EgoForce: Robust Online Egocentric Motion Reconstruction via Diffusion Forcing

Inwoo Hwang[1], **Donggeun Lim**[1], **Hojun Jang**[1], **Young Min Kim**[1,2†]

[1] Dept. of Electrical and Computer Engineering, Seoul National University

[2] Interdisciplinary Program in Artificial Intelligence and INMC, Seoul National University

https://inwoohwang.me/EgoForce

Abstract

With recent advances in embodied agents and AR devices, egocentric observations are readily available as input for real-world interactive online applications. However, egocentric viewpoints can only sporadically observe hands, in addition to the estimated head trajectory. We propose **EgoForce**, an online framework for reconstructing long-term full-body motion from noisy egocentric input. While existing generative frameworks can robustly handle noisy and sparse measurements, they assume a fixed-length observation window is available and are thus not suitable for real-time applications. Faster inference often relies on autoregressive prediction, sacrificing robustness. In contrast, we adopt a diffusion-based method with a temporally asymmetric noise schedule inspired by Diffusion Forcing. Specifically, our approach models temporally evolving uncertainty and incrementally denoises states as new streaming observations arrive. Combined with a noise-robust imputation strategy, EgoForce progressively generates stable and coherent full-body motion under strict causal constraints. Experiments demonstrate that our online framework outperforms existing online and offline methods, enabling long-horizon, full-body motion reconstruction in challenging egocentric scenarios.

1 Introduction

As an increasing number of datasets capture diverse daily interactions from AR glasses or embodied agents [51, 12, 35, 29, 2, 11, 10, 27], accurate reconstruction of human motion can accelerate advances in applications such as augmented and virtual reality, daily-life assistance, and real-time human interaction [42]. Such dataset typically contains the locations of the head-mounted devices and, when interacting hands are within the limited field of view, their locations as well, resulting in sparse and noisy body-part observations. Additionally, to fully leverage real-time data in interactive applications, the reconstruction should progressively track motions over a long time horizon without access to future observations. These sparse, noisy, and strictly causal setups pose unique challenges that fundamentally differ from offline motion recovery, which assumes access to a clean, complete temporal context.

Existing approaches struggle to satisfy competing requirements. Online models that satisfy causality are typically trained as deterministic regressors [3, 54, 23]; under sparse or corrupted observations, such regressors converge to the dominant conditional mode and smooth out high-frequency motion details, with no mechanism to represent the uncertainty of unobserved body parts. Conversely, recent diffusion-based generative models have proven effective at modeling uncertainty and recovering

†Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



Figure 1: EgoForce reconstructs full-body motion over long sequences in a strictly online manner. By progressively refining future predictions using both incoming observations and the generated motion history, it preserves temporal coherence and effectively mitigates long-term drift. We visualize two motion sequences, consisting of 610 and 660 frames, respectively.

missing signals [48, 8, 35, 15]; however, they typically rely on offline sequence-level denoising or autoregressive window-level sampling. In these scenarios, the denoising process recovers motion within a fixed temporal window, either offline or through autoregressive window-level sampling, rather than progressively refining a persistent per-frame streaming state, which is incompatible with streaming or frame-level online applications. This leaves a critical mismatch between the modeling capacity required for robust egocentric motion reconstruction and the strict latency constraints of online systems.

To bridge this gap, we formulate online egocentric motion reconstruction as a causal generation problem. As the generative model processes the motion sequence within a temporally shifted window at each time step, the motion reconstruction is conditioned only on the past and current egocentric observations. In the absence of future observations, unseen motion is estimated stochastically rather than deterministically, with uncertainty increasing monotonically with temporal distance from the present. As new observations become available, the model traverses the temporal window and updates its uncertainty estimates accordingly, thereby refining its predicted motion to smoothly connect with past motion and respect the current egocentric measurements.

Our formulation preserves the expressive power of diffusion models while making them compatible with online inference. We instantiate this formulation inspired by Diffusion Forcing [5], which controls the denoising schedule of diffusion-based generative models to process flexible constraints. Under strict causal constraints, we adapt temporally asymmetric frame-wise noise levels in which the generated past motion is conditioned on as lightly re-noised history, the present is partially observed through egocentric measurements, and uncertainty increases toward the future. The denoising process incrementally refines the current prediction using a small number of denoising steps at each time step as new streaming egocentric observations become available. In addition, we incorporate a noise-robust imputation strategy that anchors reliable observations while allowing the model to infer plausible motion for unobserved joints. Together, these components enable stable and coherent full-body motion reconstruction from sparse and corrupted egocentric inputs in an online manner.

Our main contributions can be summarized as follows: 1) We formulate online egocentric motion reconstruction as a causal generation problem with temporally evolving uncertainty. 2) We adapt diffusion-based generative modeling to the online setting by restricting denoising to future motion states through Diffusion Forcing. 3) We demonstrate stable and robust full-body motion reconstruction from sparse and noisy egocentric signals under online constraints over long time horizons.

2 Related Works

2.1 Human Motion from Control Inputs

Egocentric motion reconstruction can be framed as full-body motion reconstruction from sparse input signals. Many previous works that generate motion from partial trajectories or keyframes assume offline access to complete and clean observations [45, 13, 38, 25, 19, 50, 26, 9, 32, 14], such that models can exploit future information to maintain global coherence of motion. Recent advancements enhance robustness against noisy or missing inputs [20, 21] by adapting imputation techniques. However, these formulations remain inapplicable to online interactive applications. While RPM [3] moves toward online reconstruction through a causal transformer, it is trained as a

deterministic regressor and therefore tends to collapse to the dominant conditional mode, smoothing out the high-frequency details that stochastic generative models can represent.

2.2 Egocentric Motion Reconstruction

The development of head-mounted devices [29] has introduced diverse datasets of egocentric measurements paired with full-body motions [51, 35, 12]. Prior approaches to egocentric motion reconstruction typically adopt causal sequence models, such as transformers [23, 54, 24], to enable online applications. However, they are trained as deterministic regressors that map each observation to a single pose; given the noisy and sparse egocentric input for full-body reconstruction [15, 17], such regressors converge to the dominant conditional mode and smooth out high-frequency details, lacking the generative capacity to model the inherent uncertainty of human motion. In contrast, diffusion-based generative models [28] have demonstrated strong robustness by incorporating a diffusion sampling process with guidance mechanisms [48] or conditioning on egocentric images and control signals [35]. Recent work further extends diffusion-based motion reconstruction to causal settings through autoregressive window-level inpainting [15]. However, these diffusion-based methods either perform offline inference over the full observation window or rely on autoregressive window-level sampling, whereas our goal is per-frame streaming reconstruction with incremental refinement as each new observation arrives. In particular, HMD^2 [15] is also a causal diffusion model, but it differs from ours in three respects: it applies a uniform diffusion timestep to the entire window, it runs a complete denoising chain from Gaussian noise for every new block so that poses are emitted block by block, and it inpaints previously generated motion only at inference time without experiencing this operation during training. In contrast, EgoForce assigns heterogeneous frame-wise noise levels within a single rolling buffer, emits a pose as soon as each observation arrives, advances the buffer by only a few denoising steps per update, and is trained to condition on noisy motion history. To bridge this gap, EgoForce formulates online full-body motion reconstruction as a generative framework that enables progressive inference under sparse and noisy observations, combining the causality of online systems with the robust generative capacity of diffusion models.

2.3 Causal Modeling with Diffusion Models

Recent diffusion models [41, 16] extend their capability to fulfill the requirements of streaming content generation [47, 31, 6, 40, 53, 49] and online decision making [46, 44, 7]. TEDi [52] and Rolling Diffusion Models [39] employ a rolling-window formulation, in which diffusion models generate sequences by repeatedly denoising a short temporal window. More recent efforts reformulate diffusion as a strictly causal or autoregressive process by introducing frame-wise varying noise levels and causal sampling schedules. Diffusion Forcing [5] assigns independent noise levels to individual frames during training, enabling strict causal generation while preserving the expressive power of diffusion models. SDP [22] applies a streaming diffusion framework to robotic policy learning, enabling online action generation. Autoregressive diffusion variants [43] further factorize denoising across time steps to support online inference. Our work extends this line of research to online egocentric motion reconstruction. By explicitly modeling the temporal evolution of uncertainty online, EgoForce effectively preserves the expressive capacity of generative diffusion while satisfying the strict causal constraints of real-world egocentric systems.

