned collision model. STANCE runs in real time, achieves state-of-the-art accuracy across most tracker counts, and maintains continuity under dynamic reconfiguration.

1 INTRODUCTION

Motion captured from body-worn sensors is no longer just a control signal. Sparse streams once used to drive game avatars now serve as demonstrations for embodied agents in physically interactive scenes (Ji et al., 2024; He et al., 2024). This raises the evaluation standard: avatars need to look right; demonstrations must remain valid in simulation. That validity is challenging even before considering the scene, because capture rigs vary. Users may have only a headset and two controllers—three 6-DoF signals with no lower-body observations—or a full-body suit covering every limb. Each configuration defines a different input space, so fixed-rig methods, whether based on regression (Jiang et al., 2022a; Zheng et al., 2023; Aliakbarian et al., 2023), generative modelling (Aliakbarian et al., 2022; Du et al., 2023; Castillo et al., 2023; Feng et al., 2024), or physical simulation (Winkler et al., 2022; Lee et al., 2023; Ye et al., 2022), require a new model for each rig.

Two approaches avoid per-rig retraining. DragPoser (Ponton et al., 2025) keeps sensors implicit, optimizing a latent pose until the decoded body matches the trackers; configurations differ in the objective, but harder motions require more iterations. HMD-Poser (Dai et al., 2024) and DiffusionPoser (Van Wouwe et al., 2024) instead represent sensors explicitly, using zero-padded missing channels or a binary inpainting mask. This keeps inference feed-forward with fixed per-frame cost, suiting real-time capture, and is the route we follow. A binary marker, however, indicates only measurement presence, not reliability. Two failures follow. First, worn sensors may drift or slip, yet a present channel still enters at full weight, propagating its error along the anchored chain. Second, a returning tracker is restored between consecutive frames, abruptly replacing a prior-driven joint with an unanticipated measurement; near the root, the jump spreads through the body. RPM (Barquero et al., 2025) smooths transitions between tracking and synthesis, but limits pose updates with a hand-crafted, joint-agnostic temporal function. Because this models future uncertainty rather than current measurement reliability, corrections cannot focus on joints governed by the affected tracker.

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

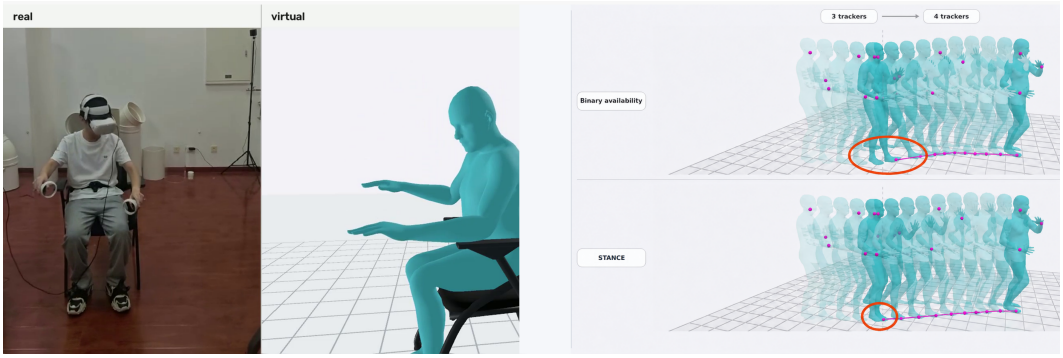

Figure 1: **Scene consistency and continuous reconfiguration.** We introduce STANCE, a single model that keeps sparse-tracker motion usable in simulation. **Left:** the hand rests on the armrest instead of sinking through it. **Right:** a fourth tracker joins three mid-sequence, where a binary availability flag makes the pose jump and STANCE does not.

To replace binary availability with measurement reliability, STANCE compares two pose estimates. A feed-forward prior provides a temporally coherent full-body pose, while trackers observe the current frame sparsely across the kinematic chain. A joint-limited partial inverse-kinematics solve toward the tracker targets yields continuous per-joint confidence from the remaining residuals: a reading that drifts enough to conflict with the body leaves a gap the solve cannot close. When a tracker reconnects, its corresponding IK target is interpolated from the prior prediction to the returning measurement to avoid discontinuities. STANCE thus follows a predict-then-correct design: a configuration-agnostic predictor maintains temporal continuity, while a single-frame residual denoiser corrects the prior with confidence-controlled per-joint amplitudes. Configuration changes affect only feature values, allowing one model to support all rigs at fixed per-frame cost.

Even a perfectly captured pose may be invalid in a new scene. It was valid only in its original room, while the prior for unobserved joints was trained on furniture-free captures; neither knows the virtual environment. Replayed against new geometry, limbs may penetrate intended contact surfaces—a small joint-angle error but a critical failure for demonstrations, since a hand passing through a grasped object provides no useful supervision. Because the denoiser predicts a residual rather than a pose, scene correction can act directly on its output. At each diffusion step, penetration depth from a signed distance field adjusts the residual, following diffusion posterior sampling (Chung et al., 2023; Tian et al., 2025), without a learned collision model or retraining.

To summarize, our contributions are: (1) A configuration-agnostic residual denoiser guided by per-joint confidence, supporting rigs from three trackers to a full-body suit at fixed per-frame cost. (2) A training-free posterior correction on that residual, removing scene penetration at every diffusion step without a learned collision model. (3) A real-time Unity prototype driving a full body from live trackers and staying continuous as trackers drop out and return.

2 Related Work

2.1 Configuration-Agnostic Input

Three 6-DoF head-and-hand poses constrain only a small fraction of body degrees of freedom, leaving the rest to tracking history via temporal attention (Jiang et al., 2022a; Zheng et al., 2023; Hu et al., 2024) or state-space models (Dong et al., 2024; Zhao et al., 2026); learned priors such as VAEs (Dittadi et al., 2021), normalising flows (Aliakbarian et al., 2022), and diffusion models (Du et al., 2023; Castillo et al., 2023); physics-based sparse tracking (Winkler et al., 2022; Lee et al., 2023); or additional signals from IMU suits (Yi et al., 2021; Jiang et al., 2022b; Ren et al., 2023). Yet existing methods fix the sensor configuration before training, requiring a separate model for each setup.

Serving arbitrary sensor sets from one model is a more recent goal. DragPoser(Ponton et al., 2025) leaves the sensors implicit, optimising inside a tracker-agnostic VAE latent space with available sensors entering only as loss terms. Layout flexibility comes for free, but configuration dependence

returns as manual tuning: the weight on the temporal prior is retuned per sensor count, and the horizon its predictor rolls out is lengthened from one step to sixty as the sensor set grows more ambiguous. The other route models sensors explicitly and marks missing ones. HMD-Poser (Dai et al., 2024) and IMUPoser (Mollyn et al., 2023) zero-pad absent sensors in a fixed-size input, while DiffusionPoser (Van Wouwe et al., 2024) uses measurements as sampling guidance and clamps observed entries during denoising, allowing variable layouts. Inference remains feed-forward with motion-independent cost, but the clamp enforces every available reading regardless of reliability.

Fixed loss weights or binary masks determine tracker influence before measurements arrive and ignore uneven kinematic reach—a pelvis tracker affects the whole body, while a foot tracker mostly affects its leg. STANCE retains a single model but infers this influence from partial-IK residuals to tracker targets, weighting each reading by how well it can be satisfied rather than by mere presence.

