ReViV: Reconstructing the Viewer and the View in 4D from Monocular Egocentric Video

Xiaozhong Lyu[1⋆], Gen Li[1⋆], Zhiyin Qian[1], Xucong Zhang[1,2], Marc Pollefeys[1,3], and Siyu Tang[1]

[1] ETH Zurich, Switzerland
[2] Delft University of Technology, Netherlands
[3] Microsoft, Switzerland

Abstract. Egocentric devices, such as wearable front-facing cameras, provide a unique perspective for capturing the continuous interaction between a human viewer and the surrounding environment. A holistic and efficient multimodal model capable of reconstructing this 4D representation is therefore highly desirable. However, existing approaches often rely on auxiliary inputs such as pre-computed camera trajectories, treat scene perception and human ego-motion modeling as separate problems despite their strong interdependency, and suffer from slow inference time. To address these limitations, we present ReViV, the first unified framework for holistic egocentric 4D reconstruction that extracts both viewer and view dynamics from a single monocular RGB video. We formulate the task as learning the full joint probability distribution over multimodal signals, including RGB video, camera trajectory, gaze direction, full-body motion, hand motion, and depth. Powered by a Masked Generative Egocentric Transformer, ReViV operates within a single feed-forward architecture to simultaneously reconstruct the temporally consistent 4D reconstruction across the viewer and the view with fast inference speed. Extensive experiments on diverse benchmarks, including HoloAssist, HOT3D, ARCTIC, Aria Digital Twin, and TACO, demonstrate that ReViV achieves state-of-the-art accuracy and efficiency across holistic ego-body, hand, and gaze reconstruction, camera tracking, while maintaining highly competitive egocentric depth estimation, without relying on heavy task-specific priors. Code and models are fully open-sourced: `https://reviv4d.github.io/`.

Keywords: Egocentric Vision · 4D Reconstruction · Human Motion

1 Introduction

Egocentric vision lies at the core of embodied intelligence, as it captures the world directly from the first-person perspective through which humans perceive, act, and interact [16, 42]. With the rapid proliferation of wearable devices such as smart glasses, body-mounted cameras, and head-mounted displays, egocentric

⋆ Equal contribution; order interchangeable on CVs. See Acknowledgements.

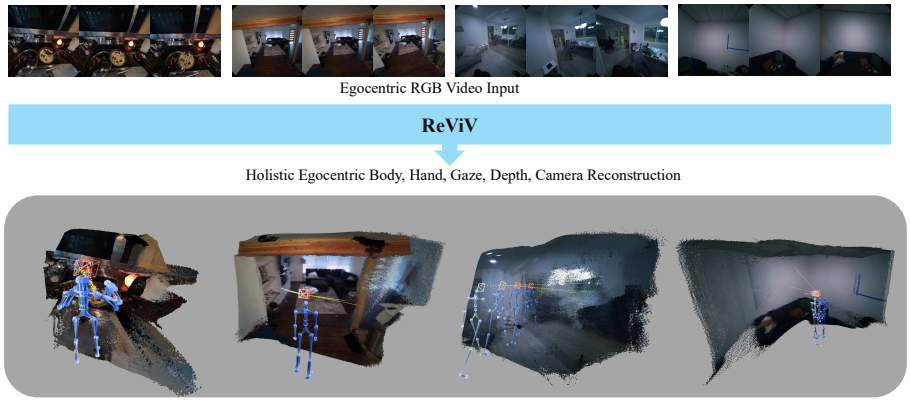

Fig. 1: We present **ReViV**, a unified framework for holistic egocentric 4D reconstruction. Given a monocular egocentric RGB video, our method jointly estimates human-centric modalities, including **body**, **hand** pose, and **gaze**, alongside scene-aware **camera trajectory** and **depth estimation**. These predictions are integrated into a temporally consistent viewer–view reconstruction from a monocular egocentric video.

video has become increasingly prevalent in augmented and virtual reality [44], assistive robotics [53], and human–computer interaction. In these scenarios, systems must not only understand the surrounding environment, but also interpret the wearer's own actions, gestures, and intentions in real time [14, 18, 33].

Despite the unique perspective offered by egocentric vision, reconstructing the viewer's actions and scene remains insufficiently explored [8,13]. Existing egocentric methods typically focus on either environmental reconstruction [22,60,63] or human pose estimation [15, 23, 24, 31, 36, 57, 61]. As a result, scene dynamics and human motion are inferred independently, ignoring the temporal constraints that naturally couple the observer and the observed world. Moreover, many of the egocentric human pose estimation approaches depend on auxiliary inputs, such as on-device SLAM trajectories [36, 57], precomputed 3D point clouds [57], or dedicated hand-tracking modules [57], limiting their applicability to specialized hardware setups. By treating scene understanding and human reconstruction in isolation, current approaches often result in temporally inconsistent view geometry and the viewer's motion [10, 56]. Therefore, a holistic framework that reconstructs the viewer's human motion while jointly modeling the contextual 4D reconstructions, such as depth sequences, gaze, and camera trajectories, from monocular egocentric video remains an open and challenging problem.

Recent advances in 4D reconstruction have achieved impressive results in third-person or multi-view settings, reconstructing dynamic scenes and human motion from multiple synchronized cameras [10, 56]. Methods such as Shape of Motion [47] enable long-range dynamic reconstruction by representing scenes with persistent 3D Gaussians and motion bases, and feed-forward architectures further improve efficiency and generalization by avoiding costly test-time optimization [6, 12, 52, 55]. However, these advances rely on third-person viewpoints

or multi-view capture, leaving the problem of reconstructing the wearer's own body dynamics in monocular egocentric settings largely unresolved.

To address this challenge, we present ReViV (see Fig. 1), a unified framework for **Re**constructing the **Vi**ewer and the **V**iew from a single monocular egocentric RGB video. Rather than treating body motion, hand dynamics, gaze, camera trajectory, and depth as independent prediction tasks, ReViV formulates egocentric reconstruction as masked generative modeling over multimodal signals. At inference time, ReViV conditions on RGB observations and predicts the missing viewer-centric and scene-centric modalities in a single feed-forward pass. Since monocular depth is inherently ambiguous in scale, ReViV predicts temporally consistent affine-invariant scene depth and uses a lightweight alignment step to place the reconstructed viewer and view into a metric 4D coordinate system when suitable geometric anchors are available.

At the core of our framework is a Masked Generative Egocentric Transformer (MGET). To model the complex and partially unobserved kinematics of the viewer, we introduce a unified tokenization scheme that encodes spatiotemporal signals from diverse modalities into compact latent representations. MGET is trained with a unified multi-task objective that learns intrinsic human motion dynamics and cross-modal correlations by predicting randomly masked tokens conditioned on available context. By masking subsets of multimodal tokens during training, the model learns to infer missing kinematics and scene geometry from visible cues, capturing the underlying joint distribution across modalities.