3 Method

We present an online, causal framework for reconstructing full-body human motion from sparse egocentric observations. We first formulate the problem setting and causal constraints (Section 3.1), then describe the training procedure based on frame-wise noise corruption under causal conditioning (Section 3.2), and finally introduce the progressive denoising strategy for causal online inference (Section 3.3).

3.1 Problem Formulation

We consider an **online egocentric setting** where an image stream $\mathcal{I} = \{\mathbf{I}_1, \ldots, \mathbf{I}_n\}$ is captured by a head-mounted device. From each image $\mathbf{I}_t$, we extract a sparse set of kinematic control signals.

Specifically, at time step t, the kinematic control signal $\mathbf{c}_t$ is defined as:

$$\mathbf{c}_t = \left(\mathbf{h}_t, (\mathbf{w}_t^l, \mathbf{v}_t^l), (\mathbf{w}_t^r, \mathbf{v}_t^r)\right), \tag{1}$$

where $\mathbf{h}_t \in SE(3)$ denotes the global head pose, represented by translation and a continuous 6D rotation representation, and $\mathbf{w}_t^{l/r} \in SE(3)$ represents the left and right wrist poses estimated from $\mathbf{I}_t$ or provided as proxy sparse controls depending on the benchmark. To account for frequent self-occlusions and the limited field-of-view of egocentric cameras, we incorporate visibility indicators $\mathbf{v}_t^{l/r} \in \{0, 1\}$. Here, $\mathbf{v}_t^{l/r} = 1$ indicates that the corresponding wrist pose is reliably observed, while $\mathbf{v}_t^{l/r} = 0$ denotes missing data.

Our goal is to learn a causal generative model $\mathcal{G}$ that reconstructs the current body pose $\mathbf{x}_t$ in an online manner where each pose $\mathbf{x}_t \in \mathbb{R}^d$ represents the full-body configuration at time step t with a **strict causal constraint**: the prediction of $\mathbf{x}_t$ must depend solely on the history of previously generated motions $\mathbf{x}_{<t}$ and the observed control signals up to the current time step, $\mathbf{c}_{\leq t}$ and $\mathbf{I}_{\leq t}$. Accordingly, we formulate the generation process by modeling the conditional distribution of the current pose $\mathbf{x}_t$ given the past motion context and available observations

$$p(\mathbf{x}_t \mid \mathbf{x}_{<t}, \mathbf{c}_{\leq t}, \mathbf{I}_{\leq t}). \tag{2}$$

This causal formulation is designed to enable online motion reconstruction, while the probabilistic formulation maintains robustness against the sparse and intermittent nature of egocentric input.

3.1.1 Motion Representation

We represent full-body motion in a canonical coordinate system centered on the head-mounted device (e.g., AR glasses) [48, 35], with the $+z$ axis aligned with its forward direction. Each pose $\mathbf{x}_t$ includes a central reference point corresponding to the device and 22 body joints, all described by global translation and continuous 6D rotation representation [21]. Foot contact indicators for both feet are also incorporated.

3.2 Training Pipeline with Frame-wise Noise Corruption under Causal Conditioning

We encode the temporal context of motion by observing a sequence of full-body 3D motion within the sliding window of length $h+1+f$. Centered at the current time step t, the temporal window contains motions of a fixed history length h and a future prediction horizon f with respect to the current time step t as $\mathbf{X}_t = \{\mathbf{x}_{t-h}^{k_{t-h}}, \ldots, \mathbf{x}_{t+f}^{k_{t+f}}\}$. Note that we are using the subscript t to denote the temporal index of motion frames, while we use the superscript $k \in \{0, \ldots, K\}$ (K denotes the maximum number of diffusion steps) to denote the diffusion time step where the denoising process $\mathbf{x}_t^K \rightarrow \mathbf{x}_t^0$ recovers the clean sequence $\mathbf{x}_t^0$ from a noisy one $\mathbf{x}_t^K$ with K maximum number of diffusion steps.

To train the online framework, we design the frame-wise noise level k_τ of the temporal window to introduce noisy samples for the diffusion model $\mathbf{x}_\tau$ paired with visibility mask $\mathbf{b}_\tau$, which are composed to serve as input to the denoising network $\mathcal{G}$, as illustrated in Figure 2.

3.2.1 Diffusion Model with Frame-wise Noise

Unlike standard diffusion training, which applies a uniform noise level across all frames in a sequence, we assign independent, varying per-frame noise levels, adopting the Diffusion Forcing paradigm [5]. For each training step, we sample a ground-truth motion window $\mathbf{X}_t^0 = \{\mathbf{x}_{t-h}^0, \ldots, \mathbf{x}_{t+f}^0\}$ centered around the current time step t. We define a vector of diffusion noise levels over the temporal window, $\mathbf{k}_t = \{k_{t-h}, \ldots, k_{t+f}\}$, where each element is independently sampled as

$$k_\tau \sim \mathcal{U}\{0, 1, \ldots, K\}, \quad \tau \in [t-h, t+f]. \tag{3}$$

Each ground-truth pose $\mathbf{x}_\tau$ is then perturbed according to its assigned noise level k_τ, producing the noisy pose $\mathbf{x}_\tau^{k_\tau}$. This approach enables the model to denoise motion trajectories under heterogeneous, frame-wise uncertainty, facilitating flexible noise scheduling during inference.

3.2.2 Causal Observation Injection

To strictly enforce online causality during training, we inject sparse body-part observations using a causal visibility mask $\mathbf{b}_\tau \in \{0, 1\}^d$, as illustrated in Figure 2. The mask follows the online

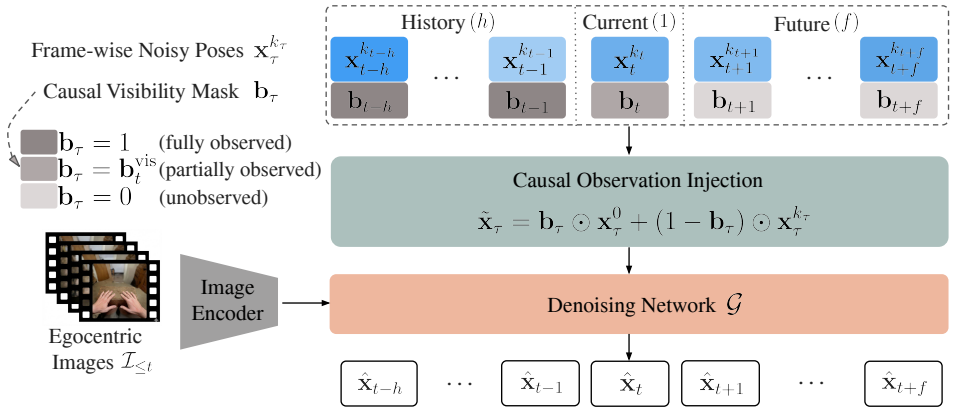

Figure 2: **Training pipeline with frame-wise noise corruption under causal conditioning.** A motion segment centered at time step t is corrupted with heterogeneous diffusion noise k_τ across all frames, including the history, so that past motion conditions the network under noise rather than as a clean prefix. Egocentric causal observations at the current frame are injected, and the denoising network $\mathcal{G}$ is trained to reconstruct the clean motion sequence conditioned on the noisy history and causal egocentric context.

information structure and marks which dimensions are available as context: $\mathbf{b}_\tau = \mathbf{1}$ for past frames ($\tau < t$), whose full-body poses are available from the generated history; $\mathbf{b}_\tau = \mathbf{b}_t^{\text{vis}}$ for the current frame ($\tau = t$), which activates only the head and the visible wrists according to the visibility indicators $\mathbf{v}_t$ while all unobserved joints remain masked; and $\mathbf{b}_\tau = \mathbf{0}$ for future frames ($\tau > t$), which are unobserved. Importantly, availability does not imply cleanliness: following Equation (3), past frames carry their own independently sampled noise levels k_τ, so the history conditions the network under noise rather than as a clean prefix.

We construct the causally injected pose

$$\tilde{\mathbf{x}}_\tau = \begin{cases} \mathbf{x}_\tau^{k_\tau}, & \tau < t \quad \text{(noisy full-body history)} \\ \mathbf{b}_t^{\text{vis}} \odot \mathbf{x}_t^0 + (1 - \mathbf{b}_t^{\text{vis}}) \odot \mathbf{x}_t^{k_t}, & \tau = t \quad \text{(anchored current observations)} \\ \mathbf{x}_\tau^{k_\tau}, & \tau > t \quad \text{(unobserved future)} \end{cases} \tag{4}$$

which anchors the reliable current-frame observations to their clean values while preserving stochastic uncertainty elsewhere. The history is injected as full-body context under its sampled noise level k_τ rather than as a clean prefix, so the network learns to stay consistent with imperfect past motion and current sensor input while predicting plausible future trajectories.