2.2 Unreliable and Time-Varying Input

Tracking failures are typically handled by bridging temporary interruptions (Aliakbarian et al., 2023; Jiang et al., 2025; Qian et al., 2024) or modelling pose ambiguity (Xia et al., 2025), both under fixed tracker configurations. RPM (Barquero et al., 2025) smooths tracking-to-synthesis transitions with a hand-crafted, joint-agnostic temporal bound; because it models future uncertainty rather than current measurement reliability, corrections cannot focus on joints governed by the affected tracker. These methods also ignore recovery magnitude. After a long dropout, the prior may diverge from the returning measurement, making larger gaps cause sharper corrections. STANCE instead acts on the target: after recovery, the IK target is interpolated from the prior prediction to the measurement, so the solver follows a continuous rather than abrupt target.

2.3 Scene Constraints and Training-Free Guidance

Scene consistency is typically enforced during training. Physics-based trackers embed the environment in simulation and let contact emerge from dynamics (Lee et al., 2023), while EnvPoser (Xia et al., 2025) conditions sparse-pose estimation on scene context and models the remaining uncertainty. Both require the scene during training and do not transfer to unseen scenes. Diffusion provides another route: inpainting fixes observations and lets the denoiser complete the rest (Lugmayr et al., 2022), as in DiffusionPoser (Van Wouwe et al., 2024). However, completion is still limited by a scene-agnostic prior, so furniture-free training cannot prevent wall penetration, while scene conditioning would require scene-dependent training. Posterior sampling avoids both by refining each clean estimate online along a measurement-likelihood gradient (Chung et al., 2023; Tian et al., 2025). STANCE applies this in residual space using only a penetration-depth gradient, allowing scene changes without a learned collision model or denoiser retraining.

3 Method

We estimate the full-body pose at frame n from tracker measurements up to n and previous pose predictions. The framework is causal, with no access to future frames or offline sequence processing; all quantities depend only on frame n and its history. Figure 2 shows one frame's processing.

3.1 Preliminaries

State and observations. We use the SMPL(Loper et al., 2015) skeleton with $J = 24$ joints. At frame n, the pose consists of the global orientations $\mathbf{R}_n^j \in SO(3)$ of all joints,

$$\mathbf{x}_n = \left[\, \phi(\mathbf{R}_n^1), \ldots, \phi(\mathbf{R}_n^J) \,\right] \in \mathbb{R}^{J\times 6}, \tag{1}$$

where ϕ is the continuous six-dimensional rotation representation(Zhou et al., 2019).

We use global rather than parent-relative rotations because trackers report global orientations, making measurements directly comparable to estimates, while global parameterisation prevents local errors from propagating down the kinematic chain and enables independent per-joint correction. Joint positions follow from forward kinematics on the subject's rest skeleton.

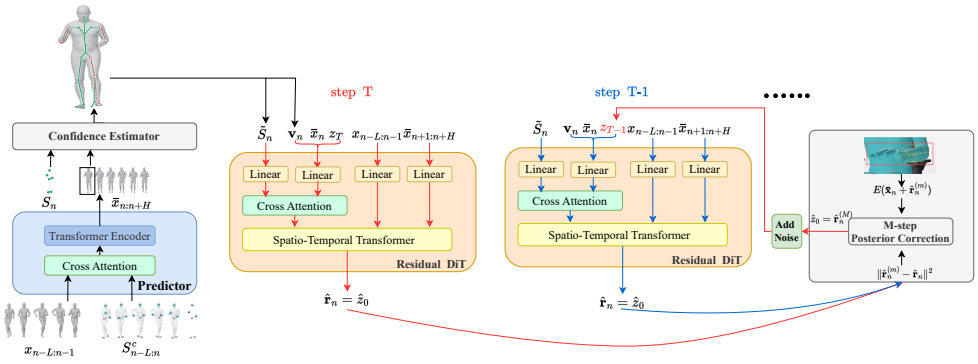

Figure 2: STANCE at frame n. **(a)** The *predictor* P_φ takes pose history $\mathbf{x}_{n-L:n-1}$ and a window of three core trackers $\mathbf{S}^{\mathrm{c}}_{n-L:n}$, producing the prior $\bar{\mathbf{x}}_n$ and future predictions $\bar{\mathbf{x}}_{n+1:n+H}$. **(b)** A partial inverse-kinematics solve compares IK targets with the solved pose to estimate observation confidence. **(c)** The *Residual DiT* f_θ attends over trackers, joints, and trajectories to produce a prior residual $\hat{\mathbf{r}}_n$. **(d)** At each denoising step, the implied pose $\bar{\mathbf{x}}_n + \hat{\mathbf{r}}_n$ is queried against the scene SDF, and the residual follows the penetration gradient while remaining near the denoiser prediction. After T steps, $\hat{\mathbf{x}}_n = \bar{\mathbf{x}}_n + \hat{\mathbf{r}}_n$ enters the history for frame $n{+}1$.

Measurements come from up to $K = 6$ rigidly attached 6-DoF trackers—head, hands, pelvis, and feet—associated with joints $\pi(1), \ldots, \pi(K)$. Tracker k provides position $\mathbf{p}^k_n$, orientation $\mathbf{R}^k_n$, and availability $a^k_n \in \{0, 1\}$:

$$\mathbf{s}^k_n = \left[\, \mathbf{p}^k_n, \phi(\mathbf{R}^k_n), a^k_n \,\right] \in \mathbb{R}^{10}. \tag{2}$$

The headset and controllers form the always-active core set $\mathcal{K}_{\mathrm{c}}$; the remaining trackers are optional, $\mathcal{K}_{\mathrm{o}}$. A *configuration* is the availability vector $\mathbf{a}_n \in \{0, 1\}^K$, whose core entries are fixed to one, so it spans the eight subsets of $\mathcal{K}_{\mathrm{o}}$ and thus all three- to six-tracker consumer setups. Unlike prior work, which fixes $\mathbf{a}$ per session, we allow $\mathbf{a}_n$ to vary over time.

Reference frame. We express all poses and tracker measurements in a local frame F_n derived from the current head measurement. Writing $\mathbf{f}_n = \mathbf{R}^{\mathrm{head}}_n \mathbf{e}_3$ for the head forward axis, obtained by rotating the canonical forward direction $\mathbf{e}_3$ by the head tracker's measured orientation, the frame is anchored on the floor beneath the head and aligned by head yaw,

$$\mathbf{o}_n = \left(p^{\mathrm{head}}_{n,x},\ y_{\mathrm{floor}},\ p^{\mathrm{head}}_{n,z}\right), \qquad \psi_n = \operatorname{atan2}\left(f_{n,x},\ f_{n,z}\right). \tag{3}$$

A world-space point $\mathbf{q}$ is transformed as $\mathbf{R}_y(\psi_n)^\top(\mathbf{q} - \mathbf{o}_n)$, and a global joint orientation as $\mathbf{R}^j_n = \mathbf{R}_y(\psi_n)^\top \mathbf{R}^j_{\mathrm{world}}$, with $\mathbf{R}_y$ a rotation about the vertical axis. Since F_n is computed from the current observation rather than integrated over time, it has no temporal state and cannot accumulate drift.

Scene. The replay environment is represented by convex primitives defining a signed distance field $s : \mathbb{R}^3 \to \mathbb{R}$, positive outside and negative inside, with ∇s as the outward normal. As only s and ∇s are queried, the scene is training-independent and can vary between sessions.

