IMU4D: 4D HUMAN-OBJECT UNDERSTANDING FROM WEARABLE IMUS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding human activities and the objects they interact with typically relies on visual perception, yet cameras pose persistent challenges in privacy, safety, energy efficiency, and scalability. We explore an alternative: 4D perception without vision. Its goal is to recover human motion and the objects the person interacts with purely from everyday wearable sensors. For this we introduce IMU4D, a framework that repurposes large language models for non-visual spatiotemporal understanding of human-object dynamics. IMU4D uses data from a few inertial sensors from earbuds, watches, or smartphones and predicts detailed 4D human motion together with surrounding objects the person interacts with. Experiments across diverse human motion datasets and human-object datasets show that IMU4D yields more coherent and temporally stable results than SoTA cascaded pipelines, suggesting wearable motion sensors alone can support 4D understanding of human–object interactions.

1 INTRODUCTION

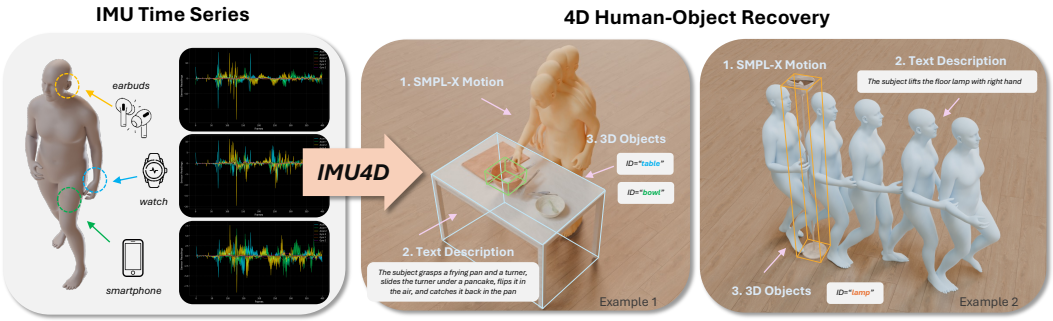

Figure 1: **IMU4D** jointly infers full 4D human–object dynamics from only a few everyday wearable IMUs. Given sparse inertial signals from devices such as earbuds, smartphones, or watches, our method predicts SMPL-X body motion, textual activity descriptions, and the category and 6-DoF pose of the 3D objects the person interacts with.

Recent advances in multimodal foundation models have expanded how AI perceives and assists in daily life. Yet their understanding of the physical world remains limited: most rely on vision-centric sensing such as cameras or LiDAR, which, while informative, introduces persistent challenges w.r.t. privacy, safety, energy efficiency, and scalability. As AI systems become more integrated into daily life, there is an increasing demand for non-visual, privacy-preserving, and low-power modalities that can still capture the richness of human-environment interaction.

Imagine an assistant that can recall how many times you drank water today, where you left your keys, or whether you've been sitting too long—without recording a single image. Such a system would extend AI's utility to personal health, memory, and accessibility. This motivates a research question: *How much can everyday motion sensors embedded in earbuds, watches, and smartphones reveal about our activities and environments?*

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

Table 1: **Motion- and object-understanding task comparison.**

Task	Input Modality	Output Modality		
		Motion	Text	3D Object
Motion-to-Text (Chen et al., 2024a; Delmas et al., 2024)	GT Motion + Visuals		✓	
Motion-Chat (Zhu et al., 2025a; Feng et al., 2024)	GT Motion	✓	✓	
Motion-to-Object (Ye et al., 2022; Nie et al., 2022)	GT Motion			✓
Ego-Understand (Hong et al., 2025; Grauman et al., 2024)	GT Poses + Visuals	✓	✓	
IMU-to-Motion (Xu et al., 2024c; Mollyn et al., 2023)	Raw IMU	✓		
IMU4D (Ours)	Raw IMU	✓	✓	✓

To answer this question, we propose **IMU4D**, a framework for vision-free 4D perception that infer detailed human motion and the 3D objects involved in the interaction purely from inertial signals. The key insight behind IMU4D: *human motion and the objects involved in an interaction are inherently coupled*—Inertial observations constrain how a person moves, while learned motion–object associations provide priors over the objects and configurations compatible with that motion. We exploit these complementary sources of information to jointly infer human motion, activity descriptions, and interacting objects.

Recent multimodal foundation models have demonstrated that such coupled structure across modalities can be effectively captured by leveraging rich pre-trained representations from LLMs and extending them to new modalities through unified architectures. Building on this intuition, IMU4D repurposes large language models (LLMs) as spatiotemporal reasoners, jointly modeling the distribution of sequential inertial signals, human motions, 3D objects the person interacts with, and semantic activities. To achieve this, we map each modality (IMU, motion, and 3D objects) into the embedding space of the LLM and formulate a unified sequential prediction problem, where the model first recovers human motion, then describes the activity in text, and finally infers the objects involved and their poses over time. The model is jointly finetuned with a multi-task loss in a two-stage training process. This unified formulation allows the model to infer coherent, context-aware 4D representations from as few as 2-3 inertial sensors—substantially improving temporal and spatial consistency compared with cascaded pipelines (IMU $\rightarrow$ motion $\rightarrow$ 3D object). Through the compositional reasoning capacity of LLMs, IMU4D bridges physical dynamics with spatial and semantic context, enabling a joint understanding of human-object interaction without relying on visual input.