To explicitly inform the network which dimensions are reliable, we concatenate the injected pose with its visibility mask, $\bar{\mathbf{x}}_\tau = [\tilde{\mathbf{x}}_\tau, \mathbf{b}_\tau]$, and construct the augmented motion window $\bar{\mathbf{X}}_t = \{\bar{\mathbf{x}}_{t-h}, \ldots, \bar{\mathbf{x}}_{t+f}\}$.

Relation to Teacher Forcing. Although conditioning on past motion resembles teacher forcing, which trains on a clean ground-truth prefix to predict the next clean state, our training never assumes a clean prefix. Each past frame carries an independently sampled noise level k_τ, so the denoiser learns to condition on imperfect, noisy history while reconstructing the current and future frames for arbitrary noise configurations $\mathbf{k}_t$, as in Diffusion Forcing [5]. This matches our online inference, where the generated history rather than the ground truth is re-noised before conditioning (Section 3.3), and avoids training on a single fixed progressive schedule, which we found to drift over long sequences.

3.2.3 Noise-Robust Causal Observation Injection

To improve robustness to noisy egocentric inputs, we synthetically corrupt the egocentric control signals $\mathbf{c}_t$ into noisy versions $\mathbf{c}_t^{\text{noisy}}$ during training and train a denoiser using the corresponding clean motion data $\mathbf{x}_t$, following [21]. Since the model prediction is unstable with a large number of diffusion steps $k \geq K^*$, a noisy observation signal is applied only during the early diffusion steps up to a threshold K^*, whereas later denoising stages rely solely on diffusion model predictions. The

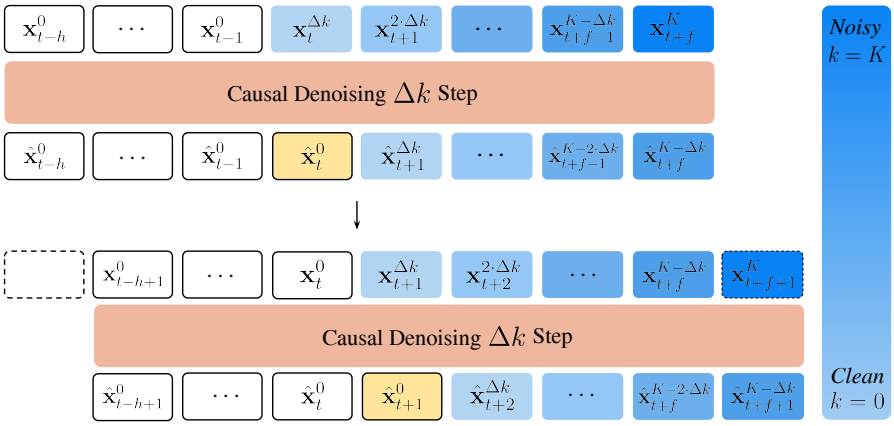

Figure 3: **Causal online inference with progressive denoising refinement.** At each time step, the temporal window is shifted forward to reuse previously denoised states as warm-starts, while a new future frame is initialized with Gaussian noise. Causal egocentric observations are injected, and the denoising network performs a fixed Δk refinement step to fully denoise the current pose while progressively refining future predictions under increasing uncertainty, as detailed in Alg. 1 of the Appendix.

causally injected pose at the current frame is then defined as

$$\tilde{\mathbf{x}}_t = \begin{cases} \mathbf{b}_t^{\text{vis}} \odot \mathbf{x}_t^0(\mathbf{c}_t^{\text{noisy}}) + (1 - \mathbf{b}_t^{\text{vis}}) \odot \mathbf{x}_t^{k_t}, & k_t \geq K^* \\ \mathbf{x}_t^{k_t}, & k_t < K^*. \end{cases} \tag{5}$$

3.2.4 Conditioning and Objective

In addition to the sparse kinematic observations $\mathbf{c}_t$ that are injected directly into the motion representation, the causal egocentric image stream

$$\mathbf{I}_{\leq t} = \{\mathbf{I}_{t-h}, \ldots, \mathbf{I}_t\}$$

is encoded by an image encoder [33] and fused into the denoising network $\mathcal{G}$ via cross-attention. This visual context resolves ambiguities caused by missing joints and provides semantic cues for motion reconstruction.

The model is trained to directly predict the ground truth motion window $\mathbf{X}_t^0$ by minimizing the expected L_2 reconstruction error

$$\mathcal{L} = \mathbb{E}_{\mathbf{X}_t^0, \mathbf{k}} \left[\left\| \mathcal{G}(\bar{\mathbf{X}}_t, \mathbf{k}, \mathbf{I}_{\leq t}) - \mathbf{X}_t^0 \right\|_2^2 \right]. \tag{6}$$

3.3 Causal Online Inference with Progressive Denoising Refinement

After the model is trained to denoise signals under heterogeneous noise conditions, we leverage its behavior and design the inference step for online generation. The inference process is formulated as a **progressive denoising refinement** procedure, with a carefully designed per-frame noise level. The temporal sequence $\mathbf{x}_t \in \mathbf{X}_t$ ensures coherence to the previous generation $\hat{\mathbf{x}}_{t-h:t-1}$, while respecting the current, and possibly noisy, egocentric observations $(\mathbf{c}_t, \mathbf{I}_t)$. Simultaneously, it anticipates future motion trajectories within a consistent prediction horizon f. The process is illustrated in Figure 3 and also described below.

3.3.1 Initializing the Sequence

At the very first step of inference ($t = 0$), the system initializes the entire buffer with Gaussian noise. We denoise the initial sequence to recover a stable trajectory sample. During this stage, the historical frames (h) and the current frame are fully denoised ($k = 0$), while the future horizon (f) is partially denoised to follow the predefined noise for the future horizon as follows.

Noise Scheduling for Future Horizon. To capture the increasing uncertainty across the future prediction horizon f, we assign a noise schedule that increases monotonically with temporal distance. For a frame at a relative offset $i \in \{0, \dots, f\}$ from the current step t, the target noise level k_{t+i} is defined as

$$k_{t+i} = i \cdot \Delta k, \quad \Delta k = \frac{K}{f+1}. \tag{7}$$

This temporal scheduling ensures the current frame $\mathbf{x}_t$ reaches a clean state ($k_t = 0$), while future predictions are increasingly noisy states, with the furthest terminal frame $\mathbf{x}_{t+f}$ approaching the maximum diffusion step K.

3.3.2 Online Inference Pipeline

As time advances from t to $t+1$, the observed motion sequence $\mathbf{X}_t$ is shifted by a single time step, sharing $h+f$ temporal interval. The inference loop consists of three stages below, which is visualized in Figure 3.

1. *Temporal Shift and Warm-start.* As time evolves, the temporal window slides forward by one frame. The oldest history frame $\mathbf{x}_{t-h}$ is discarded, and all latents in the buffer are shifted to the left. These shifted frames serve as high-quality warm-starts because they retain the denoising progress from the previous time step. To maintain the horizon, a new terminal frame $\mathbf{x}_{t+f+1}$ is introduced and initialized with standard normal noise: $\mathbf{x}^K_{t+f+1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

2. *Causal Condition Injection.* Upon receiving new egocentric observations at $t+1$, $(\mathbf{c}_{t+1}, \mathbf{I}_{t+1})$, we update the reliability masks $\mathbf{b}_\tau$ and inject the causal signals following Equation (4): the current-frame observations are anchored from $\mathbf{c}_{t+1}$, while the generated history $\hat{\mathbf{x}}_{\tau<t+1}$ takes the place of the ground-truth history and is re-noised in the next stage, mirroring the noisy history seen during training. Then, we construct the sequence $\bar{\mathbf{X}}_{t+1}$ augmented with the binary mask $\mathbf{b}_\tau$.