Diffusion. A diffusion model(Ho et al., 2020) corrupts a clean variable $\mathbf{z}_0$ over T steps under a fixed variance schedule $\{\bar{\alpha}_t\}$, $q(\mathbf{z}_t \mid \mathbf{z}_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\,\mathbf{z}_0,\ (1{-}\bar{\alpha}_t)\mathbf{I}\right)$, until $\mathbf{z}_T$ is indistinguishable from Gaussian noise. Learning reverses this corruption: a transformer G_θ takes the corrupted variable $\mathbf{z}_t$, a timestep embedding t and a conditioning signal $\mathbf{c}$, and returns an estimate of the clean data,

$$\hat{\mathbf{z}}_0 = G_\theta(\mathbf{z}_t,\ t,\ \mathbf{c}). \tag{4}$$

Predicting $\mathbf{z}_0$ directly rather than the added noise is standard in motion generation(Tevet et al., 2023; Du et al., 2023), since it permits geometric losses on joint positions. At inference G_θ is applied from $t = T$ down to 1, each estimate being re-noised to the preceding step; STANCE uses DDIM sampling(Song et al., 2021) to keep this loop short enough for real time. Crucially, STANCE does not diffuse the absolute pose: the diffused variable is a single-frame residual on a previously predicted prior, as the next section makes precise.

3.2 CONFIGURATION-INDEPENDENT MOTION PRIOR

Assigning confidence to a tracker requires a pose estimate that does not already depend on it. We obtain one from a predictor P_φ that reads the recent pose history together with a window of core-tracker features and returns the current frame along with a short horizon,

$$\big[\,\bar{\mathbf{x}}_n,\ \ldots,\ \bar{\mathbf{x}}_{n+H}\,\big] = P_\varphi\big(\mathbf{x}_{n-L:n-1},\ \mathbf{S}^{\mathrm{c}}_{n-L:n}\big), \qquad L = H = 10, \tag{5}$$

where $\bar{\mathbf{x}}_n$ is the *prior* that the rest of the method corrects.

Expressed in the local frame F_n, a core tracker $k \in \mathcal{K}_{\mathrm{c}}$ contributes at each frame m of the window $n-L, \ldots, n$ its orientation, its frame-to-frame rotation, its position and its displacement,

$$\mathbf{u}^k_m = \big[\ \phi(\mathbf{R}^k_m),\ \phi\big(\mathbf{R}^{k\top}_{m-1}\mathbf{R}^k_m\big),\ \mathbf{p}^k_m,\ \mathbf{p}^k_m - \mathbf{p}^k_{m-1}\ \big] \in \mathbb{R}^{18}, \tag{6}$$

and the three concatenate into $\mathbf{S}^{\mathrm{c}}_m \in \mathbb{R}^{54}$. A linear map lifts $\mathbf{S}^{\mathrm{c}}_{n-L:n} \in \mathbb{R}^{(L+1)\times 54}$ to tracker tokens $\mathbf{H}^{\mathrm{s}} \in \mathbb{R}^{(L+1)\times D}$. The motion side is lifted the same way, but only its history exists yet: the L past poses occupy the leading token positions and the $H+1$ frames still to be predicted are held by zero placeholders, giving $\mathbf{H}^{\mathrm{m}} \in \mathbb{R}^{(L+H+1)\times D}$. Following the rolling prediction architecture of RPM(Barquero et al., 2025), the motion tokens read the tracker tokens once by cross-attention before a transformer encoder processes them jointly,

$$\mathbf{H} = \mathrm{Enc}\Big(\mathrm{LN}\big(\mathbf{H}^{\mathrm{m}} + \mathrm{CA}(\mathbf{H}^{\mathrm{m}},\ \mathbf{H}^{\mathrm{s}})\big)\Big), \tag{7}$$

after which $\bar{\mathbf{x}}_n, \ldots, \bar{\mathbf{x}}_{n+H}$ are decoded from the trailing $H+1$ tokens of $\mathbf{H}$.

We restrict P_φ to the three core trackers; the optional ones are withheld deliberately. The exclusion leaves the architecture of P_φ identical across configurations: the availability vector $\mathbf{a}_n$ never appears in Eq. 5, entering only in the two modules that follow, where it drives the single-frame correction of the prior under any configuration drawn from $\mathcal{K}_{\mathrm{o}}$.

Following predictive world models(Hafner et al., 2025; Assran et al., 2025), we use P_φ to predict H frames. Forecasting helps the latent state to capture motion dynamics beyond the current frame.

During training, P_φ is fed its own previous predictions rather than ground truth(Barquero et al., 2025), exposing it to the error distribution encountered when its outputs become inputs. It is trained solely with the reconstruction loss $\|\bar{\mathbf{x}}_n - \mathbf{x}_n\|^2$ and fully trained before the denoiser.

3.3 OBSERVATION CONFIDENCE IN SPACE AND TIME

The trackers constrain six body points while the prior covers all J joints, so the readings must first be propagated across the skeleton. A partial inverse-kinematics solve performs this propagation, and the per-joint distances between the solved pose and its IK targets provide the confidence signal.

Partial inverse kinematics. The solve is *partial* because only part of the pose is ever inverted: a joint a tracker sits on takes the measured rotation by assignment, and FABRIK closes five kinematic chains—torso, two arms, two legs—onto their endpoint trackers, each solved only when the measurements it needs are present, which for a limb means its own endpoint tracker and for the torso a measured root as well. Initialised at the prior, the solve returns a pose together with a label per joint,

$$\big(\mathbf{x}^{\mathrm{ik}}_n,\ \mathbf{c}_n\big) \ \leftarrow\ \mathrm{IK}\Big(\bar{\mathbf{x}}_n,\ \big\{\mathbf{p}^k_n\big\}_{k\,:\,a^k_n=1}\Big), \tag{8}$$

where c^j_n is DIRECT for tracker-provided rotations, CHAIN for IK-chain joints adjusted by FABRIK toward a tracker, and INHERITED for unobserved joints retaining the prior.

Tracker drift and noise can make the target it sets inconsistent with the body or unreachable. The solve-to-target distance therefore indicates how much the reading should be trusted, $\rho^k_n = \big\|\mathrm{FK}^{\pi(k)}(\mathbf{x}^{\mathrm{ik}}_n) - \mathbf{p}^k_n\big\|$, where the chain for tracker k ends at joint $\pi(k)$, on which the device is worn.

From gap to confidence. The measured distance is site-specific, whereas confidence is required at every joint. Propagation is therefore performed using the constraint label,

$$\rho^j_n = \frac{\rho^k_n}{\omega(c^j_n)}, \qquad \kappa^j_n = \exp\Big(-\frac{\rho^j_n}{\ell^j\,\sigma}\Big), \tag{9}$$

for every joint reached by the chain to tracker k; unreached INHERITED joints take $\kappa_n^j = 0$. Here ω depends only on the label: it is one for tracker-provided rotations and smaller for FABRIK-inferred rotations, whose bone direction leaves twist unconstrained. The exponential normalises the miss by the joint chain length ℓ^j and a calibration-fixed scale σ. A floor ϵ then turns the trust into the amplitude the denoiser is allowed at joint j, leaving no joint frozen,

$$m_n^j = \epsilon + (1-\epsilon)\,\kappa_n^j, \qquad \epsilon = 0.05. \tag{10}$$

Confidence measures how well a joint is observed, not how much it needs to change: that is already carried by the residual the denoiser learns. Alongside m_n, the solve yields $\boldsymbol{\delta}_n = \mathcal{Z}\big(\mathbf{x}_n^{\mathrm{ik}}\big) - \mathcal{Z}\big(\bar{\mathbf{x}}_n\big)$, where $\mathcal{Z}$ standardises the solved skeleton and prior channel-wise into the denoiser space, making their discrepancy a signed direction. This is used as conditioning rather than a target.