Extensive experiments on multiple benchmarks demonstrate that ReViV achieves state-of-the-art performance in holistic ego-body, hand, and gaze reconstruction while maintaining competitive accuracy in depth estimation and camera tracking. By scaling our multimodal dataset to 7B unique tokens and optimizing over 500B training tokens, the model learns robust cross-modal dependencies that yield temporally consistent viewer–view dynamics. ReViV's feed-forward architecture enables notably faster inference than optimization-based methods [17, 60]. Our contributions are:

- A unified framework for egocentric 4D reconstruction from a single monocular RGB video, holistically modeling scene, body, hand, and gaze.
- A large-scale feed-forward pretrained model that explicitly models human motion, improving scalability and generalization in egocentric settings.
- Comprehensive benchmarking on downstream egocentric tasks, demonstrating that ReViV achieves state-of-the-art accuracy in human body, hand, and gaze reconstruction along with highly competitive depth estimation.

2 Related Work

2.1 4D Reconstruction

Dynamic 4D reconstruction recovers temporally coherent 3D structures of dynamic scenes over time. Early works mainly focused on capturing 3D motion trajectories of objects [29, 48]. These approaches have evolved from static 3D scene

reconstruction toward dynamic scene modeling using synchronized multi-view video inputs [10,30,51,56]. Recent advances further relax the multi-view requirement by performing 4D reconstruction from monocular videos [7, 21, 26, 43, 49].

To model long-term dynamics, [47] represents dynamic scenes as persistent 3D Gaussians with compact motion bases, achieving efficient long-range reconstruction. However, Gaussian-based representations still rely on per-scene test-time optimization, limiting their generalization. To address this, recent feed-forward methods jointly reconstruct and track dynamic scenes, offering fast and generalized solutions without optimization at inference time [12, 52, 55]. Despite these advances, existing 4D reconstruction methods mainly target third-person or multi-view settings, leaving egocentric domain largely unexplored.

2.2 Scene Reconstruction from Egocentric Videos

Scene reconstruction from egocentric perspectives poses unique challenges due to the limited field of view, frequent camera motion, and strong occlusions. EgoM2P [22] introduces a multitask pretraining framework for egocentric vision that jointly models RGB, depth, gaze, and camera trajectories, enabling multimodal scene understanding. EgoGaussian [63] extends this idea by simultaneously reconstructing 3D scenes and dynamically tracking object motion directly from monocular egocentric RGB inputs. EgoMono4D [60] applies self-supervised learning for point cloud sequence reconstruction to the label-scarce egocentric field. These works demonstrate the potential of egocentric video for holistic scene modeling, yet they focus mainly on scene or object-level geometry, without explicitly reconstructing full human motion and interaction dynamics.

2.3 Human Motion Estimation from Egocentric Videos

Estimating human motion from egocentric videos has been explored in several directions. Some works employ fisheye cameras to predict full-body pose [9,19,25, 45, 46, 54], though image distortion and self-occlusion remain major challenges. EgoEgo [24] estimates 3D human motion by first predicting head movement from the egocentric video and then inferring the full-body motion from it, however, it does not recover hand poses. Ego-Pose [61] estimates and forecasts a person's pose sequence from egocentric input using reinforcement learning, while [31] introduces an object-aware 3D egocentric pose estimation framework. More recently, generative models have significantly advanced egocentric pose estimation. HMD^2 [15] and EgoAllo [57] employ conditional diffusion models to generate full-body kinematics; additionally, EgoAllo estimates hand movements by integrating off-the-shelf hand priors [38] through diffusion guidance. Similarly, UniEgoMotion [36] proposes a unified diffusion framework for motion reconstruction and forecasting. While these methods achieve impressive kinematic accuracy, they fundamentally rely on supplementary inputs, such as pre-computed SLAM trajectories, 3D point clouds, or external hand trackers. This reliance on specialized hardware limits their applicability to casual, in-the-wild monocular videos.

Table 1: Large-Scale Multi-Modal Pretraining Dataset. Building upon the scene-centric foundation of EgoM2P [22], our dataset includes dense hand and full-body kinematics to capture holistic human-scene interactions. By scaling the total pretraining corpus from 4B to 7B unique tokens, ReViV establishes a new data foundation for egocentric 4D understanding. (Note: ✓: available, ×: unavailable. ✓*: our generated pseudo-labels. Grey: baseline usage. Color: newly integrated datasets/modalities.

Datasets \ Modalities	RGB	Depth	Gaze	Camera	Hand	Body
EgoExo4D [14]	✓	×	✓	✓	×	×
HoloAssist [50]	✓	✓*	✓	✓	✓	×
HOT3D (Aria) [3]	✓	✓*	✓	✓	✓	×
HOT3D (Quest) [3]	✓	✓*	×	✓	✓	×
ARCTIC [11]	✓	✓*	×	✓	✓	×
TACO [27]	✓	✓*	×	✓	✓	×
H2O [20]	✓	✓	×	✓	✓	×
EgoGen [23]	✓	✓	×	✓	×	×
Nymeria [32]	✓	✓*	✓	✓	×	✓

Furthermore, gaze cues have been integrated into egocentric motion understanding. [62] forecasts future gaze in the 3D scene from 2D visual inputs, and [35] proposes a gaze-regularized attention mechanism that enhances vision-language models (VLMs) for egocentric behavior understanding, particularly for activity recognition and future prediction.

Together, these efforts highlight the growing interest in egocentric human motion and scene modeling, yet a unified approach for 4D reconstruction that jointly models human body, hand, and gaze dynamics along with scene geometry remains an open challenge. In this work, we propose ReViV, the first unified generative framework for 4D reconstruction in the egocentric domain.

3 Data Engine

One major limitation of prior work lies in the lack of a unified data organization for multimodal pretraining, which prevents training a single framework across heterogeneous egocentric datasets. To train ReViV at scale (Sec. 4), we construct an automated multi-modal data engine that harmonizes spatial representations and compensates for missing modalities across diverse datasets (Tab. 1). Building upon the scene-centric foundation of EgoM2P [22], we extend the corpus with dense hand and full-body kinematics, increasing the pretraining set from 4B to a unified 7B-token dataset tailored for holistic egocentric 4D modeling.

3.1 Temporally Consistent Geometric Pseudo-Labeling

Many egocentric datasets lack dense 3D geometry due to hardware constraints or incomplete sensor streams. To address this, we generate high-quality, temporally consistent video depth pseudo-labels using Video Depth Anything [5]. This

automated pipeline enables us to scale geometric supervision beyond the constrained hand-object interactions used in EgoM2P [22] to diverse, unconstrained real-world environments. The resulting large-scale depth annotations provide consistent geometric signals over time, allowing ReViV to learn robust 4D structural priors. As shown in our evaluations, this large-scale geometric pretraining improves video depth prediction performance in complex scenes compared to existing baselines.

3.2 Unified Kinematic Representation

Egocentric datasets differ significantly in coordinate conventions and sensor setups, leading to inconsistencies in kinematic annotations. To enable unified modeling, we standardize all motion signals into two complementary reference frames that disentangle local manipulation from global navigation.

Camera-Space Hand Kinematics. We project the hand joints into the camera space of their current respective frames. While dependent on the immediate camera pose, this provides a trajectory-invariant representation relative to the user's viewpoint, encouraging the model to learn interaction priors independent of global body translation.