3. *Progressive Denoising.* The denoising network $\mathcal{G}$ performs a fixed number of Δk refinement steps. To mitigate compounding errors and distribution drift during long-horizon rollouts, we apply a stabilization trick by re-injecting a minimal noise level $n > 0$ into the historical latents $\mathbf{x}_{\tau \in [t-h+1, t]}$ before the denoising step

$$\mathbf{x}_\tau^{\max(k-\Delta k, 0)} \leftarrow \mathcal{G}(\text{InjectNoise}(\mathbf{x}_\tau^k, n), \bar{\mathbf{X}}_{t+1}, \mathcal{I}_{\leq t+1}). \tag{8}$$

As we reuse the prediction from the shared period, our progressive refinement reverses only Δk diffusion steps, rather than denoising the entire sequence from scratch. Such a design significantly reduces the number of diffusion steps per progression, enabling real-time estimation of full-body motion. We provide the formal algorithmic procedure in Alg. 1 of the appendix to facilitate a clearer understanding of our inference pipeline.

4 Experiments

In this section, we present a comprehensive experimental evaluation of our method for online egocentric motion reconstruction. We first outline the evaluation setup (Section 4.1). Next, we compare our method with online and offline baselines for egocentric motion reconstruction (Section 4.2), and demonstrate that our approach achieves strong performance under strict online constraints. Finally, we evaluate the robustness of our approach under noisy input conditions (Section 4.3).

4.1 Evaluation Details

Dataset and Implementation Details. Our pipeline is implemented in PyTorch [34] and trained on a single NVIDIA RTX 3090 GPU. We use the EE4D-Motion dataset [35] as our primary benchmark, which is sampled at 10 FPS. During training, we randomly sample sliding windows consisting of $h+1+f$ frames, where h and f denote the history length and future lookahead horizon, respectively. In our default setting, we use $h = 5$ and $f = 19$, resulting in 25-frame windows. We use $K = 100$ diffusion timesteps. Each causal denoising step spans $\Delta k = K/(f+1) = 5$ diffusion steps. For stabilization sampling, we set the noise injection level to $n = 2$. When training with noisy egocentric observations (Section 4.3), we inject zero-mean noise with a standard deviation into the input signals. The noise standard deviation is controlled by a noise level l, defined as $(l^\circ, l \text{ cm})$ for rotational and

Table 1: Quantitative egocentric motion reconstruction results on the EE4D-Motion dataset, comparing online and offline baselines. Our method achieves state-of-the-art performance under strict online causal constraints. Best values are highlighted in blue and second-best values in red. HMD^2 (task-adapted) denotes our re-implementation adapted to the egocentric inputs of this benchmark (Section C), as the official code is not publicly available.

Methods	Reconstruction ↓				Quality		Smoothness		Sampling	
	Head PE	Wrist PE	MPJPE	MPJRE-F	Semantic ↑	FID ↓	PJ →	AUJ ↓	Time (s) ↓	Online
Real motions	-	-	-	-	0.999	0.001	0.329	10.7	-	-
AvatarJLM [54]	-	0.187	0.119	0.388	0.829	0.096	1.743	18.8	0.01	✓
RPM [3]	-	0.108	0.115	0.364	0.907	0.042	0.453	11.1	0.005	✓
HMD^2 (task-adapted) [15]	-	0.131	0.127	0.245	0.913	0.032	0.373	11.0	0.35	✓
EgoAllo [48]	-	0.079	0.168	0.612	0.869	0.045	0.912	14.3	84.9	✗
UniEgoMotion [35]	0.034	0.089	0.091	0.497	0.917	0.024	0.738	12.7	2.05	✗
EgoForce (Ours)	0.072	0.076	0.073	0.138	0.987	0.016	0.318	10.8	0.08	✓
w/o image feat.	0.073	0.078	0.077	0.141	0.986	0.018	0.313	11.0	0.07	✓
w/o stabilize.	0.074	0.081	0.075	0.143	0.985	0.019	0.315	10.9	0.08	✓
$h = 30$	0.138	0.117	0.129	0.271	0.933	0.086	0.298	12.9	0.08	✓
offline ver.	0.036	0.044	0.042	0.059	0.991	0.012	0.322	10.8	3.19	✗

translational signals, respectively, following [21]. We set $l = 2$ and $K^* = 3$. Additional details on the data processing and implementation are provided in Section C of the Appendix.

Baselines. We compare our method against several baselines, including online egocentric motion reconstruction models (AvatarJLM [54] and RPM [3]), an online window-level diffusion baseline (HMD^2 (task-adapted) [15], our re-implementation adapted to the egocentric inputs of this benchmark as detailed in Section C), and recent diffusion-based models (UniEgoMotion [35] and EgoAllo [48]). For the offline methods, we train each model using 80 randomly sampled frames per window. During generation, the reconstructed windows are stitched together to produce the full motion sequence. On the other hand, our online method processes the long-term trajectory progressively. Evaluation metrics are computed over the full motion sequence. More details on baseline implementations are provided in the Appendix.

Evaluation Metrics. We evaluate our method across four dimensions: (i) Reconstruction Accuracy via MPJPE (m) and rotation error (MPJRE-F, Frobenius norm) for the full body, along with Head Position Error (Head PE) and Wrist Position Error (Wrist PE); (ii) Motion Quality using TMR-based [37] semantic similarity and FID to assess realism; (iii) Motion Smoothness via Peak Jerk (PJ) and Area Under the Jerk (AUJ) to measure temporal stability; and (iv) Online Capability by reporting causality and per-frame inference latency. Detailed definitions and configurations are provided in Section D of the Appendix.

4.2 Online Egocentric Motion Reconstruction

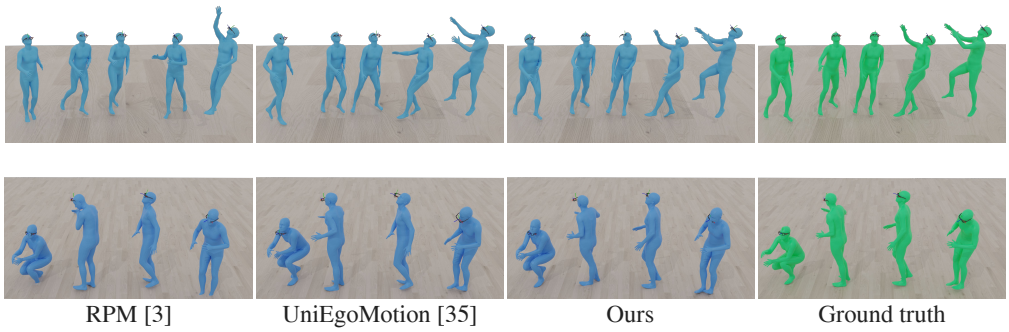

Figure 4: Existing online methods (e.g., RPM [3]) suffer from limited motion fidelity, whereas offline approaches (e.g., UniEgoMotion [35]) rely on window-based generation and stitching, often leading to discontinuous motion at window boundaries. In contrast, our method generates globally coherent and smooth motion under strict causal constraints.

Table 1 presents a quantitative comparison of our approach against various baseline configurations, including strong online and offline baselines. The supplementary video features the reconstructed motion sequences.

Baselines Analysis. Existing online reconstruction methods satisfy strict causality and online constraints, but their limited modeling capacity often leads to inferior reconstruction accuracy and lower overall motion quality. In contrast, offline reconstruction methods achieve higher motion quality and lower reconstruction errors; however, they rely heavily on future control signals and extensive iterative denoising or additional guidance inference. This incurs substantial latency, making them unsuitable for online egocentric reconstruction settings. The causal window-level diffusion baseline HMD^2 (task-adapted) is more accurate than the deterministic online models, but it re-samples every block from Gaussian noise with a uniform diffusion timestep and conditions on previously generated motion only at inference time, which increases the per-update cost and leaves a clear gap to our frame-level progressive refinement.

Reconstruction Accuracy and Motion Quality under Online Constraints. Our method effectively bridges this gap between online capability and motion quality. Under strict online settings, it consistently outperforms existing online baselines by producing more accurate and temporally coherent full-body motions. Specifically, our approach achieves significantly lower FID, higher semantic similarity, and reduced full-body and wrist reconstruction errors. Note that the head trajectory is used as a soft conditioning signal rather than a hard kinematic target, so our online model does not reproduce the observed head trajectory exactly; we discuss the resulting trade-off between head-following accuracy and motion naturalness in Section F. Moreover, our approach remains competitive with offline diffusion-based methods while requiring substantially fewer inference steps and strictly operating without future information. These results demonstrate that our formulation enables both high-quality motion reconstruction and practical online deployment. In Figure 4, our results more faithfully follow the ground-truth motion while strictly adhering to causal online constraints. The motion quality is better visualized in the supplementary video.