Confidence through a reconfiguration. A tracker that drops out and recovers re-enters the solve between two consecutive frames, handing it a target the prior never anticipated, and the resulting pose discontinuity travels through the body when the site is near the root. We therefore smooth the target position itself, so that the predicted motion stays continuous: the solve receives an interpolation $\alpha_n^k \in [0,1]$ between what the prior places at the site and what the tracker reports,

$$\tilde{\mathbf{R}}_n^k = \mathrm{slerp}\left(\bar{\mathbf{R}}_n^{\pi(k)},\, \mathbf{R}_n^k,\, \alpha_n^k\right), \qquad \tilde{\mathbf{p}}_n^k = (1-\alpha_n^k)\,\bar{\mathbf{p}}_n^{\pi(k)} + \alpha_n^k\,\mathbf{p}_n^k, \tag{11}$$

The weight follows the bit at a bounded rate,

$$\alpha_n^k = \alpha_{n-1}^k + \mathrm{clip}\left(a_n^k - \alpha_{n-1}^k,\, -r_n^k,\, r_n^k\right), \qquad r_n^k = \frac{r_0}{1+d_n^k}, \tag{12}$$

where r_0 is a rate constant and $d_n^k = \mathrm{smoothstep}\left(\theta_n^k;\, \theta_{\mathrm{lo}},\, \theta_{\mathrm{hi}}\right) \in [0,1]$ increases with the geodesic angle $\theta_n^k = \arccos\left(\frac{1}{2}(\mathrm{tr}(\bar{\mathbf{R}}_n^{\pi(k)\top}\mathbf{R}_n^k) - 1)\right)$ between the prior and measured rotations. After a long dropout, this angle may be large; the resulting lower rate lengthens the transition when the returning measurement would cause the largest jump.

3.4 Confidence-Gated Residual Diffusion

The denoiser corrects only what the prior got wrong, and the confidence decides how much of that error it may claim. The variable it diffuses is therefore the gated residual:

$$\mathbf{r}_n = \mathbf{m}_n \odot \left(\mathbf{x}_n - \bar{\mathbf{x}}_n\right), \tag{13}$$

Noising follows the schedule of §3.1 with $\mathbf{r}_n$ in place of $\mathbf{z}_0$, and a transformer f_θ inverts it,

$$\hat{\mathbf{r}}_n = f_\theta\big(\mathbf{z}_t,\, t,\, C_n\big), \tag{14}$$

where the conditioning C_n gathers the predictor's context $(\mathbf{x}_{n-L:n-1},\, \bar{\mathbf{x}}_{n:n+H})$, the solve's output $\mathbf{v}_n = (\boldsymbol{\delta}_n,\, \mathbf{m}_n,\, \mathbf{c}_n)$, and the frame's tracker record, $\tilde{\mathbf{S}}_n = [\,\tilde{\mathbf{p}}_n,\, \phi(\tilde{\mathbf{R}}_n),\, \mathbf{a}_n\,]$. Joint tokens read the tracker tokens once by cross-attention, absent trackers masked out, after which a spatio-temporal transformer attends among the joints and along each joint's own trajectory. Training minimises the residual term together with a geometric term on the reconstructed pose, $\mathcal{L} = \mathcal{L}_{\mathrm{res}}(\mathbf{r}_n, \hat{\mathbf{r}}_n) + \lambda\,\mathcal{L}_{\mathrm{geo}}(\bar{\mathbf{x}}_n + \hat{\mathbf{r}}_n, \mathbf{x}_n)$, with random noise added to the tracker readings to simulate the drift and jitter of a worn device, so that $\mathbf{m}_n$ is exercised over the range it will meet at deployment.

Finally, each 6-D pair is re-orthonormalised into a rotation and the pose is read as $\hat{\mathbf{x}}_n = \bar{\mathbf{x}}_n + \hat{\mathbf{r}}_n$.

3.5 Training-Free Posterior Correction

The pose is answerable only to the trackers, but once it is replayed in an interactive simulation the contact geometry lies outside the motion prior and motion data cannot supply it, so we impose it at sampling time as a posterior correction. At each diffusion step, the model already commits to a $\hat{\mathbf{z}}_0$, and posterior sampling moves that estimate along the gradient of a measurement likelihood before the step is completed(Chung et al., 2023; Tian et al., 2025).

In this task we model that measurement as follows. Let $\mathbf{b}_q(\cdot)$, $q = 1, \ldots, Q$, be the body points differentiable forward kinematics places along the limbs, and let τ be the clearance their thickness demands. Penetration is what the signed distance field of §3.1 reports below that clearance,

$$E\big(\mathbf{x}\big) = \sum_{q=1}^{Q} \Big(\max\left\{0,\; \tau - s(\mathbf{b}_q(\mathbf{x}))\right\}\Big)^2, \tag{15}$$

which vanishes wherever the body stays outside the geometry and grows with the square of the depth elsewhere, so that its gradient pushes a point out in proportion to how far it has sunk in.

The correction acts on the residual rather than the pose. Let $\hat{\mathbf{r}}_n^{(m)}$ be an inner iterate started from it, $\hat{\mathbf{r}}_n^{(0)} = \hat{\mathbf{r}}_n$. The posterior score splits into a prior and a measurement term, $\nabla \log p(\mathbf{z}_0 \mid \mathbf{z}_t, \mathbf{y}) = \nabla \log p(\mathbf{z}_0 \mid \mathbf{z}_t) + \nabla \log p(\mathbf{y} \mid \mathbf{z}_0)$. Parameterised as in PostEdit(Tian et al., 2025), the first density is $\mathcal{N}(\hat{\mathbf{r}}_n, \sigma_t^2 \mathbf{I})$ and the scene is a measurement of likelihood $\propto \exp(-E/2\gamma^2)$, so ascending the posterior is a descent on two quadratics,

$$\hat{\mathbf{r}}_n^{(m+1)} = \hat{\mathbf{r}}_n^{(m)} - h\, \mathbf{m}_n \odot \nabla_{\hat{\mathbf{r}}_n^{(m)}} \left(\underbrace{\frac{\left\| \hat{\mathbf{r}}_n^{(m)} - \hat{\mathbf{r}}_n \right\|^2}{2\sigma_t^2}}_{\text{motion fidelity}} + \underbrace{\frac{E\left(\bar{\mathbf{x}}_n + \hat{\mathbf{r}}_n^{(m)}\right)}{2\gamma^2}}_{\text{scene consistency}} \right). \tag{16}$$

The first term keeps the iterate near the denoiser prediction, while the second pushes the body out of the geometry, preserving the predicted motion without scene penetration. After M correction steps, the result replaces $\hat{\mathbf{r}}_n$ and is re-noised to level $t-1$ for the next denoising step. Applying the correction at every step, rather than only at the end, lets subsequent steps pull the pose back onto the motion manifold. Unlike PostEdit(Tian et al., 2025), we add no Langevin noise, keeping each frame a deterministic function of its inputs.

Nothing in this subsection is learned. The denoiser is trained once on furniture-free captures and never sees a scene, and a session that changes the room changes only the field that Eq. 15 queries. Distances are read from colliders the Unity runtime already maintains, so the correction adds no model and costs one backward pass through forward kinematics.

4 Experiments

4.1 Dataset and Metrics

We evaluate on AMASS(Mahmood et al., 2019) under the cross-dataset A-P2 protocol of RPM(Barquero et al., 2025), resampled to 30 fps: 11,634 training sequences from twelve subsets and 138 test sequences from HumanEva and Transitions. Tracker readings are synthesised from ground-truth joints and perturbed with random noise to simulate tracker drift, and dropped at random intervals to simulate signal loss and recovery.

Accuracy is reported as MPJRE (deg), MPJPE (cm), and MPJVE (cm/s), and smoothness as jitter, the mean positional jerk($10^2\,\mathrm{m/s^3}$). Because sequence-averaged jitter obscures reconfiguration events, we report PJ and AUJ using RPM's evaluator over the half second after each dropout and recovery(Barquero et al., 2025; 2024), evaluating the two phases separately. Scene consistency is measured by penetration depth into the convex-primitives scene, using the signed distance field.