Gravity-Aligned Global Body Kinematics. Anchoring trajectories to a raw initial camera pose introduces pitch and roll artifacts. Instead, we establish a stable world reference frame by taking the first frame's camera pose and projecting its up-vector to align opposite to gravity. This creates a ground-parallel reference frame that isolates true global navigation from the camera's initial tilt.

4 Method

We propose a unified framework to jointly reconstruct human-centric (gaze, body, hand motion) and scene-centric (camera trajectories, depth maps) features from monocular egocentric video (Fig 2). Because the viewer's body is heavily occluded in egocentric views, deterministic estimation is inherently ambiguous. To address this, we formulate the task as a generative modeling problem and approximate the multimodal joint distribution through masked token prediction.

4.1 Problem Formulation

Let $\mathcal{X} = \{\mathbf{x}_{\text{rgb}}\}$ represent the observed egocentric video context, and $\mathcal{Y} = \{\mathbf{y}_{\text{hand}}, \mathbf{y}_{\text{body}}, \mathbf{y}_{\text{gaze}}, \mathbf{y}_{\text{depth}}, \mathbf{y}_{\text{cam}}\}$ denote the human and scene dynamic states to reconstruct. Directly learning a deterministic mapping $f : \mathcal{X} \to \mathcal{Y}$ is ill-posed due to the severe occlusion of the ego-body, which renders the underlying solution space highly multimodal. To address this, we cast the problem as learning the *joint probability distribution* across all observed and unobserved modalities:

$$p(\mathcal{X}, \mathcal{Y}) = p(\mathbf{x}_{\text{rgb}}, \mathbf{y}_{\text{hand}}, \mathbf{y}_{\text{body}}, \mathbf{y}_{\text{gaze}}, \mathbf{y}_{\text{depth}}, \mathbf{y}_{\text{cam}}). \tag{1}$$

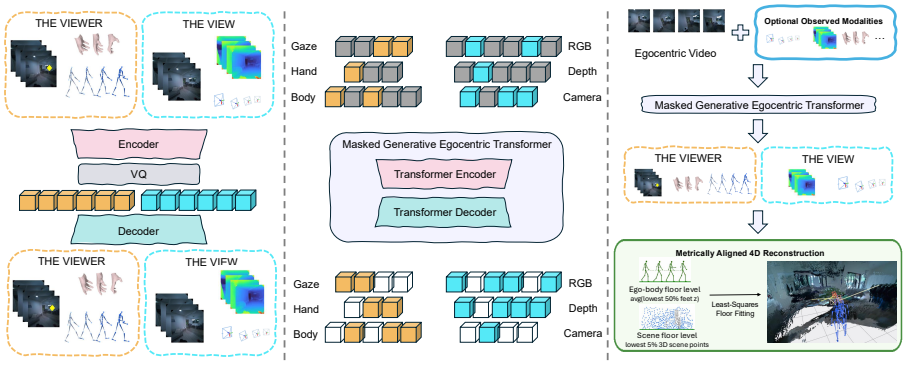

Fig. 2: ReViV Architecture. **(Left)** Modality-specific VQ-VAEs discretize heterogeneous viewer (gaze, hand, body) and view (RGB, depth, camera) continuous modalities into a unified token sequence. **(Middle)** A Masked Generative Egocentric Transformer (MGET) learns the joint probability distribution by predicting randomly masked tokens (grey blocks), capturing both intra- and cross-modal dynamics. **(Right)** During inference, MGET conditions on available observed tokens (e.g., RGB) to iteratively decode and reconstruct the unobserved human and scene states. Floor fitting aligns the reconstruction to a metric 4D coordinate system.

In practice, we approximate this joint distribution through masked multi-modal token prediction, which enables the model to learn both intra-modal temporal dynamics and cross-modal correlations. During inference, this allows us to flexibly sample from the conditional distribution $p(\mathcal{Y} \mid \mathcal{X})$ to reconstruct the human and scene states.

Metric Alignment of 4D Reconstruction. To achieve a metric-aligned 4D reconstruction, we introduce a lightweight alignment step to map the relative scene depth into the metric domain. Specifically, we fit a stable floor plane from the reconstructed scene geometry and use it to regularize the metric scale, with details in supplementary Sec. D. For RGB inputs that do not contain a visible floor, we leverage off-the-shelf metric depth predictions from VIPE [17] as an optional geometric anchor. These alignment steps allow the reconstructed viewer and view to be placed within a shared, metric-aligned 4D coordinate system.

4.2 Unified Discrete Representation

Modeling the joint distribution $p(\mathcal{X}, \mathcal{Y})$ in the continuous domain is challenging due to the heterogeneity of the modalities, which range from dense pixel arrays to sparse kinematic vectors. We therefore leverage a unified discrete latent interface to homogenize the modalities.

For each modality $m \in \mathcal{X} \cup \mathcal{Y}$ with continuous input dimension D_m, we employ a modality-specific Vector Quantized Variational Autoencoder (VQ-VAE) to learn a quantization function $Q_m : \mathbb{R}^{D_m} \to \mathcal{Z}_m$. This function maps the continuous signal to a sequence of L_m discrete tokens $\mathbf{z}_m = (z_{m,1}, \ldots, z_{m,L_m})$. These

tokens are indices drawn from a learned codebook $\mathcal{C}_m \in \mathbb{R}^{K_m \times d}$, where K_m denotes the vocabulary size for modality m and d is the latent feature dimension. The full multi-modal state is represented as a unified sequence of tokens $\mathbf{Z}$:

$$\mathbf{Z} = [\mathbf{z}_{\text{rgb}}, \mathbf{z}_{\text{depth}}, \mathbf{z}_{\text{cam}}, \mathbf{z}_{\text{gaze}}, \mathbf{z}_{\text{hand}}, \mathbf{z}_{\text{body}}] \tag{2}$$

For human body and hand tokenization, to address the severe scarcity of high-quality, unified whole-body datasets, we design a decoupled, dual-stream tokenization framework capable of leveraging disjoint body-only and hand-only motion captures. Given an input 3D joint sequence $\mathbf{y}_m \in \mathbb{R}^{N \times J \times 3}, m \in \{\text{hand}, \text{body}\}$ consisting of N frames and J joints, our body and hand tokenizers independently apply 2D convolutions to extract local kinematic priors. These features are subsequently processed by 12-block Transformer encoders, projecting the raw kinematics into a dense, continuous latent space of 210-dimensional embeddings. To quantize these representations while ensuring high dictionary utilization across highly varied datasets, we leverage a spherical quantization codebook. By projecting the continuous embeddings onto a unit sphere and pairing this with an Exponential Moving Average (EMA) codebook update strategy, we effectively prevent codebook collapse and maintain an expressive discrete latent space for both full-body and fine-grained hand domains. Symmetric 12-block Transformer decoders then reconstruct the continuous sequences, optimized via standard L_2 reconstruction and codebook commitment losses. For the dense visual modalities, we quantize RGB and depth videos using Cosmos Tokenizers [1]. Furthermore, to capture a broader distribution of other ego-state dynamics, we re-train the gaze and camera trajectory tokenizers from EgoM2P [22] on our scaled-up 7B-token dataset (Tab. 1). This discretization transforms the continuous regression problem into a sequence modeling task defined over modality-specific discrete vocabularies ($\mathcal{C}_m$), ensuring that predictions for each modality remain constrained to their respective codebook spaces.