Long-Term Stability and Smoothness. Furthermore, our approach exhibits exceptional stability in reconstructing long-term motion sequences. With future prediction conditioned on historical states, which is progressively refined with incoming egocentric observations, our method effectively mitigates error accumulation over time. This continuous refinement leads to highly stable and smooth motion generation. In contrast to online approaches like RPM that attempt to enforce smoothness, or offline methods that typically generate fixed-length temporal windows and stitch them together (often resulting in discontinuous transitions at boundaries), our method ensures seamless and naturally smooth movements. This superior temporal stability is quantitatively validated by our Peak Jerk (PJ) and significantly lowered Area Under Jerk (AUJ) metrics. Figure 1 presents representative long-horizon reconstructions, highlighting consistent global structure and smooth temporal transitions across extended motion sequences.

Ablation Study. We compare our model against several ablated variants to evaluate the impact of our specific design choices. By conditioning on image features, our model reconstructs more semantically similar motions, which is quantitatively verified by improved semantic similarity and motion quality scores. Although the average gain is modest, visual context matters most in two situations: when the wrists are occluded or leave the field of view for extended periods, and in scene-dependent interactions where the activity is dictated by the surrounding environment. For instance, in bicycle-repair sequences the wrists are unobserved for long stretches, and the egocentric images supply the action and object cues that kinematic tracking alone cannot provide. Furthermore, incorporating our stabilization trick improves reconstruction accuracy by reducing error accumulation during the sampling process. We additionally find that conditioning on a longer history context ($h = 30$) accumulates errors and degrades performance; therefore, we limit the history length (h) to 5.

Upper-Bound Performance with Offline Variant. We also evaluate an offline variant of our model, which observes the full conditioning signal $\mathbf{c}_{[t-h:t+f]}$. This offline version provides an upper-bound estimate for our architecture, while our standard online model remains competitive despite strictly satisfying causal online constraints.

4.3 Noise Robust Reconstruction

Next, we evaluate the robustness of our method under noisy egocentric observations in Table 2. Under strict online constraints, our approach steadily generates high-quality motions (e.g., Semantic Score of 0.971) while maintaining low reconstruction error (MPJPE of 0.084m) despite noisy egocentric signals. These results demonstrate that our noise-robust causal diffusion framework, trained with observation corruption, achieves strong robustness while maintaining high reconstruction accuracy, motion quality, and real-time feasibility.

Table 2: Quantitative egocentric motion reconstruction results on the noisy EE4D-Motion dataset.

Methods	Reconstruction ↓				Quality		Smoothness	
	Head PE	Wrist PE	MPJPE	MPJRE-F	Semantic ↑	FID ↓	PJ →	AUJ ↓
Real motions	-	-	-	-	0.999	0.001	0.329	10.7
AvatarJLM [54]	-	0.228	0.171	0.517	0.813	0.118	1.882	20.8
RPM [3]	-	0.147	0.134	0.386	0.899	0.054	0.436	12.1
HMD^2 (task-adapted) [15]	-	0.149	0.155	0.283	0.884	0.046	0.391	11.3
EgoAllo [48]	-	0.091	0.152	0.635	0.846	0.048	0.937	15.3
UniEgoMotion [35]	**0.053**	0.134	0.145	0.523	0.902	0.030	0.874	14.7
EgoForce (Ours)	0.081	**0.086**	**0.084**	**0.163**	**0.971**	**0.021**	**0.323**	**11.0**
w/o noise robust.	0.093	0.101	0.102	0.201	0.943	0.027	0.412	11.3

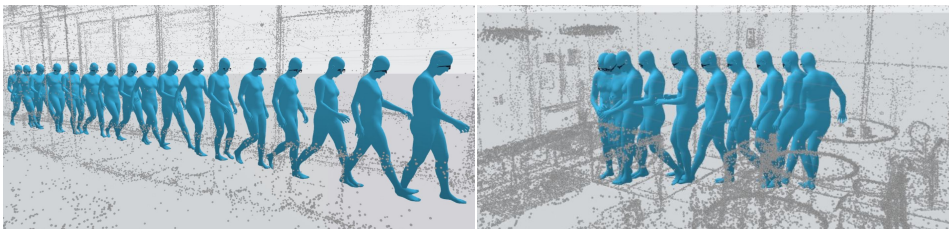

Figure 5: Qualitative Ego-Exo4D examples using Project Aria SLAM trajectories and HaMeR hand estimates as streaming egocentric inputs. Despite real sensing noise and domain shift from training, our framework produces temporally coherent and plausible full-body motion.

To further assess robustness beyond EE4D-Motion, we additionally train our model on the AMASS dataset [30] with synthetic egocentric control signals. We provide additional results on the AMASS dataset in the Appendix (Sections C and E). We then apply the trained model to real-world Ego-Exo4D [12] sequences using actual SLAM trajectories and hand estimation [36] results from streaming egocentric inputs. Since the full-body annotations of Ego-Exo4D are sparse, noisy, and temporally discontinuous, we use it for qualitative evaluation only (see Section C). Our method produces stable and plausible motion reconstructions in these real-world settings, highlighting the framework's effectiveness in practice and its consistent performance across diverse data sources.

5 Conclusion

In this paper, we presented EgoForce, a novel diffusion-based framework for online egocentric motion reconstruction. By formulating egocentric reconstruction as a causal generation problem with temporally evolving uncertainty, we successfully adapted the expressive power of diffusion models to an online streaming setting. Experimental results demonstrate that EgoForce generates high-quality motion from streaming egocentric observations, preserves long-term motion connectivity, and outperforms prior offline diffusion approaches, achieving state-of-the-art performance. We expect EgoForce to promote the development of practical, real-world interactive applications in which continuous and reliable human motion understanding is essential for seamless daily-life assistance and interaction. We discuss remaining limitations, including the trade-off between head-following accuracy and motion naturalness under causal constraints and the reliance on the learned motion prior, in Section F.

Acknowledgments

This work was partially supported by the BK21 FOUR Program (Education and Research Program for Future ICT Pioneers) at Seoul National University in 2026, the Creative-Pioneering Researchers Program through Seoul National University, and an Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. RS-2021-II211343, Artificial Intelligence Graduate School Program, Seoul National University).

References

[1] Yutong Bai, Danny Tran, Amir Bar, Yann LeCun, Trevor Darrell, and Jitendra Malik. Whole-body conditioned egocentric video prediction, 2025.

[2] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. *CVPR*, 2025.

[3] German Barquero, Nadine Bertsch, Manojkumar Marramreddy, Carlos Chacón, Filippo Arcadu, Ferran Rigual, Nicky 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.

[4] German Barquero, Sergio Escalera, and Cristina Palmero. Seamless human motion composition with blended positional encodings. 2024.

[5] Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion, 2024.

[6] Rui Chen, Mingyi Shi, Shaoli Huang, Ping Tan, Taku Komura, and Xuelin Chen. Taming diffusion probabilistic models for character control. In *ACM SIGGRAPH 2024 Conference Papers*, SIGGRAPH '24, New York, NY, USA, 2024. Association for Computing Machinery.

[7] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[8] Kyungwon Cho and Hanbyul Joo. Hand-aware egocentric motion reconstruction with sequence-level context. *arXiv preprint arXiv:2512.19283*, 2025.

[9] Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. In *ECCV*, pages 390–408, 2025.

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision. *International Journal of Computer Vision*, 130(1):33–55, 2022.

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

[12] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, and et al. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives, 2024.

[13] Chuan Guo, Inwoo Hwang, Jian Wang, and Bing Zhou. Snapmogen: Human motion generation from expressive texts, 2025.

[14] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. 2023.

[15] Vladimir Guzov, Yifeng Jiang, Fangzhou Hong, Gerard Pons-Moll, Richard Newcombe, C. Karen Liu, Yuting Ye, and Lingni Ma. Hmd2: Environment-aware motion generation from single egocentric head-mounted device. In *International Conference on 3D Vision (3DV)*, March 2025.

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

[17] Fangzhou Hong, Vladimir Guzov, Hyo Jin Kim, Yuting Ye, Richard Newcombe, Ziwei Liu, and Lingni Ma. Egolm: Multi-modal language model of egocentric motions. *arXiv preprint arXiv:2409.18127*, 2024.

[18] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

[19] Inwoo Hwang, Jinseok Bae, Donggeun Lim, and Young Min Kim. Goal-driven human motion synthesis in diverse task. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, pages 2920–2930, June 2025.

[20] Inwoo Hwang, Jinseok Bae, Donggeun Lim, and Young Min Kim. Motion synthesis with sparse and flexible keyjoint control, 2025.