4.2 Comparison with the State of the Art

Table 1 compares STANCE with the state of the art in the two configurations consumer hardware ships. The baselines differ between the two blocks because each supports only one configuration; STANCE does not. Both STANCE rows come from a single set of weights, evaluated against specialists on the configuration each was designed for. At six points it ranks first on all four metrics; at three it ranks first on MPJRE, MPJVE and jitter, and second on MPJPE by only 0.08 cm.

The two cost columns show the price of this generality. Five of eight three-point baselines run faster than 29 ms, but all underperform STANCE on MPJRE, MPJVE, and jitter. Only one baseline improves on MPJPE, at $4.7\times$ more parameters and $2.7\times$ more time. Capacity is not decisive either: the two largest baselines are slower and, aside from MPJPE, less accurate. Although STANCE is heavier than the fastest baselines, its 29 ms still meets the frame budget of 30 fps capture.

Table 2 shows STANCE better on both rejoining metrics, 6.89 and 11.86 versus 7.31 and 11.93 for RPM-Smooth, while reducing MPJPE by 1.11 cm and MPJVE by 5.51 cm/s. The exception is dropout AUJ, 22.79 versus 6.43. Yet peak jerk remains the lowest, indicating no abrupt jump but a slower settling: as α_n^k decays over several frames, the same motion is spread over a wide window, increasing AUJ.

Table 1: **Comparison with the state of the art on A-P2.** The upper and lower blocks compare baselines under three- and six-point tracking, respectively. **Bold** and underline denote the best and second-best values in each block.

Method	Params (M)	Lat. (ms)	MPJRE↓ (deg)	MPJPE↓ (cm)	MPJVE↓ (cm/s)	Jit.↓ ($10^2\mathrm{m/s^3}$)
3-point: head and hands						
AvatarPoser(Jiang et al., 2022a)	**4.1**	80.7	5.18	6.65	32.74	2.67
AGRoL(Du et al., 2023)	7.5	18.3	5.21	6.60	50.11	9.65
EgoPoser(Jiang et al., 2025)	**4.1**	**1.7**	5.49	6.88	34.19	3.52
SAGE(Feng et al., 2024)	120.3	72.4	5.36	6.17	34.62	3.68
AvatarJLM(Zheng et al., 2023)	63.8	77.5	5.21	**5.23**	22.72	1.77
HMD-Poser(Dai et al., 2024)	17.4	4.9	5.17	5.58	24.51	2.15
RPM-Reactive(Barquero et al., 2025)	9.9	4.3	5.15	6.56	27.16	1.55
RPM-Smooth(Barquero et al., 2025)	9.9	4.3	5.40	7.14	30.36	1.39
STANCE (ours)	13.7	29.0	**4.66**	5.31	**21.23**	**0.96**
6-point: head, hands, hip and feet						
Analytic IK	—	—	15.02	3.13	10.44	1.45
SparsePoser(Ponton et al., 2023)	**7.6**	**9.3**	3.94	2.79	22.76	2.18
+ learned IK	7.7	15.6	3.73	2.12	17.81	2.17
+ learned IK, $H{=}7$	7.7	15.6	3.53	1.82	11.51	1.81
STANCE (ours)	13.7	29.1	**3.37**	**1.07**	**8.29**	**1.16**

Table 2: **How smoothly a tracker is lost and regained.** Peak jerk (PJ) and AUJ over the half second after a signal drops (*losing*) and after it returns (*rejoining*), with PJ in $10^2\mathrm{m/s^3}$.

				Losing it		Rejoining it	
Method	**MPJPE**	**MPJVE**	**Jit.**	**PJ**	**AUJ**	**PJ**	**AUJ**
HMD-Poser(Dai et al., 2024)	**6.42**	29.35	2.52	9.04	55.58	71.13	174.80
RPM-Reactive(Barquero et al., 2025)	7.80	31.54	1.64	6.06	11.73	10.92	32.81
RPM-Smooth(Barquero et al., 2025)	8.37	34.04	1.44	4.74	**6.43**	7.31	11.93
STANCE (ours)	7.26	**28.53**	**1.00**	**4.51**	22.79	**6.89**	**11.86**

In Figure 3, both baselines leave the hands short of the readings given to them. The kick is the sharper case: neither lifts the leg at all. STANCE meets the readings at the measured joints and, at the unobserved ones, follows a prior the denoiser may move only by the floor ϵ.

4.3 Ablation Study

Table 3: **Ablations.** The upper block removes the predicted horizon from the denoiser context; the lower removes the denoiser and reports the prior alone. Both use A-P2 in the stated configuration.

Variant	MPJRE↓ (deg)	MPJPE↓ (cm)	MPJVE↓ (cm/s)	Jitter↓
Predicted horizon in the denoiser context				
3-point, $H{=}0$	4.763	5.682	26.093	1.060
3-point, $H{=}10$	**4.659**	**5.305**	**21.229**	**0.958**
6-point, $H{=}0$	**3.351**	**1.062**	8.853	1.209
6-point, $H{=}10$	3.374	1.073	**8.285**	**1.162**
Contribution of the denoiser, three-point				
Predictor only	**4.468**	**5.259**	23.348	1.995
Predictor + denoiser	4.659	5.305	**21.229**	**0.958**

As shown in Table 3, withholding the predicted horizon costs little under six-point tracking and a great deal under three: position error rises from 5.305 to 5.682 cm and velocity by more than 4 cm/s. The horizon resolves ambiguity only where joints are under-constrained, and with six trackers the body is already determined. It improves smoothness in both, however, from 1.060 to 0.958 and from 1.209 to 1.162, because correcting a joint along its own trajectory prevents abrupt reversals.

The second block isolates the denoiser's contribution at three points, with the predictor using only head and hands. Static accuracy is slightly better without it, but jitter drops from 1.995 to 0.958 and velocity error by 2.119 cm/s. This reflects the residual design: preserve an accurate prior while correcting its temporal deficiencies.



Figure 3: **Qualitative comparison under three-point tracking.** A dance (a) and a high kick (b), compared with AGRoL (Du et al., 2023) and RPM (Barquero et al., 2025). Both baselines produce footwork inconsistent with the upper body and underpredict how far the kicking leg travels.

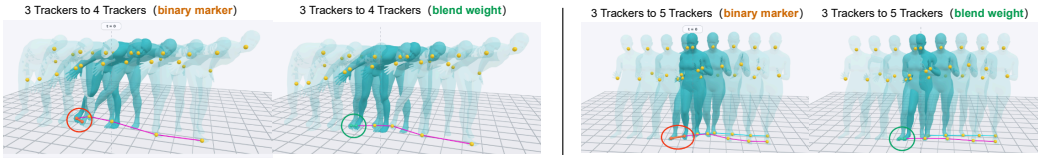

Figure 4: **Foot trackers returning mid-motion.** A throw with one foot tracker returning (left) and a boxing sequence with both returning simultaneously (right). A binary marker causes an immediate jump on recovery, whereas a ramped blend weight restores the feet smoothly.

To assess reconnect robustness, Figure 4 compares a throw with single-foot reconnection and a boxing sequence with dual-foot reconnection. A binary availability mask causes a one-frame snap and abrupt whole-body jitter, whereas our smoothly increasing blend weight α_n^k spreads corrections across frames. Since r_n^k decreases with disagreement, larger discrepancies are corrected more gradually, reducing sudden contact breaks and foot sliding.