4.3 Masked Generative Egocentric Transformer

We approximate the joint distribution $p(\mathbf{Z})$ using a Masked Generative Egocentric Transformer (MGET) parameterized by θ. Specifically, we adapt an encoder-decoder architecture based on T5 [41], comprising 12 Transformer blocks in both the encoder and decoder, with a hidden dimension of 768. Because the unified sequence $\mathbf{Z}$ comprises highly heterogeneous data, we first map the discrete tokens to continuous feature vectors via modality-specific embedding layers. To explicitly inject structural and semantic context, we add learnable modality-type embeddings and fixed sinusoidal spatiotemporal positional encodings to the sequence before processing. To compensate for the fine-grained pixel details inevitably lost during discrete quantization, we introduce an additional trainable Vision Transformer (ViT) that directly encodes the raw egocentric video into continuous spatially detailed features $\mathbf{x}_{\text{ViT}}$. These features are incorporated as additional visual context, yielding $\mathcal{X} = \{\mathbf{x}_{rgb}, \mathbf{x}_{\text{ViT}}\}$.

Following the paradigm of masked modeling [4,58], we define a training objective that learns the inter-modal dependencies across the unified token sequence

$\mathbf{Z}$ representing the full state $\mathcal{X} \cup \mathcal{Y}$. During training, we randomly sample a binary mask $\mathbf{M} \in \{0,1\}^{|\mathbf{Z}|}$, partitioning the sequence into masked tokens $\mathbf{Z}_{\mathcal{M}}$ and visible tokens $\mathbf{Z}_{\mathcal{V}}$. By randomly sampling the mask $\mathbf{M}$ across all modalities [2, 22], the network is forced to learn an ensemble of arbitrary conditional distributions $p(\mathbf{Z}_{\mathcal{M}} \mid \mathbf{Z}_{\mathcal{V}})$. Because a joint distribution can be factorized into a sequence of conditionals, optimizing over all possible mask configurations allows this objective to act as an efficient proxy for learning the underlying joint distribution $p(\mathbf{Z}) \approx p(\mathcal{X}, \mathcal{Y})$:

$$\mathcal{L}_{\text{mask}}(\theta) = -\mathbb{E}_{\mathbf{Z},\mathbf{M}} \left[\sum_{i \in \mathcal{M}, z_i \notin \mathbf{x}_{\text{ViT}}} \log p_\theta(z_i \mid \mathbf{Z}_{\mathcal{V}}) \right]. \tag{3}$$

Crucially, although optimized with a single cross-entropy loss, this random multimodal masking strategy serves as a multi-task training objective. First, by masking partial sequences within a single modality, the network learns intra-modal motion dynamics (e.g., $p(\mathbf{z}_{\text{body},\mathcal{M}} \mid \mathbf{z}_{\text{body},\mathcal{V}}), p(\mathbf{z}_{\text{hand},\mathcal{M}} \mid \mathbf{z}_{\text{hand},\mathcal{V}})$). This forces the model to encode strong temporal kinematic priors, enabling it to generate continuous and plausible motions without relying on explicit anatomical or smoothness regularizers. Second, by masking across different modalities, the network learns robust cross-modal correlations (e.g., inferring unobserved kinematics from visible context, $p(\mathbf{z}_{\text{body},\mathcal{M}} \mid \mathbf{z}_{\text{rgb}})$). Together, these two implicit tasks enable MGET to capture human kinematics and environment-conditioned interaction patterns directly from the multimodal training distribution.

The Transformer processes the input sequence to capture spatiotemporal dependencies to output latent representations. These features are projected through modality-specific linear heads to predict probability distributions over their respective vocabularies $\mathcal{C}_m$. At inference time, we achieve our primary reconstruction goal by formulating a specific masking condition that provides the observed egocentric video context as the visible tokens ($\mathbf{Z}_{\mathcal{V}} = \mathbf{Z}_{\mathcal{X}}$) and treat the human and scene state modalities as fully masked ($\mathbf{Z}_{\mathcal{M}} = \mathbf{Z}_{\mathcal{Y}}$). We then employ an iterative parallel decoding strategy to progressively sample masked tokens from these distributions, reconstructing the complete, temporally coherent 4D state. More implementation details are in Supplementary Material Sec. B and C.

5 Experiment

Evaluation Details. For all experiments, we split each sequence into 2-second clips. For video modalities (RGB and depth), we downsample the original 30 FPS to 8 FPS, and resize spatial resolution to 256× 256. For non-video modalities, including camera, gaze, body and hand, we retain the original 30 FPS.

5.1 Egocentric Body Motion Reconstruction

Baselines. We compare ReViV against EgoAllo [57] and UniEgoMotion [36], two recent SOTA diffusion-based models for egocentric full-body motion recovery.

Table 2: Egocentric Body Motion Reconstruction on ADT [34]. ReViV achieves SOTA performance and efficiency purely from monocular video. Even without explicit camera tracking, ReViV yields superior motion quality (FID), pose accuracy (PA-MPJPE), and pose similarity compared to baselines utilizing either estimated (VIPE) or ground-truth camera trajectories. ↓: lower is better; ↑: higher is better.

Methods	Input Cam Traj	GA-MPJPE ↓	PA-MPJPE ↓	Similarity ↑	FID ↓	Time (s) ↓
EgoAllo [57]	None	–	–	–	–	–
	VIPE [17]	227.6	167.5	0.529	1.069	101.0
	GT Cam	198.4	136.8	0.556	1.160	72.5
UniEgoMotion [36]	None	–	–	–	–	–
	VIPE [17]	130.5	98.4	0.572	1.087	37.6
	GT Cam	**105.8**	88.7	0.591	1.098	9.1
Ours	None	<u>111.5</u>	**88.6**	**0.751**	**0.442**	**0.7**

Evaluation Protocol. Since our large-scale pretraining corpus lacks ground-truth SMPL(-X) [28,37] annotations, we cannot retrain the baseline models under identical supervision. To ensure a fair comparison and assess generalization to unseen domains, we conduct all quantitative evaluations on the Aria Digital Twin dataset [34], which was strictly excluded from our training data. Specifically, we use 3,185 video clips with corresponding 3D body annotations. Unlike ReViV, existing baselines cannot operate purely from monocular RGB and require explicit camera trajectories. To ensure a fair comparison when ground-truth trajectories are unavailable, we provide these baselines with consistent camera poses estimated using the state-of-the-art camera tracking method VIPE [17].

Metrics. Absolute joint position errors can be heavily biased by global translation offsets, which come from both the ambiguity of monocular camera tracking and the dynamic articulation between the head-mounted camera and the body's root joint. To isolate true pose accuracy, we report two aligned variants of the Mean Per Joint Position Error (MPJPE). `GA-MPJPE` applies a single Procrustes alignment (optimizing rotation, translation, and scale) across the entire motion sequence. In contrast, `PA-MPJPE` applies this alignment independently on a per-frame basis to capture strictly local pose accuracy.