Experiments on diverse datasets and real-world IMU records show that IMU4D achieves more coherent and temporally consistent reconstructions than prior approaches. Our results demonstrate that ubiquitous wearable sensors alone can support privacy-preserving 4D understanding of human activities and the surrounding objects involved in the interaction—paving the way for embodied AI that perceives the world through motion rather than sight.

Our contributions are threefold:

- We introduce the first framework that recovers 4D human motion, textual descriptions, and the categories and poses of objects the person interacts with from as few as three IMU sensors by repurposing LLMs for cross-modal structural reasoning.
- We formulate IMU-based 4D perception as a unified sequence prediction problem that jointly decodes human motion, text, and a variable number of objects together with their per-frame trajectories.
- Experiments show that our joint reasoning framework outperforms pipelined SoTA methods and enables recognition and localization of interacted objects without any visual input.

2 Related Work

IMU-Based Human Motion Estimation. Early works (Sony Corporation, 2025; Von Marcard et al., 2017; Huang et al., 2018; Yi et al., 2021; Dittadi et al., 2021; Yi et al., 2022b; Jiang et al., 2022a;b; Zuo et al., 2024; Yi et al., 2024; Van Wouwe et al., 2024; Xiao et al., 2024; Yi et al., 2025b; Zuo et al., 2025) explored IMU-based motion capture as a cost-effective alternative to optical systems (Moeslund & Granum, 2001), but typically rely on specialized body-worn sensors (Schepers et al., 2018). Recent works (Mollyn et al., 2023; Xu et al., 2024c; Zhang et al., 2025) instead

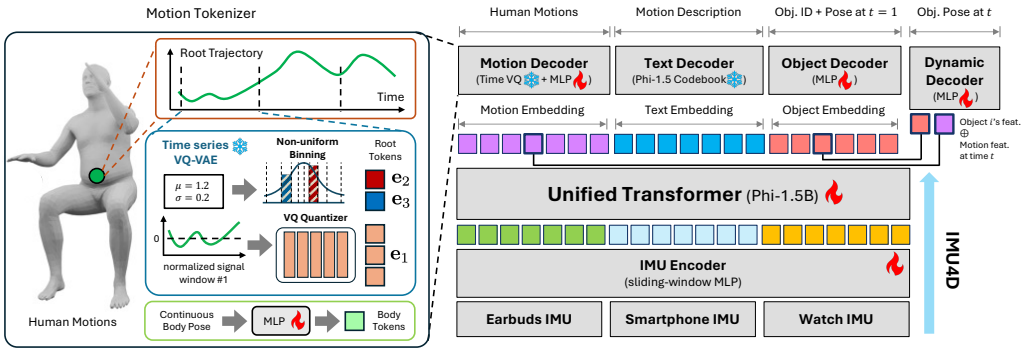

Figure 2: **Overview of IMU4D.** (Left) Motion Tokenizer. The root trajectory is chunked into fixed-length windows and normalized, where normalized chunks are VQ-quantized and normalization parameters μ and σ are separately quantized via non-uniform binning, yielding compact discrete root tokens. Body poses are continuously encoded via an MLP to produce body tokens. (Right) Multi-modal Transformer. Inertial signals from earbuds, smartphone, and watch are encoded and fed into a unified transformer, which jointly decodes human motions, motion descriptions, object IDs and their first-frame poses, and optionally object dynamics.

leverage commodity devices such as smartphones, smartwatches, and wireless earbuds, enabling more accessible motion sensing but remaining largely task-specific and lacking scene awareness. Other approaches incorporate egocentric video (Hong et al., 2025; Zhang et al., 2024b; Yi et al., 2025a; Lee & Joo, 2024; Patel et al., 2025) to provide richer contextual cues, at the cost of additional sensing and privacy concerns. In contrast, IMU4D jointly reasons about human motion, interacting 3D objects, and semantics using only sparse IMU signals from everyday devices, without visual input.

Motion-Object Synthesis. Human motion and interacting objects are strongly coupled both semantically and geometrically. Existing human-object interaction datasets (Bhatnagar et al., 2022; Li et al., 2023a; Zhang et al., 2024a; Xu et al., 2025; Lu et al., 2025a; Kim et al., 2025) enable generating motion conditioned on objects (Hassan et al., 2021; Li et al., 2023a; Zhao et al., 2023; Kulkarni et al., 2024; Jiang et al., 2024b; Ghosh et al., 2023) and inferring object configurations from motion (Nie et al., 2022; Ye et al., 2022; Yi et al., 2022a; 2023; Li et al., 2024b). Recent methods further jointly model humans and objects (Petrov et al., 2025; Xu et al., 2025; 2024b), with physics (Merel et al., 2020; Lee & Joo, 2023; Hassan et al., 2023; Braun et al., 2024) or language (Wang et al., 2024c; Jiang et al., 2024a; Li et al., 2024a; Xu et al., 2024b) providing additional constraints. However, these methods typically assume clean human poses from motion capture or simulation. In contrast, IMU4D directly reasons about human motion and interacting 3D objects from sparse, raw inertial measurements.