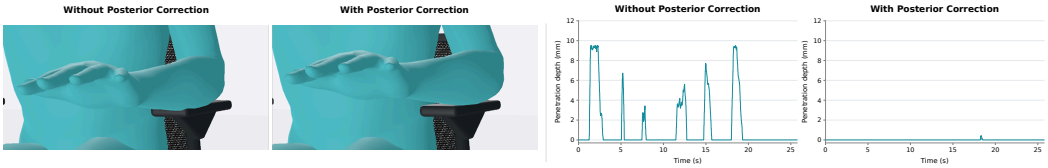

Figure 5: **Armrest tapping with and without posterior correction. Left:** a seated sequence where the left forearm repeatedly taps the armrest. Without correction, the arm penetrates the surface; with correction, it stops on it. **Right:** penetration depth over the same sequence, uncorrected (left) and corrected (right), with one peak per tap.

The metrics above cannot capture scene correction because MPJPE and jitter use furniture-free references. Figure 5 therefore replays a seated sequence where the forearm taps an armrest and measures penetration depth at each contact. Without correction, every tap intersects the geometry because the pipeline has no knowledge of the armrest. With Eq. 16 applied at every denoising step, the arm rests on the surface rather than passing through it. Repeated taps show that the constraint holds consistently across contacts.

5 Conclusion

In this paper, we introduce STANCE, which replaces binary availability with per-joint confidence derived from the residual left by a partial inverse-kinematics solve against tracker targets. This confidence controls how much the denoiser may correct the prior and how gradually a returning tracker is reintegrated. Because it is computed online, a single model supports all tracker configurations at fixed per-frame cost. A training-free posterior correction further prevents scene penetration during sampling, allowing the environment to change between sessions.

AI USE STATEMENT

We used generative AI tools to assist with language editing and translation, literature searches, and the organization and presentation of this manuscript. These tools also provided suggestions on experimental design and result interpretation, and assisted with code development and debugging. We evaluated these suggestions and made the final decisions on the method, experimental protocols, and conclusions. All reported numerical results were obtained by running the experimental code. We checked AI-assisted text and references against the relevant sources and reviewed and tested AI-assisted code before using it in the experiments. The authors take full responsibility for the content of this paper, including material prepared with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

Sections 3.2–3.5 describe the model and training objectives, Section 4.1 specifies the dataset and evaluation protocol, and Appendix A.1 gives implementation details. We plan to publicly release the source code for the proposed method, together with preprocessing scripts, training and inference configurations, and evaluation code. The release will include instructions for obtaining the required datasets from their original providers and preparing them for use with our implementation. These resources are intended to support independent evaluation of the method and reproduction of the reported experiments.

REFERENCES

Sadegh Aliakbarian, Pashmina Cameron, Federica Bogo, Andrew Fitzgibbon, and Thomas J Cashman. Flag: Flow-based 3d avatar generation from sparse observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13253–13262, 2022.

Sadegh Aliakbarian, Fatemeh Saleh, David Collier, Pashmina Cameron, and Darren Cosker. Hmd-nemo: Online 3d avatar motion generation from sparse observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9622–9631, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

German Barquero, Sergio Escalera, and Cristina Palmero. Seamless human motion composition with blended positional encodings. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 457–469, 2024.

German Barquero, Nadine Bertsch, Manojkumar Marramreddy, Carlos Chacón, Filippo Arcadu, Ferran Rigual, Nicky Sijia He, Cristina Palmero, Sergio Escalera, Yuting Ye, and Robin Kips. From sparse signal to smooth motion: Real-time motion generation with rolling prediction models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Géry Casiez, Nicolas Roussel, and Daniel Vogel. 1 euro filter: A simple speed-based low-pass filter for noisy input in interactive systems. In *Proc. SIGCHI Conf. Human Factors in Computing Systems (CHI)*, pp. 2527–2530, 2012.

Angela Castillo, Maria Escobar, Guillaume Jeanneret, Albert Pumarola, Pablo Arbeláez, Ali Thabet, and Artsiom Sanakoyeu. Bodiffusion: Diffusing sparse observations for full-body human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4221–4231, 2023.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023. arXiv:2209.14687.

Peng Dai, Yang Zhang, Tao Liu, Zhen Fan, Tianyuan Du, Zhuo Su, Xiaozheng Zheng, and Zeming Li. Hmd-poser: On-device real-time human motion tracking from scalable sparse observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 874–884, 2024.

Andrea Dittadi, Sebastian Dziadzio, Darren Cosker, Ben Lundell, Thomas J Cashman, and Jamie Shotton. Full-body motion from a single head-mounted device: Generating smpl poses from partial observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11687–11697, 2021.

Kun Dong, Jian Xue, Zehai Niu, Xing Lan, Ke Lv, Qingyuan Liu, and Xiaoyu Qin. Realistic full-body motion generation from sparse tracking with state space model. In *Proceedings of the ACM International Conference on Multimedia*, 2024.

Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 481–490, 2023.

Han Feng, Wenchao Ma, Quankai Gao, Xianwei Zheng, Nan Xue, and Huijuan Xu. Stratified avatar generation from sparse observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 153–163, 2024.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8944–8951. IEEE, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Lei Hu, Zihao Zhang, Chongyang Zhong, Boyuan Jiang, and Shihong Xia. Pose-aware attention network for flexible motion retargeting by body part. In *Proc. IEEE Conf. Virtual Reality and 3D User Interfaces (VR)*, 2024.

Mazeyu Ji, Xuanbin Peng, Fangchen Liu, Jialong Li, Ge Yang, Xuxin Cheng, and Xiaolong Wang. Exbody2: Advanced expressive humanoid whole-body control. *arXiv preprint arXiv:2412.13196*, 2024.

Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *European Conference on Computer Vision*, pp. 443–460. Springer, 2022a.

Jiaxi Jiang, Paul Streli, Manuel Meier, and Christian Holz. Egoposer: Robust real-time egocentric pose estimation from sparse and intermittent observations everywhere. In *European Conference on Computer Vision*, pp. 277–294. Springer, 2025.

Yifeng Jiang, Yuting Ye, Deepak Gopinath, Jungdam Won, Alexander W Winkler, and C Karen Liu. Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation. In *SIGGRAPH Asia Conference Papers*, pp. 1–9, 2022b.

Sunmin Lee, Sebastian Starke, Yuting Ye, Jungdam Won, and Alexander Winkler. Questenvsim: Environment-aware simulated motion tracking from sparse sensors. In *ACM SIGGRAPH 2023 conference proceedings*, pp. 1–9, 2023.

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):1–16, 2015.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11461–11471, 2022.

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5442–5451, 2019.

Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. Imuposer: Full-body pose estimation using imus in phones, watches, and earbuds. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2023.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proc. Int. Conf. Machine Learning (ICML)*, pp. 8162–8171, 2021.

Jose Luis Ponton, Haoran Yun, Andreas Aristidou, Carlos Andujar, and Nuria Pelechano. Sparseposer: Real-time full-body motion reconstruction from sparse data. *ACM Transactions on Graphics*, 43(1):1–14, 2023.

Jose Luis Ponton, Eduard Pujol, Andreas Aristidou, Carlos Andujar, and Nuria Pelechano. Dragposer: Motion reconstruction from variable sparse tracking signals via latent space optimization. *Computer Graphics Forum*, 2025. Eurographics 2025.

Bo Qian, Zhenhuan Wei, Jiashuo Li, and Xing Wei. Reliaavatar: A robust real-time avatar animator with integrated motion prediction. *arXiv preprint arXiv:2407.02129*, 2024.

Yiming Ren, Chengfeng Zhao, Yannan He, Peishan Cong, Han Liang, Jingyi Yu, Lan Xu, and Yuexin Ma. LiDAR-aid inertial poser: Large-scale human motion capture by sparse inertial and LiDAR sensors. *IEEE Trans. Visualization and Computer Graphics*, 29(5):2337–2347, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. In *International Conference on Learning Representations*, 2023.