Beyond per-joint errors, we evaluate motion realism and distributional similarity. Following prior generative motion models, we project the sequences into a latent space using the pretrained TMR motion encoder [39]. From these embeddings, we compute the Fréchet Inception Distance (`FID`) to measure distributional alignment, and the cosine `Similarity` between paired embeddings to quantify motion-level correspondence.

Results. As shown in Table 2, ReViV establishes a new SOTA for monocular egocentric body motion reconstruction. Compared to baselines using estimated camera trajectories from VIPE [17], our method dominates across all metrics while being 100× faster than EgoAllo and over 10× faster than UniEgoMotion during inference. We attribute this to our unified framework: unlike baselines that suffer from compounding camera-tracking errors, ReViV directly models the joint distribution $p(\mathcal{X}, \mathcal{Y})$, bypassing error-prone intermediate tracking entirely.

Table 3: Quantitative Comparison of Egocentric Hand Motion Reconstruction. ReViV achieves state-of-the-art accuracy across four benchmarks while operating orders of magnitude faster than continuous optimization baselines (e.g., Dyn-HaMR). Metrics evaluated include Global-Aligned (GA-MPJPE), Root-Aligned (RA-MPJPE), and Procrustes-Aligned (PA-MPJPE) errors. ↓ indicates lower is better.

Methods	Time (s) ↓	HoloAssist [50]			HOT3D [3]		
		GA-MPJPE ↓	RA-MPJPE ↓	PA-MPJPE ↓	GA-MPJPE ↓	RA-MPJPE ↓	PA-MPJPE ↓
HaMeR [38]	72	59.5	57.9	32.9	94.4	74.1	48.1
Dyn-HaMR [59]	280	49.6	31.5	23.0	107.7	**46.9**	32.0
Ours	**0.7**	**23.5**	**29.7**	**10.5**	**38.2**	48.1	**13.6**

Methods	Time (s) ↓	ARCTIC [11]			TACO [27]		
		GA-MPJPE ↓	RA-MPJPE ↓	PA-MPJPE ↓	GA-MPJPE ↓	RA-MPJPE ↓	PA-MPJPE ↓
HaMeR [38]	72	95.9	39.2	28.6	52.0	39.3	29.1
Dyn-HaMR [59]	280	76.9	33.3	22.4	48.0	32.5	23.5
Ours	**0.7**	**35.1**	**33.0**	**13.7**	**20.3**	**25.2**	**9.4**

Notably, even without camera trajectories as input, ReViV outperforms baselines that rely on ground-truth cameras across local pose accuracy (`PA-MPJPE`), semantic correspondence (`Similarity`), and motion realism (`FID`). This advantage stems from our discrete tokenization and multimodal masking objective, which forces the model to learn robust, biomechanically plausible kinematic priors rather than over-relying on rigid trajectory inputs. While baselines equipped with perfect external tracking yield slightly better global alignment (`GA-MPJPE`), ReViV remains highly competitive, demonstrating that global ego-motion can be effectively learned implicitly from pure monocular video. See additional qualitative visualizations in Supplementary Material Sec. E.2.

5.2 Egocentric Hand Motion Reconstruction

Baselines. We compare ReViV against two recent state-of-the-art (SOTA) methods for egocentric hand motion reconstruction: HaMeR [38] and Dyn-HaMR [59]. We use the official implementations of both methods and follow their recommended evaluation settings whenever applicable.

Evaluation Protocol. We evaluate our method on four public egocentric hand motion benchmarks: HoloAssist [50] (27,910 validation clips), HOT3D [3] (1,519 clips), ARCTIC [11] (449 clips), and TACO [27] (998 clips). In addition to our body motion evaluation, we report Global-Aligned MPJPE and Procrustes-Aligned MPJPE, along with Root-Aligned MPJPE (`RA-MPJPE`), which offsets the hand joints by per-frame root joint positions before computing the mean per-joint error. To assess computational efficiency, we record the average inference time per clip. Since the Dyn-HaMR baseline relies on computationally intensive test-time optimization, evaluating it across the massive HoloAssist dataset is prohibitively expensive. Therefore, to ensure a fair and tractable comparison, we randomly sample 3,000 clips from the HoloAssist [50] validation set and report both accuracy and runtime statistics on this representative subset.

Results. Table 3 shows ReViV outperforms both frame-based (HaMeR) and dynamic (Dyn-HaMR) baselines across four diverse datasets. Notably, ReViV

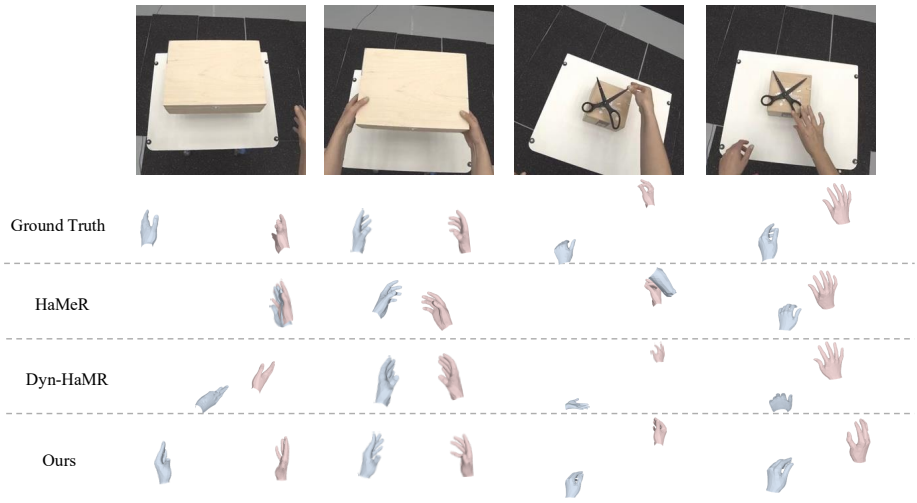

Fig. 3: Qualitative Results of Egocentric Hand Reconstruction. ReViV learns robust spatiotemporal motion priors to generate plausible hand motions, even under severe occlusion or when the hands are out of view. We visualize sampled, non-consecutive frames across two distinct sequences to demonstrate temporal stability.

Table 4: Egocentric camera tracking, gaze estimation and depth estimation on ADT. We additionally report the average runtime per clip for each method over the camera tracking task, where EgoM2P and our method achieve real-time performance.

		Camera			Gaze	Depth	
Methods	Time (s) ↓	ATE ↓	RTE ↓	RRE ↓	MSE ↓	Abs Rel ↓	$\delta_{1.25}$ ↑
EgoM2P [22]	**0.7**	0.030	**0.009**	1.290	0.0311	0.458	30.3
EgoMono4D [60]	14.2	0.051	0.015	1.307	-	**0.150**	**83.9**
VIPE [17]	25.8	**0.005**	**0.009**	1.307	-	-	-
Ours	**0.7**	<u>0.015</u>	**0.009**	**1.279**	**0.0211**	<u>0.265</u>	<u>56.5</u>

achieves SOTA local pose accuracy (`PA-MPJPE`) and global alignment (`GA-MPJPE`) while being 100× faster than HaMeR and over 400× faster than Dyn-HaMR.