[21] Inwoo Hwang, Bing Zhou, Young Min Kim, Jian Wang, and Chuan Guo. Scenemi: Motion in-betweening for modeling human-scene interactions, 2025.

[22] Sigmund H. Høeg, Yilun Du, and Olav Egeland. Streaming diffusion policy: Fast policy synthesis with variable noise diffusion models, 2024.

[23] 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*. Springer, 2024.

[24] 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 *Proceedings of European Conference on Computer Vision*. Springer, 2022.

[25] Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Suwajanakorn, and Siyu Tang. Optimizing diffusion noise can serve as universal motion priors. In *arxiv:2312.11994*, 2023.

[26] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2151–2162, 2023.

[27] Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Egohumans: An egocentric 3d multi-human benchmark, 2023.

[28] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17142–17151, 2023.

[29] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, Kevin Bailey, David Soriano Fosas, C. Karen Liu, Ziwei Liu, Jakob Engel, Renzo De Nardi, and Richard Newcombe. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *the 18th European Conference on Computer Vision (ECCV)*, 2024.

[30] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, October 2019.

[31] Vongani H. Maluleke, Kie Horiuchi, Lea Wilken, Evonne Ng, Jitendra Malik, and Angjoo Kanazawa. Diffusion forcing for multi-agent interaction sequence modeling, 2025.

[32] Zichong Meng, Zeyu Han, Xiaogang Peng, Yiming Xie, and Huaizu Jiang. Absolute coordinates make motion generation easy. *arXiv preprint arXiv:2505.19377*, 2025.

[33] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[35] Chaitanya Patel, Hiroki Nakamura, Yuta Kyuragi, Kazuki Kozuka, Juan Carlos Niebles, and Ehsan Adeli. Uniegomotion: A unified model for egocentric motion reconstruction, forecasting, and generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10318–10329, 2025.

[36] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024.

[37] Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *International Conference on Computer Vision (ICCV)*, 2023.

[38] Ekkasit Pinyoanuntapong, Muhammad Saleem, Korrawe Karunratanakul, Pu Wang, Hongfei Xue, Chen Chen, Chuan Guo, Junli Cao, Jian Ren, and Sergey Tulyakov. Maskcontrol: Spatio-temporal control for masked motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9955–9965, 2025.

[39] David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. Rolling diffusion models, 2024.

[40] Yi Shi, Jingbo Wang, Xuekun Jiang, Bingkun Lin, Bo Dai, and Xue Bin Peng. Interactive character control with auto-regressive motion diffusion models. *ACM Trans. Graph.*, 43, jul 2024.

[41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, October 2020.

[42] Kewei Sui, Anindita Ghosh, Inwoo Hwang, Jian Wang, and Chuan Guo. A survey on human interaction motion generation, 2025.

[43] Mingzhen Sun, Weining Wang, Gen Li, Jiawei Liu, Jiahui Sun, Wanquan Feng, shanshan Lao, SiYu Zhou, Qian He, and Jing Liu. Ar-diffusion: Asynchronous video generation with auto-regressive diffusion. 2025.

[44] Takara Everest Truong, Michael Piseno, Zhaoming Xie, and Karen Liu. Pdp: Physics-based character animation via diffusion policy. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–10, 2024.

[45] Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. Tlcontrol: Trajectory and language control for human motion synthesis. *arXiv preprint arXiv:2311.17135*, 2023.

[46] Yan Wu, Korrawe Karunratanakul, Zhengyi Luo, and Siyu Tang. Uniphys: Unified planner and controller with diffusion for flexible physics-based character control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[47] Lixing Xiao, Shunlin Lu, Huaijin Pi, Ke Fan, Liang Pan, Yueer Zhou, Ziyong Feng, Xiaowei Zhou, Sida Peng, and Jingbo Wang. Motionstreamer: Streaming motion generation via diffusion-based autoregressive model in causal latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10086–10096, October 2025.

[48] Brent Yi, Vickie Ye, Maya Zheng, Yunqi Li, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7072–7084, 2025.

[49] Qing Yu, Akihisa Watanabe, and Kent Fujiwara. Causal motion diffusion models for autoregressive motion generation. In *CVPR*, 2026.

[50] Siwei Zhang, Bharat Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, Petr Kadlecek, Siyu Tang, and Federica Bogo. Rohm: Robust human motion reconstruction via diffusion. In *CVPR*, 2024.

[51] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European Conference on Computer Vision*, pages 180–200. Springer, 2022.

[52] Zihan Zhang, Richard Liu, Kfir Aberman, and Rana Hanocka. Tedi: Temporally-entangled diffusion for long-term motion synthesis. In *SIGGRAPH, Technical Papers*, 2024.

[53] Kaifeng Zhao, Gen Li, and Siyu Tang. DartControl: A diffusion-based autoregressive motion model for real-time text-driven motion control. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[54] 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 (ICCV)*, 2023.

In the appendix, we present comprehensive technical details and extended experimental results. First, we provide the algorithmic pseudocode for our causal online inference pipeline (Section A) and a detailed description of the offline variant of our framework (Section B). Second, we offer further details regarding the experimental setup, including dataset configurations and implementation specifics (Section C), as well as the formal definitions of our evaluation metrics (Section D). Third, we report additional quantitative results on the AMASS dataset, along with an expanded set of ablation studies, to further validate our design choices (Section E). Furthermore, we discuss the limitations of our approach and potential future directions (Section F). Finally, we invite readers to visit our project page for extensive qualitative demonstrations and visual comparisons (Section G).

A Algorithmic Details

In this section, we provide a formal algorithmic description of our proposed inference pipeline to supplement the discussion in the main manuscript. Alg. 1 details the step-by-step procedure for causal online inference, illustrating how the system manages the temporal buffer and the progressive denoising refinement that enables online motion reconstruction.

Algorithm 1 Causal Online Inference with Progressive Refinement

Require: Denoiser $\mathcal{G}$, history h, horizon f, step $\Delta k = K/(f+1)$
Input: Stream of egocentric observations $(\mathbf{c}_t, \mathbf{I}_t)$
Initialize $\mathbf{X}_0 = \{\mathbf{x}_{-h}, \dots, \mathbf{x}_f\} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
Fully denoise $\mathbf{x}_{\tau \in [-h,0]}$ and partially denoise $\mathbf{x}_{i \in [1,f]}$ to $k_i = i \cdot \Delta k$. ▷ Initialization
for each time step $t = 0, 1, 2, \dots$ **do**
    ▷ Temporal Shift and Warm-start
    $\mathbf{X}_{t+1} \leftarrow \mathbf{X}_t \setminus \{\mathbf{x}_{t-h}\}$
    Shift buffer left: $\mathbf{x}_\tau^{(t+1)} \leftarrow \mathbf{x}_\tau^{(t)}$
    Append $\mathbf{x}_{t+f+1}^K \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    ▷ Causal Condition Injection
    Receive $(\mathbf{c}_{t+1}, \mathbf{I}_{t+1})$ and update $\mathbf{b}_\tau$
    Inject causal signals via Eq. 4
    Construct augmented sequence $\bar{\mathbf{X}}_{t+1}$
    ▷ Progressive Denoising by Δk steps
    $\mathbf{x}_\tau^{\max(k-\Delta k,0)} \leftarrow \mathcal{G}(\text{InjectNoise}(\mathbf{x}_\tau^k, n), \bar{\mathbf{X}}_{t+1}, \mathcal{I}_{\leq t+1})$
    Output: Clean current pose $\hat{\mathbf{x}}_{t+1}$ (where $k = 0$)
end for

B Offline Version Variant

B.1 Training

Unlike our default online setting, the offline variant assumes access to the full conditioning signal $\mathbf{c}_{[t-h:t+f]}$ within a temporal window of length T. Instead of directly generating the high-dimensional full-body motion $\mathbf{x}_\tau \in \mathbb{R}^d$, the offline model first generates movements in a low-dimensional key-joint space $\mathbf{z}_\tau \in \mathbb{R}^{d_J}$ [20], which is subsequently used to reconstruct the full-body motion $\mathbf{x}$.