Feng Tian, Yixuan Li, Yichao Yan, Shanyan Guan, Yanhao Ge, and Xiaokang Yang. Postedit: Posterior sampling for efficient zero-shot image editing. In *International Conference on Learning Representations*, volume 2025, pp. 579–609, 2025.

Tom Van Wouwe, Seunghwan Lee, Antoine Falisse, Scott Delp, and C Karen Liu. Diffusionposer: Real-time human motion reconstruction from arbitrary sparse sensors using autoregressive diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2513–2523, 2024.

Alexander Winkler, Jungdam Won, and Yuting Ye. Questsim: Human motion tracking from sparse sensors with simulated avatars. In *SIGGRAPH Asia*, pp. 1–8, 2022.

Songpengcheng Xia, Yu Zhang, Zhuo Su, Xiaozheng Zheng, Zheng Lv, Guidong Wang, Yongjie Zhang, Qi Wu, Lei Chu, and Ling Pei. Envposer: Environment-aware realistic human motion estimation from sparse observations with uncertainty modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Dongseok Yang, Doyeon Kim, and Sung-Hee Lee. LoBSTr: Real-time lower-body pose prediction from sparse upper-body tracking signals. *Computer Graphics Forum*, 40(2):265–275, 2021.

Dongseok Yang, Jiho Kang, Lingni Ma, Joseph Greer, Yuting Ye, and Sung-Hee Lee. Divatrack: Diverse bodies and motions from acceleration-enhanced three-point trackers. *Computer Graphics Forum*, 2024. Eurographics 2024.

Yongjing Ye, Libin Liu, Lei Hu, and Shihong Xia. Neural3Points: Learning to generate physically realistic full-body motion for virtual reality users. *Computer Graphics Forum*, 41(8):183–194, 2022.

Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics*, 40(4):1–13, 2021.

Shuting Zhao, Zeyu Xiao, and Xinrong Chen. Kinest: A kinematics-guided spatiotemporal state space model for human motion tracking from sparse signals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Xiaozheng Zheng, Zhuo Su, Chao Wen, Zhou Xue, and Xiaojie Jin. Realistic full-body tracking from sparse observations via joint-level modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14678–14688, 2023.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5745–5753, 2019.

A Appendix

A.1 Implementation Details

The predictor is a four-layer transformer encoder (4 heads, feed-forward width 1024) reading $L = 10$ past poses and an eleven-frame window of the three core trackers to produce the current pose and $H = 10$ future frames. It is trained for 100k steps with AdamW (learning rate 3×10^{-4}, weight decay 10^{-4}), gradients clipped at unit norm and the learning rate decayed thirtyfold after 50k steps.

The denoiser is four DiT blocks (width 192, 6 heads) over 50 forward steps on a cosine schedule(Nichol & Dhariwal, 2021) with $\mathbf{z}_0$-prediction, trained in bfloat16 for 100k steps at batch size 256 with AdamW (learning rate 10^{-4}) and EMA decay 0.995. Configurations are drawn uniformly from the eight subsets of $\mathcal{K}_o$, and tracker readings carry synthetic drift and noise as described in §3.4. Both modules are then fine-tuned jointly for 20k steps at 10^{-6} and 10^{-5}. Inference runs ten DDIM steps(Song et al., 2021) with $\eta = 0$.

FABRIK runs two iterations per chain. In Eq. 10 the floor is $\epsilon = 0.05$. The scale σ is a fraction of chain length rather than a distance, taken as the median relative endpoint residual over a held-out pass of the trained predictor and floored at 10^{-3}; the measured median falls below that floor, so $\sigma = 10^{-3}$. The posterior correction of Eq. 16 runs $M = 5$ inner steps of size $h = 2 \times 10^{-4}$ at each of the ten diffusion steps, with scene tolerance $\gamma = 2.8$ mm.

A.2 One Model, Every Configuration

Table 4: **All eight configurations from one set of weights.** No retraining and no per-configuration tuning; only the availability vector $\mathbf{a}_n$ changes between rows.

#	Optional trackers	MPJRE↓ (deg)	MPJPE↓ (cm)	MPJVE↓ (cm/s)	Jitter↓
3	—	4.659	5.305	21.229	**0.958**
4	Hip	4.006	3.363	16.574	1.015
4	Left foot	4.488	3.656	16.700	1.040
4	Right foot	4.473	3.669	16.511	1.049
5	Both feet	4.256	2.220	12.517	1.119
5	Hip, left foot	3.620	2.022	12.065	1.089
5	Hip, right foot	3.659	2.129	11.959	1.094
6	Hip, both feet	**3.374**	**1.073**	**8.285**	1.162

Table 4 evaluates one checkpoint eight times with only $\mathbf{a}_n$ varying. Position error falls monotonically from 5.31 to 1.07 cm, so the model absorbs additional measurements without configuration-specific variants.

The four-point rows quantify something prior work has only assumed. Adding a pelvis tracker or a foot tracker is the same hardware decision for the user, yet the pelvis gives lower position (3.36 against ∼3.66 cm) and rotation (4.01° against $\sim 4.49^\circ$) error, and the hierarchy survives at five points, where a pelvis with one foot (2.02 cm) beats both feet (2.22 cm) despite constraining one limb rather than two. That the pelvis is privileged is implicit in earlier designs—LoBSTr(Yang et al., 2021) makes it the sole required lower-body input, three-point methods spend capacity estimating it(Jiang et al., 2022a; Yang et al., 2024), full-body systems root the kinematic chain there(Ponton et al., 2023)—but none measures it against another sensor at equal count. The margin may be method-dependent; its direction follows from skeletal topology rather than from our architecture.

Jitter moves the other way, from 0.96 to 1.16 as trackers are added. Under three-point tracking the pose is largely governed by a smooth prior, whereas six trackers earn high confidence and admit their own noise. We report unfiltered numbers: a one-euro(Casiez et al., 2012) or Kalman stage would suppress the increase, and omitting it keeps the comparison about the framework.

Figure 6 shows where the extra measurements are spent. The same weights reconstruct a walk and a high kick from three trackers and from six, and the visible difference is confined to the lower body: the placement of the feet in the walk, and how far the kicking leg travels. Two properties of the design account for that. The predictor never reads the optional trackers, so the prior is the same pose in both columns; and the chains the added trackers close terminate below the waist, so the joints