Furthermore, qualitative visualizations shown in Fig. 3 highlight ReViV's robustness to severe occlusions common in egocentric vision. While baselines typically collapse or predict broken kinematics under these conditions, ReViV leverages strong spatiotemporal priors learned from the joint distribution. This enables ReViV to produce biomechanically plausible, temporally consistent hand movements across temporal gaps (e.g., between frames 4 and 10 in the left sequence, and frames 2 and 6 in the right), even when the hands are entirely unobservable from the egocentric view.

5.3 Egocentric Camera Tracking

Baselines. We compare ReViV with egocentric-focused 4D reconstruction methods (EgoM2P [22] and EgoMono4D [60]), as well as the SOTA general-scene 4D reconstruction method VIPE [17]. EgoMono4D relies on off-the-shelf optical flow predictions and reformulates camera tracking as depth alignment across confidence-aware correspondences. Similarly, VIPE performs dense bundle adjustment over off-the-shelf predictions of dense optical flow, sparse tracks, and metric depth. In contrast, ReViV does not rely on explicit geometric constraints and predicts camera poses directly from egocentric RGB input end-to-end.
Evaluation Protocol. We evaluate on all sequences from the unseen ADT dataset. For camera tracking, we report standard error metrics, including Absolute Translation Error (ATE), Relative Translation Error (RTE), and Relative Rotation Error (RRE). All predicted camera trajectories are aligned to the ground truth using a Sim(3) transformation estimated over the entire clip.
Results. See Tab. 4 for quantitative results. Our method outperforms all baselines except VIPE on the ATE metric, while achieving the lowest RTE and RRE among all methods. The performance gap with VIPE on ATE can be attributed to its use of dense bundle adjustment over the entire input clip, whereas our method predicts camera poses in a single feedforward pass without geometric post-processing. Qualitative results are in the Supplementary Material Sec. E.

5.4 Egocentric Gaze Estimation

We compare ReViV with EgoM2P [22] on the ADT dataset to predict 2D gaze locations from egocentric videos. Both the predicted gaze locations and ground truth are normalized to $[0, 1]$ with respect to the input resolution. We report the Mean Squared Error (MSE) as the evaluation metric. As reported in Tab. 4, our method achieves significantly lower error compared to EgoM2P on unseen data. Qualitative results are provided in the Supplementary Material Sec. E.

5.5 Egocentric Video Depth Estimation

Baselines. We compare ReViV against egocentric-focused 4D reconstruction methods EgoM2P [22] and EgoMono4D [60] for depth estimation. EgoMono4D leverages UniDepth [40], a specialized monocular depth estimation model, by finetuning on egocentric data. In contrast, our method is trained from scratch and does not rely on expert priors.
Evaluation Protocol. We use all sequences with depth annotations on ADT [34] and filter out clips where the timestamp difference between RGB and depth frames exceeds 10 ms to avoid misalignment. For all methods, predictions are aligned with the ground truth with a clip-wise scale and translation factor. We then report relative metrics, including Absolute Relative Error (Abs Rel) and the percentage of predicted depths within a 1.25 factor of the ground truth ($\delta_{1.25}$).
Results. The quantitative results are reported in Tab. 4. ReViV significantly outperforms EgoM2P on the unseen dataset but falls behind EgoMono4D on

Table 5: Ablation on Ego Body Motion, Camera Tracking & Depth.

	Body Motion				Camera		Depth	
Methods	GA ↓	PA ↓	Sim ↑	FID ↓	ATE ↓	RRE ↓	Rel ↓	$\delta_{1.25}$ ↑
ReViV	**111.5**	**88.6**	**0.751**	**0.442**	**0.015**	1.279	**0.265**	**56.5**
ReViV-BodyData	134.0	109.3	0.645	0.478	–	–	–	–
ReViV-BodySpec.	200.3	139.9	0.545	0.623	–	–	–	–
ReViV-w/oNy	–	–	–	–	0.031	1.293	0.432	32.0
ReViV-w/oNy&Holo	–	–	–	–	0.032	1.335	0.452	30.1
ReViV-1to1Mask	158.3	120.6	0.615	0.548	0.036	**1.267**	0.401	36.8
ReViV-Uniform	172.7	119.8	0.598	0.665	0.053	1.529	0.494	26.7
ReViV-Tiny	143.1	116.7	0.634	0.486	0.026	1.269	0.337	44.7

both metrics. We attribute this gap to two main factors: (1) EgoMono4D benefits from strong depth-specific priors inherited from UniDepth pretrained weights, whereas ours is trained without depth expert initialization; and (2) the discrete tokenization from the Cosmos tokenizer [1] introduces quantization errors that can affect fine-grained depth prediction. While there is a gap, our method is substantially more efficient, achieving over **20**× faster inference than EgoMono4D.

5.6 Ablation Studies

We conduct systematic ablation studies to validate each core design choice of ReViV, as summarized in Tab. 5. Complementary hand reconstruction ablations are provided in the Supplementary Material Sec. F.

Task-Specific vs. Unified Model. Training a task-specific body specialist (ReViV-BodySpec.) with only RGB input suffers a massive accuracy drop across all metrics compared to ReViV. Isolated single-modality training fails to exploit the rich cross-modal supervision signals captured by our joint formulation.

Model Capacity. A lightweight variant (ReViV-Tiny, ∼127M parameters) exhibits noticeably degraded performance across body motion, camera tracking, and depth estimation. This confirms that the more parameters is critical for approximating the complex joint distribution underlying our large-scale multimodal dataset, and that reconstruction quality scales with model capacity.

Masking Strategy. We trained an independent 1-to-1 modality masking by setting Dirichlet α to 0.001 (ReViV-1to1Mask), missing deep cross-modal info. Another ablation using uniform sampling on all modalities (ReViV-Uniform), which causes sparse body tokens with unbalanced dense video tokens, collapsing accuracy. Our sampling strategy mitigates both by producing all possible mask permutations for an ensemble of arbitrary conditional distributions.

Data Scaling. To disentangle data volume, we incrementally removed training datasets (Nymeria, and Nymeria + HoloAssist), yielding ReViV-w/oNy and ReViV-w/oNy&Holo . Experiment results confirm performance scales with data volume. Models (ReViV-BodyData) only trained on datasets with body annotations degrade accuracy compared to our full model. This demonstrates that even without explicit annotations, joint training on massive datasets significantly enhances reconstruction by enriching generalized cross-modal priors.