In this offline setting, we remove the causal asymmetry described in Sec. 3.2 and apply the visibility mask directly to the key-joint representation. Similar to the online formulation, the mask $\mathbf{b}_\tau \in \{0,1\}^{d_J}$ is derived from the egocentric visibility signal $\mathbf{b}_\tau^{\text{vis}}$, applied over the key-joint space $\mathbf{z}$ at each time τ:

$$\mathbf{b}_\tau = \mathbf{b}_t^{\text{vis}}$$

Similar to the online setting, we assign independent, varying per-frame noise levels. Each ground-truth key-joint pose $\mathbf{z}_\tau^0$ is then independently perturbed according to its specifically assigned noise level k_τ, yielding the noisy key-joint representation $\mathbf{z}_\tau^{k_\tau}$. Then, we perform the masked observation injection and concatenate the resulting representation with its visibility mask as follows:

$$\tilde{\mathbf{z}}_\tau = \mathbf{b}_\tau \odot \mathbf{z}_\tau^0 + (1 - \mathbf{b}_\tau) \odot \mathbf{z}_\tau^{k_\tau}, \quad \bar{\mathbf{z}}_\tau = [\tilde{\mathbf{z}}_\tau, \mathbf{b}_\tau]. \tag{9}$$

By constructing the augmented motion window $\bar{\mathbf{Z}}_t = \{\bar{\mathbf{z}}_\tau\}$, we train the key-joint diffusion model $\mathcal{G}_z$ with the following objective:

$$\mathcal{L}_{\mathbf{z}} = \mathbb{E}_{\mathbf{Z}_t^0, \mathbf{k}} \left[\left\| \mathcal{G}_z(\bar{\mathbf{Z}}_t, \mathbf{k}, \mathbf{I}_{\leq t}) - \mathbf{Z}_t^0 \right\|_2^2 \right]. \tag{10}$$

Following the generation of key-joint trajectories via $\mathcal{G}_z$, we train a subsequent full-body completion model $\mathcal{G}_x$ to reconstruct the full-body motion $\mathbf{x} \in \mathbb{R}^d$. In this stage, the generated key-joint values $\mathbf{z}_\tau$ are anchored into their corresponding indices within the full-body vector $\mathbf{x}_\tau$:

$$\mathbf{x}_\tau^{\texttt{keyjoints}} = \mathbf{z}_\tau. \tag{11}$$

The full-body completion model is trained via a full-sequence diffusion paradigm. We apply a consistent noise level k across the entire temporal window rather than varying it per frame. This formulation encourages the model to capture global spatial correlations between the anchored key joints and the remaining degrees of freedom. Let $\bar{\mathbf{X}}_t$ denote the augmented motion window; the full-body completion model $\mathcal{G}_x$ is trained using the following objective:

$$\mathcal{L}_{\mathbf{x}} = \mathbb{E}_{\mathbf{X}_t^0, k} \left[\left\| \mathcal{G}_x(\bar{\mathbf{X}}_t, k, \mathbf{I}_{\leq t}) - \mathbf{X}_t^0 \right\|_2^2 \right]. \tag{12}$$

B.2 Inference

To achieve stable, long-term motion synthesis beyond the fixed training window T, we employ a chunk-wise autoregressive inference strategy integrated with causal uncertainty sampling. This approach ensures temporal coherence and prevents error accumulation by allowing the model to refine transitional regions between consecutive temporal windows.

Instead of a standard sliding window with hard boundaries, we introduce a progressive uncertainty schedule within each chunk of length T. For a given chunk, the initial $T - L$ frames are treated as clean observations (noise level $k = 0$), providing a stable anchor for the sequence. For the remaining L frames—defined as the uncertainty region—we apply a monotonically increasing noise schedule. Specifically, the noise level for the last i-th frame within this region ($i \in \{1, \ldots, L\}$) is defined as:

$$k_{T-L+i} = i \cdot \Delta k, \quad \text{where } \Delta k = \frac{K}{L+1}. \tag{13}$$

This design encourages the model to transition from deterministic reconstruction in the confident region to a probabilistic, generative state toward the end of the chunk, effectively mimicking the signal distribution the model encounters during the next window's transition.

The generation proceeds by shifting the temporal window forward. To maintain continuity, the subsequent chunk reuses the final $M + L$ frames from the preceding window as temporal context. Here, the L frames represent the previously generated uncertainty region, which the model now "denoises" into a clean state, while the M frames provide additional stable, clean history. These context frames are concatenated with the fresh egocentric conditioning signals for the remaining $T - (M + L)$ frames of the new window.

By iteratively refining these overlapping uncertainty regions, we produce a seamless and stable long-term trajectory in the key-joint space $\mathbf{z}$. Once the full key-joint trajectory $\mathbf{z}$ is generated, the full-body completion model $\mathcal{G}_x$ is applied to lift the low-dimensional trajectory into the final full-body motion $\mathbf{x}$.

C Implementation Details

C.1 Model Details

We model online egocentric motion reconstruction with a Diffusion Forcing denoiser. The denoiser is a 12-layer Transformer decoder with latent dimension 768, 12 attention heads, pre-norm self-attention, cross-attention to visual DINOv2 features, and SwiGLU feed-forward blocks. The per-frame diffusion level is embedded with a sinusoidal/MLP timestep embedding of dimension 64 and concatenated to the motion features before the input projection.

We train all models using AdamW with a learning rate of 3×10^{-5}, weight decay of 0.01, and default momentum parameters $(\beta_1, \beta_2) = (0.9, 0.999)$. The batch size is set to 64. We train the models for 500 epochs. Gradients are clipped with a global norm of 1.0, and we maintain an exponential moving average of model parameters with decay 0.999. For diffusion training, we use 100 diffusion steps with a cosine noise schedule and optimize the x_0 prediction objective using an MSE loss. Models are trained with PyTorch Lightning on a single NVIDIA RTX 3090 GPU.

C.2 Data Processing

Following prior egocentric motion works [48, 54], we construct proxy head-mounted device (HMD) and wrist trajectories as sparse egocentric control signals for both EE4D-Motion [35] and AMASS [30]. For EE4D-Motion, the proxy trajectories are obtained from the body motion annotations provided with the dataset. For AMASS, we construct the same egocentric control signals from motion-capture sequences. All trajectories are sampled at 10 FPS. Following the technical specifications of Aria glasses, we apply a Field-of-View (FoV) filter to the wrist joints; only joints within the operational FoV are treated as observed, while those outside the sensor range are masked. For noisy-input experiments, we independently perturb both HMD and wrist trajectories with predefined rotational and translational noise distributions as described in Section 4.1.

For real-world experiments on Ego-Exo4D [12], we use egocentric inputs captured by Project Aria glasses. Specifically, we use SLAM camera poses as head trajectories and estimate wrist poses from egocentric images using HaMeR [36], following the protocol of EgoAllo [48]. Although Ego-Exo4D provides 3D body joint annotations for a subset of sequences, these annotations are sparse, noisy, temporally discontinuous, and do not include joint-angle information [35]. Therefore, we use Ego-Exo4D only for qualitative visualization rather than quantitative motion reconstruction evaluation.

C.3 Baseline Details

The baseline methods RPM [3], AvatarJLM [54], HMD^2 [15], and EgoAllo [48] take the head trajectory as input and predict local body motion; their global body translation is obtained by anchoring the predicted local motion to the provided head trajectory. Accordingly, we omit head-tracking metrics for these methods.

We adapt wrist conditioning according to each baseline's original design. For EgoAllo [48], we follow the original inference procedure and use wrist observations in the post-optimization stage through diffusion guidance. For AvatarJLM and RPM, we condition on wrist observations following their original conditioning formulations. For UniEgoMotion [35] and HMD^2, we additionally provide wrist observations as conditioning signals to match our egocentric input setting. We refer to our HMD^2 baseline as HMD^2 (task-adapted) throughout the paper, since it is a task-matched adaptation rather than a reproduction of its native setting. EE4D-Motion does not provide the SLAM point-cloud modality used by the original HMD^2; we therefore remove the point-cloud encoder, use DINOv2 [33] image features as the visual encoder, and condition the model on the same egocentric signals as all other evaluated methods (HMD pose, images, and wrist observations). The adaptation preserves the core algorithmic design of HMD^2, namely window-level causal diffusion with a uniform diffusion timestep, autoregressive overlap inpainting, and complete per-block denoising, so that the comparison isolates window-level block diffusion from our frame-level progressive refinement. Because the official HMD^2 codebase is not publicly available, we re-implement the method as described in the paper.

D Evaluation Metrics

We evaluate our method from four complementary perspectives: (i) reconstruction accuracy, (ii) motion quality, (iii) long-term stability (motion smoothness), and (iv) online capability.