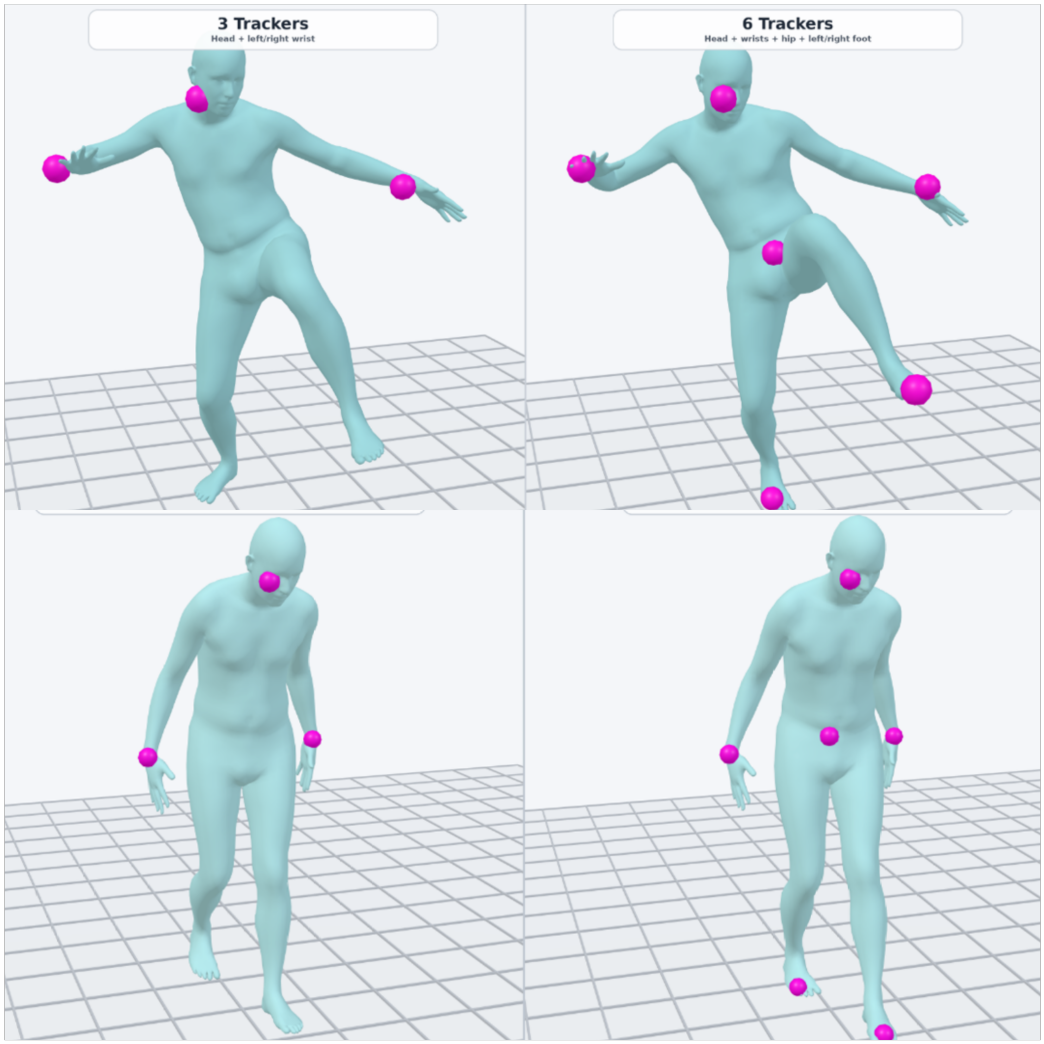

Figure 6: **One model at three and at six trackers.** A high kick (top) and a walk (bottom), reconstructed by the same weights from both configurations. Spheres mark the active trackers. The upper bodies are unchanged; what differs is the stance of the walk and the height the kicking leg reaches.

whose confidence they raise are the legs. A model that zero-pads absent channels offers no such separation, since every channel shares one latent space and a foot observation may perturb an arm. What the three-tracker column loses is precision along the branch it cannot see, not the identity of the motion: a lower kick rather than a different action.

A.3 Additional Comparisons

DragPoser is the other route to configuration independence, so it belongs in the comparison, but no checkpoint is public and its benchmark is not the one Table 1 uses. We therefore evaluate on its corpus and quote the published figures, keeping the result apart so that nothing is read across a boundary the numbers do not cross.

STANCE lowers the rotation error from $22.00°$ to $6.19°$ at three trackers and from $7.66°$ to $4.34°$ at six, at ten times the parameter count and a comparable average cost per frame. On average is the operative qualifier. DragPoser reaches a pose by optimising in a latent space until its constraints are satisfied, so a frame costs whatever that frame demands: 25.3 ms typically and 173.8 ms at the iteration limit, a spread of nearly seven to one. STANCE runs the same ten denoising steps whatever

Table 5: **Comparison with DragPoser on its own benchmark.** No checkpoint is public, so we evaluate STANCE on DragPoser's dataset and compare against the figures its authors report; the corpus therefore differs from Table 1, and jitter is not reported in(Ponton et al., 2025). †DragPoser optimises until its constraints are met, so the frame cost varies: 25.3 ms on average, 173.8 ms at the iteration limit its authors use.

Method	Params (M)	Lat. (ms)	MPJRE↓ (deg)	MPJPE↓ (cm)	MPJVE↓ (cm/s)
DragPoser(Ponton et al., 2025), 3-pt	**1.4**	**25.3**†	22.00	15.50	34.00
STANCE (ours), 3-pt	13.7	29.0	**6.19**	**7.61**	**20.52**
DragPoser(Ponton et al., 2025), 6-pt	**1.4**	**25.3**†	7.66	2.18	10.70
STANCE (ours), 6-pt	13.7	29.1	**4.34**	**1.60**	**9.12**

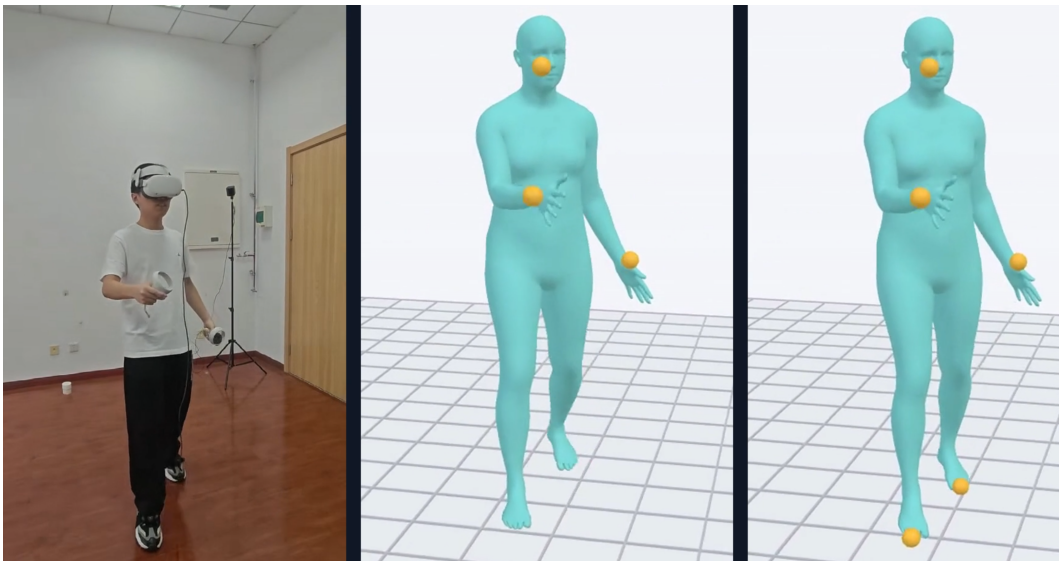

Figure 7: **Live capture on consumer hardware.** A session driven by a headset, two controllers and two foot trackers (right), beside the three-point reconstruction obtained by withholding the foot readings from the same recording (centre). Spheres mark the active trackers.

the frame contains, so 29 ms is its average and its bound at once. For capture that is the more useful guarantee, since it is the late frame rather than the mean that breaks a real-time loop.

A.4 Live Capture on Consumer Hardware

The evaluation so far runs on synthesised readings. Figure 7 closes that gap with a live session on consumer hardware, a headset and two controllers joined by two foot trackers, reconstructed by the same weights used throughout. Withholding the foot readings from the same stream gives the three-point reconstruction beside it, so the two columns differ only in what the model was allowed to see. In still frames they agree closely.

Where they separate is instructive. A forward kick performed with a stationary upper body is recovered at five trackers and missed at three: with no lower-body sensor, leg motion can only be inferred from upper-body cues, and a kick the torso does not commit to leaves nothing to infer from. When the upper body does swing into the kick, as in Figure 6, three trackers recover it. Optional trackers are therefore worth exactly what the prior cannot anticipate, which is the quantity the confidence of §3.3 is built to measure.