6 Conclusion

In this work, we presented ReViV, the first unified generative framework to bridge the longstanding gap between egocentric scene reconstruction and unobserved human kinematics. By casting the inherently ambiguous task of ego-body estimation as a joint probability learning problem, we demonstrated that a single monocular RGB video is sufficient to reconstruct both the dynamic 4D environment and the full-body, hand, and gaze behaviors of the camera wearer. At the core of our approach is the Masked Generative Egocentric Transformer (MGET), which successfully aligns these heterogeneous modalities into a shared, temporally coherent latent space. Through large-scale pretraining, ReViV establishes a new foundational baseline for holistic egocentric perception, achieving state-of-the-art performance without relying on specialized hardware or pre-computed SLAM. Ultimately, by unifying the viewer and the view within a single feed-forward architecture, this work provides a stepping stone for next-generation embodied AI systems capable of natural, human-centric reasoning.

Limitations and Future Directions. We leverage VQ-VAEs to resolve the heterogeneity of egocentric multimodalities. By mapping these diverse continuous signals into a unified discrete token space, we enable a single Transformer to jointly model both the viewer and the view, benefiting from the stable scaling laws that drive modern foundation models. However, this unified discrete representation introduces an inherent trade-off for dense regression tasks. Specifically, the quantization process inevitably discards some high-frequency spatial details, which slightly limits the model's capacity for depth estimation compared to task-specific continuous regression works. Exploring the integration of continuous generative priors, such as conditional diffusion models, into the decoding stage could be a future direction. This would allow our framework to retain its robust multimodal reasoning capabilities while effectively recovering the fine-grained geometric details necessary for high-fidelity 4D scene reconstruction.

Acknowledgements. Xiaozhong Lyu and Gen Li contributed equally to this work. Gen Li formulated the core idea, developed the main codebase, implemented the hand tokenizer, trained the main model (without the body component), and led the paper writing. Xiaozhong Lyu drove the system scale-up and finalization by expanding the multimodal dataset to 7B tokens, developing the body tokenizer, retraining the gaze and camera tokenizers, training the final full-scale main model, conducting experiments, and creating visualizations.

Xiaozhong Lyu was supported by an ETH Zurich Research Grant. Gen Li was supported by a Microsoft Spatial AI Zurich Lab PhD scholarship. This work was also supported as part of the Swiss AI Initiative by a grant from the Swiss National Supercomputing Centre (CSCS) under project IDs a143, a144 on Alps. We thank Zinuo You and Malte Prinzler for proofreading.

References

1. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025)
2. Bachmann, R., Kar, O.F., Mizrahi, D., Garjani, A., Gao, M., Griffiths, D., Hu, J., Dehghan, A., Zamir, A.: 4m-21: An any-to-any vision model for tens of tasks and modalities. Advances in Neural Information Processing Systems **37**, 61872–61911 (2024)
3. Banerjee, P., Shkodrani, S., Moulon, P., Hampali, S., Han, S., Zhang, F., Zhang, L., Fountain, J., Miller, E., Basol, S., Newcombe, R., Wang, R., Engel, J.J., Hodan, T.: HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. CVPR (2025)
4. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2022)
5. Chen, S., Guo, H., Zhu, S., Zhang, F., Huang, Z., Feng, J., Kang, B.: Video depth anything: Consistent depth estimation for super-long videos. In: CVPR (2025)
6. Chen, Y., Chen, X., Xue, Y., Chen, A., Xiu, Y., Pons-Moll, G.: Human3r: Everyone everywhere all at once. arXiv preprint arXiv:2510.06219 (2025)
7. Chu, W.H., Ke, L., Fragkiadaki, K.: Dreamscene4d: Dynamic multi-object scene generation from monocular videos. NeurIPS **37**, 96181–96206 (2024)
8. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: The epic-kitchens dataset: Collection, challenges and baselines. IEEE TPAMI **43**(11), 4125–4141 (2020)
9. Deshmukh, M., Akada, H., Rhodin, H., Theobalt, C., Golyanik, V.: E-3dpsm: A state machine for event-based egocentric 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14017–14026 (2026)
10. Duan, Y., Wei, F., Dai, Q., He, Y., Chen, W., Chen, B.: 4d-rotor gaussian splatting: towards efficient novel view synthesis for dynamic scenes. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024)
11. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
12. Feng, H., Zhang, J., Wang, Q., Ye, Y., Yu, P., Black, M.J., Darrell, T., Kanazawa, A.: St4rtrack: Simultaneous 4d reconstruction and tracking in the world. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8503–8513 (2025)
13. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: CVPR. pp. 18995–19012 (2022)
14. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, e.a.: Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19383–19400 (June 2024)
15. Guzov, V., Jiang, Y., Hong, F., Pons-Moll, G., Newcombe, R., Liu, C.K., Ye, Y., Ma, L.: Hmd 2: Environment-aware motion generation from single egocentric head-mounted device. In: 2025 International Conference on 3D Vision (3DV). pp. 1394–1405. IEEE (2025)

16. He, Y., Huang, Y., Chen, G., Lu, L., Pei, B., Xu, J., Lu, T., Sato, Y.: Bridging perspectives: A survey on cross-view collaborative intelligence with egocentric-exocentric vision. International Journal of Computer Vision **134**(2), 62 (2026)
17. Huang, J., Zhou, Q., Rabeti, H., Korovko, A., Ling, H., Ren, X., Shen, T., Gao, J., Slepichev, D., Lin, C.H., Ren, J., Xie, K., Biswas, J., Leal-Taixe, L., Fidler, S.: Vipe: Video pose engine for 3d geometric perception. In: NVIDIA Research Whitepapers arXiv:2508.10934 (2025)
18. Jiang, H., Grauman, K.: Seeing invisible poses: Estimating 3d body pose from egocentric video. In: CVPR. pp. 3501–3509. IEEE (2017)
19. Jiang, H., Ithapu, V.K.: Egocentric pose estimation from human vision span. In: ICCV. pp. 10986–10994. IEEE (2021)
20. Kwon, T., Tekin, B., Stühmer, J., Bogo, F., Pollefeys, M.: H2o: Two hands manipulating objects for first person interaction recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10138–10148 (October 2021)
21. Lei, J., Weng, Y., Harley, A.W., Guibas, L., Daniilidis, K.: Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. In: CVPR. pp. 6165–6177 (2025)
22. Li, G., Chen, Y., Wu, Y., Zhao, K., Pollefeys, M., Tang, S.: Egom2p: Egocentric multimodal multitask pretraining. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10830–10843 (October 2025)
23. Li, G., Zhao, K., Zhang, S., Lyu, X., Dusmanu, M., Zhang, Y., Pollefeys, M., Tang, S.: Egogen: An egocentric synthetic data generator. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14497–14509 (June 2024)
24. Li, J., Liu, K., Wu, J.: Ego-body pose estimation via ego-head pose estimation. In: CVPR. pp. 17142–17151 (2023)
25. Li, Z., Dwivedi, S.K., Maric, F., Chacon, C., Bertsch, N., Arcadu, F., Hodan, T., Ramamonjisoa, M., Wonka, P., Zhao, A., et al.: Egoposeformer v2: Accurate egocentric human motion estimation for ar/vr. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21121–21131 (2026)
26. LIU, Q., Liu, Y., Wang, J., Lyu, X., Wang, P., Wang, W., Hou, J.: MoDGS: Dynamic gaussian splatting from casually-captured monocular videos with depth priors. In: The Thirteenth International Conference on Learning Representations (2025)
27. Liu, Y., Yang, H., Si, X., Liu, L., Li, Z., Zhang, Y., Liu, Y., Yi, L.: Taco: Benchmarking generalizable bimanual tool-action-object understanding. arXiv preprint arXiv:2401.08399 (2024)
28. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015)
29. Lu, J., Huang, T., Li, P., Dou, Z., Lin, C., Cui, Z., Dong, Z., Yeung, S.K., Wang, W., Liu, Y.: Align3r: Aligned monocular depth estimation for dynamic videos. In: CVPR. pp. 22820–22830 (2025)
30. Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In: 2024 International Conference on 3D Vision (3DV). pp. 800–809. IEEE (2024)
31. Luo, Z., Hachiuma, R., Yuan, Y., Kitani, K.: Dynamics-regulated kinematic policy for egocentric pose estimation. NeurIPS **34**, 25019–25032 (2021)