Reconstruction Accuracy. We evaluate reconstruction accuracy by measuring how closely the predicted motion aligns with the ground-truth motion. For each test scenario, we reconstruct the full motion sequence (average length: 173 frames) and compute errors across the entire sequence. We report both global translation and rotation errors for all body joints and the head. Translational errors

are quantified using the mean per-joint position error (MPJPE, meters), while rotational errors are measured via the mean per-joint rotation error (MPJRE-F), calculated as the Frobenius norm between predicted and ground-truth rotation matrices. Head PE and Wrist PE correspond to the translational errors of the head and wrist, respectively.

Motion Quality. Motion quality evaluates the realism and distributional consistency of generated motion. We embed both generated and ground-truth motions into a shared latent space using a pretrained motion encoder from TMR [37]. Semantic similarity measures how well the generated motion matches the ground-truth motion at a semantic level by computing the cosine distance between their latent embeddings. In addition, *Fréchet Inception Distance* (FID) captures the distributional discrepancy between generated and real motions in the latent space. Extensive motion sequences can hinder the extraction of meaningful latent representations; therefore, we randomly sample segments of 80 frames from the reconstructed motions and evaluate motion quality on these segments.

Motion Smoothness. Motion smoothness evaluates the temporal dynamics of long-term reconstructed motion, computed over the entire reconstructed sequence beyond what the latent distribution (FID) can capture. To explicitly quantify such effects, we adopt Peak Jerk (PJ) to characterize temporal fluctuations and Area Under the Jerk (AUJ) from [4] to measure cumulative deviations from natural motion statistics. For PJ, performance is optimal when values are close to those of the ground-truth motion, as excessively low values may indicate over-smoothed dynamics. In contrast, lower AUJ values reflect smoother and more natural behavior.

Online Capability. Online capability indicates whether a method supports causal, streaming inference using only past and current observations. We report this as a binary attribute in the evaluation table, along with the per-frame update time for online methods and the batch inference time for offline methods.

E Additional Results

E.1 Results on AMASS dataset

Following Section 4.3, we provide additional quantitative evaluations on the AMASS dataset. We evaluate our framework against representative baselines from both the causal online [3] and offline [48] regimes. Since AMASS does not provide egocentric images, we remove image-feature conditioning. The results demonstrate that our framework generalizes effectively across diverse motion datasets while maintaining high-fidelity reconstruction quality under strict causal online constraints.

Table 3: Quantitative egocentric motion reconstruction results on the AMASS dataset, comparing online and offline baselines. Our method shows generalizability across diverse datasets.

Methods	Reconstruction ↓				Quality
	Head PE	Wrist PE	MPJPE	MPJRE-F	FID ↓
Real motions	-	-	-	-	0.001
RPM [3]	-	0.342	0.339	0.638	0.821
EgoAllo [48]	-	0.172	0.134	0.532	0.455
EgoForce (Ours)	0.136	**0.146**	**0.129**	**0.238**	**0.423**

E.2 Evaluation across Multiple Seeds

Since our motion synthesis framework is generative, we evaluate the stochastic stability of our method by repeating the experiments from Table 1 and Table 2 across 20 independent random seeds. As reported in Table 4, the mean performance is presented alongside 95% confidence intervals. The observed variance across all runs is negligible, which demonstrates that our diffusion-based approach consistently produces high-quality reconstructions.

Table 4: Evaluation across multiple random seeds. We report the mean performance and 95% confidence intervals for key metrics over 20 runs.

Methods	Reconstruction ↓				Quality		Smoothness	
	Head PE	Wrist PE	MPJPE	MPJRE	Semantic	FID	PJ	AUJ
Tab. 1 (Ours)	0.072	0.076	0.073	0.138	0.987	0.016	0.318	10.8
+ 20 runs	$0.073^{\pm 0.001}$	$0.077^{\pm 0.002}$	$0.073^{\pm 0.002}$	$0.137^{\pm 0.002}$	$0.987^{\pm 0.001}$	$0.016^{\pm 0.001}$	$0.317^{\pm 0.002}$	$10.8^{\pm 0.003}$
Tab. 2 (Ours)	0.081	0.086	0.084	0.163	0.971	0.021	0.323	11.0
+ 20 runs	$0.080^{\pm 0.002}$	$0.085^{\pm 0.003}$	$0.084^{\pm 0.002}$	$0.163^{\pm 0.002}$	$0.970^{\pm 0.001}$	$0.021^{\pm 0.001}$	$0.322^{\pm 0.003}$	$11.0^{\pm 0.003}$

E.3 Results without Wrist-Observation Conditioning

We further investigate the impact of wrist-signal conditioning by evaluating a variant of our model that excludes egocentric hand observations, using the same experimental setup as Table 1 in the main paper. As shown in Table 5, the performance degradation is marginal. This suggests that the egocentric image features already capture contextual cues regarding hand coordination and proximity [35], allowing the model to reasonably infer hand motion even without direct skeletal input.

Table 5: Ablation study on wrist-signal conditioning.

Methods	Reconstruction ↓				Quality		Smoothness	
	Head PE	Wrist PE	MPJPE	MPJRE-F	Semantic ↑	FID ↓	PJ →	AUJ ↓
EgoForce (Ours)	0.072	**0.076**	**0.073**	**0.138**	**0.987**	**0.016**	**0.318**	**10.8**
w/o wrist signal	**0.064**	0.087	0.085	0.143	0.981	0.018	0.306	10.7

F Limitations

Our framework can be further improved in several aspects. (i) Although our method supports online causal generation, it currently requires $\Delta k = 5$ denoising steps per update. While this is sufficient for 10 FPS motion streams on [35], reducing the number of diffusion steps [18] would further improve scalability to higher frame rates and real-time applications. (ii) Head-following versus motion naturalness. The HMD trajectory is used as a soft conditioning signal rather than a hard kinematic target, so the reconstructed head does not exactly coincide with the observed trajectory, and our online model shows a higher Head PE than offline methods even though it leads on the remaining metrics. Our offline variant closes most of this gap (Table 1), indicating that the residual error mainly stems from the causal constraint. Hard anchoring of the head to the observed trajectory removes this error, but in our experiments it degraded full-body realism. The residual gap thus reflects a trade-off between strict condition following under causal constraints and motion naturalness, and tighter cross-modal alignment that preserves both remains an open problem. (iii) Our framework requires external hand poses, which can be obtained from egocentric images for real-world use, similar to prior work [48]. Future work could leverage more advanced image or video encoders designed for egocentric motion understanding [1], potentially enabling more robust and integrated perception-to-motion modeling. (iv) Failure cases. Our reconstruction can occasionally miss subtle motion details and may produce head-pose errors during large global movements. We provide representative failure cases and qualitative examples on the project page. (v) Reliance on the learned motion prior. In our rolling buffer, each frame receives its current observation only during the last Δk denoising steps before it reaches the current time step, so the reconstruction is partly driven by the motion prior learned from the training data. Allocating more denoising steps per update strengthens condition following but shortens the future horizon and increases latency. Consequently, how well the method generalizes to motions that are rare or absent in the training data remains an open question that we have not evaluated quantitatively, and we leave a systematic study on out-of-distribution motions to future work.

G Project Page

We provide a project page (`https://inwoohwang.me/EgoForce`) with supplementary qualitative results. The project page includes videos of long-horizon online reconstructions, comparisons with

baseline methods, and additional examples under noisy egocentric inputs. We also include real-world Ego-Exo4D examples.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract and introduction clearly outline the challenges of online egocentric motion reconstruction and our proposed diffusion-based framework.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We discuss the limitations in Section F of the appendix, which is referenced from the conclusion, and provide failure cases on the project page.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This work does not involve theoretical innovations.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide implementation details in the Section C of appendix. We also include baseline implementation details.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper provides implementation and data-processing details in the appendix, but the current submission does not include an anonymized code release or pretrained checkpoints. We plan to release code and models upon publication, subject to the terms of the underlying datasets and third-party assets.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We discussed the architecture of the models and training details in the Section C of the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the mean values along with 95% confidence intervals calculated from 20 independent runs to ensure statistical significance in Table 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide details in Section C of the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and ensured our submission to conform the regulations.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is not much societal impact of this work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: Our work does not have such risk.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have reviewed the license of the used datasets and models, and carefully follow their license terms. All of them permit research use.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: We provide a project page, which includes additional results and details. If code, checkpoints, or processed annotations are released later, they will be documented with usage instructions, licenses, and limitations.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The authors do not conduct new crowdsourcing experiments or collect new human-subject data. The experiments use existing datasets cited in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our work does not study particular human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were only utilized for grammar correction and language polishing.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.