32. Ma, L., Ye, Y., Hong, F., Guzov, V., Jiang, Y., Postyeni, R., Pesqueira, L., Gamino, A., Baiyya, V., Kim, H.J., et al.: Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In: ECCV. pp. 445–465. Springer (2024)
33. Ng, E., Xiang, D., Joo, H., Grauman, K.: You2me: Inferring body pose in egocentric video via first and second person interactions. In: CVPR. pp. 9890–9900 (2020)
34. Pan, X., Charron, N., Yang, Y., Peters, S., Whelan, T., Kong, C., Parkhi, O., Newcombe, R., Ren, Y.C.: Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20133–20143 (October 2023)
35. Pani, A., Yang, Y.: Gaze-vlm: Bridging gaze and vlms through attention regularization for egocentric understanding. In: NeurIPS (2025)
36. Patel, C., Nakamura, H., Kyuragi, Y., Kozuka, K., Niebles, J.C., Adeli, E.: Uniegomotion: A unified model for egocentric motion reconstruction, forecasting, and generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10318–10329 (2025)
37. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 10975–10985 (2019)
38. Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., Malik, J.: Reconstructing hands in 3D with transformers. In: CVPR (2024)
39. Petrovich, M., Black, M.J., Varol, G.: Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9488–9497 (2023)
40. Piccinelli, L., Yang, Y.H., Sakaridis, C., Segu, M., Li, S., Van Gool, L., Yu, F.: UniDepth: Universal monocular metric depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
41. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020)
42. Rodin, I., Furnari, A., Mavroeidis, D., Farinella, G.M.: Predicting the future from first person (egocentric) vision: A survey. Computer Vision and Image Understanding **211**, 103252 (2021)
43. Stearns, C., Harley, A., Uy, M., Dubost, F., Tombari, F., Wetzstein, G., Guibas, L.: Dynamic gaussian marbles for novel view synthesis of casual monocular videos. In: SIGGRAPH Asia. pp. 1–11 (2024)
44. Tome, D., Alldieck, T., Peluse, P., Pons-Moll, G., Agapito, L., Badino, H., De la Torre, F.: Selfpose: 3d egocentric pose estimation from a headset mounted camera. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(6), 6794–6806 (2020)
45. Tome, D., Peluse, P., Agapito, L., Badino, H.: xr-egopose: Egocentric 3d human pose from an hmd camera. In: ICCV. pp. 7728–7738 (2019)
46. Wang, J., Liu, L., Xu, W., Sarkar, K., Theobalt, C.: Estimating egocentric 3d human pose in global space. In: ICCV. pp. 11500–11509 (2021)
47. Wang, Q., Ye, V., Gao, H., Zeng, W., Austin, J., Li, Z., Kanazawa, A.: Shape of motion: 4d reconstruction from a single video. In: ICCV. pp. 9660–9672 (2025)
48. Wang, Q., Zhang, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3d perception model with persistent state. In: CVPR. pp. 10510–10522 (2025)

49. Wang, S., Yang, X., Shen, Q., Jiang, Z., Wang, X.: Gflow: recovering 4d world from monocular video. In: Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence. pp. 7862–7870 (2025)
50. Wang, X., Kwon, T., Rad, M., Pan, B., Chakraborty, I., Andrist, S., Bohus, D., Feniello, A., Tekin, B., Frujeri, F.V., Joshi, N., Pollefeys, M.: Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20270–20281 (October 2023)
51. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: CVPR. pp. 20310–20320 (2024)
52. Xiao, Y., Wang, J., Xue, N., Karaev, N., Makarov, Y., Kang, B., Zhu, X., Bao, H., Shen, Y., Zhou, X.: Spatialtrackerv2: 3d point tracking made easy. arXiv preprint arXiv:2507.12462 (2025)
53. Xin, J., Wang, L., Xu, K., Yang, C., Yin, B.: Learning interaction regions and motion trajectories simultaneously from egocentric demonstration videos. IEEE Robotics and Automation Letters **8**(10), 6635–6642 (2023)
54. Xu, W., Chatterjee, A., Zollhoefer, M., Rhodin, H., Fua, P., Seidel, H.P., Theobalt, C.: Mo 2 cap 2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. IEEE Transactions on Visualization and Computer Graphics **25**(5), 2093–2101 (2019)
55. Xu, Z., Li, Z., Dong, Z., Zhou, X., Newcombe, R., Lv, Z.: 4dgt: Learning a 4d gaussian transformer using real-world monocular videos. arXiv preprint arXiv:2506.08015 (2025)
56. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In: CVPR. pp. 20331–20341 (2024)
57. Yi, B., Ye, V., Zheng, M., Li, Y., Müller, L., Pavlakos, G., Ma, Y., Malik, J., Kanazawa, A.: Estimating body and hand motion in an ego-sensed world. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7072–7084 (2025)
58. Yu, L., Cheng, Y., Sohn, K., Lezama, J., Zhang, H., Chang, H., Hauptmann, A.G., Yang, M.H., Hao, Y., Essa, I., Jiang, L.: MAGVIT: Masked generative video transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
59. Yu, Z., Zafeiriou, S., Birdal, T.: Dyn-hamr: Recovering 4d interacting hand motion from a dynamic camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2025)
60. Yuan, C., Chen, G., Yi, L., Gao, Y.: Self-supervised monocular 4d scene reconstruction for egocentric videos. arXiv preprint arXiv:2411.09145 (2024)
61. Yuan, Y., Kitani, K.: Ego-pose estimation and forecasting as real-time pd control. In: ICCV. pp. 10082–10092 (2019)
62. Yun, H., Na, J., Kim, J., Murdock, C., Kim, G.: Gaze beyond the frame: Forecasting egocentric 3d visual span. In: NeurIPS (2025)
63. Zhang, D., Li, G., Li, J., Bressieux, M., Hilliges, O., Pollefeys, M., Van Gool, L., Wang, X.: Egogaussian: Dynamic scene understanding from egocentric video with 3d gaussian splatting. In: 2025 International Conference on 3D Vision (3DV). pp. 1091–1102. IEEE (2025)