Large Multimodal Foundation Models. Large multimodal foundation models unify diverse modalities into shared representations for joint understanding and generation (Liu et al., 2023; Xie et al., 2024; Deng et al., 2025), achieving success across vision (Liu et al., 2023; Xie et al., 2024; Deng et al., 2025; Chen et al., 2025; Wang et al., 2024a; Wu et al., 2024b; Zhou et al., 2024; Liu et al., 2024; Qu et al., 2025; Wu et al., 2024a; Ge et al., 2024; Bai et al., 2025; Team, 2024; Chen et al., 2024b), audio (Ghosh et al., 2024; Zhang et al., 2023), and robotics (Kim et al., 2024; Intelligence et al., 2025). Recent works have extended this paradigm to inertial signals (Hong et al., 2025; Zhang et al., 2026) and human motion (Zhu et al., 2025a; Zhang et al., 2024d; Chen et al., 2024a; Zhang et al., 2024c; Wang et al., 2024b; Lu et al., 2025b; Li et al., 2025; Shi et al., 2025; Fan et al., 2025; Guo et al., 2025). However, their use for ambient sensing and human-object perception remains largely underexplored. IMU4D bridges this gap by jointly reasoning about inertial sensing, human motion, and surrounding 3D objects within a unified multimodal framework.

3 METHOD

IMU4D takes sparse IMU readings and jointly predicts 4D human motions, motion descriptions, and objects involved in the interaction. We repurpose large language models and train on diverse

motion–text–object datasets to enable unified 4D ambient human-object understanding. Fig. 2 shows the overall architecture. We describe the details next.

3.1 Problem Formulation

Formally, our goal is to learn a model $f_{\boldsymbol{\theta}}$ that takes IMU signals $\mathbf{S}_{\text{imu}}$ as input, and outputs human motions $\mathbf{M}$, motion descriptions $\mathbf{T}$, and 3D objects the person interacted with together with their trajectories $\mathbf{L}$:

$$\mathbf{M}, \mathbf{T}, \mathbf{L} = f_{\boldsymbol{\theta}}(\mathbf{S}_{\text{imu}}).$$

The input **IMU signal** is denoted as $\mathbf{S}_{\text{imu}} \in \mathbb{R}^{T \times N_{\text{imu}} \times d_{\text{imu}}}$, where T is the total number of time frames and N_{imu} is the number of IMU sensors. Each sensor provides linear acceleration $\boldsymbol{a} \in \mathbb{R}^3$, angular velocity $\boldsymbol{\omega} \in \mathbb{R}^3$ from a gyroscope, and a relative rotation $\mathbf{R} \in \mathbb{R}^9$ computed by integrating the angular velocity, giving $d_{\text{imu}} = 15$ readings per sensor per frame. Sensors are typically mounted at the ears, wrists, and thighs — corresponding to earbuds, watches, and smartphones placed in pockets. Since some devices may not always be available in practice, we set $N_{\text{imu}} \in [1, 5]$ to accommodate varying sensor availability. We assume a 6-DoF IMU configuration without magnetometer readings, as many wearable devices do not provide reliable and accessible magnetometer data, and such readings are often corrupted by environmental magnetic interference.

The output consists of three modalities. The **4D human motion** refers to motion trajectory $\mathbf{M} \in \mathbb{R}^{T \times 69}$, including global translation $\mathbf{t} \in \mathbb{R}^3$, global orientation $\mathbf{r}_{\text{root}} \in \mathbb{R}^3$, and body joint rotations $\mathbf{r}_{\text{joint}} \in \mathbb{R}^{N_{\text{joints}} \times 3}$, where $N_{\text{joints}} = 21$. We adopt SMPL-X (Pavlakos et al., 2019) as the body representation and convert all rotations into a 6D representation (Zhou et al., 2019) for stable training. The **text description** $\mathbf{T} \in \mathbb{Z}^{N_{\text{text}}}$ is a sequence of N_{text} discrete token indices encoding a natural language description of the observed activity. Finally, the **3D object** (category + pose, not geometry) is defined as a set of objects $\mathbf{L} = \{\mathbf{O}_i\}_{i=1}^{N_{\text{obj}}}$, where N_{obj} is the number of predicted objects. We support both static and dynamic objects, with an object-type label $\mathbf{o}_{\text{type}}$ indicating whether an object is static or dynamic. Each object $\mathbf{O}_i$ is characterized by a class label $\mathbf{o}_{\text{class}} \in \{1, \ldots, C\}$, where C denotes the number of object categories. For a static object, i.e., $\mathbf{o}_{\text{type}} =$ static, its pose is represented by an orientation $\mathbf{o}_{\text{orient}} \in \mathbb{R}^6$ and a translation $\mathbf{o}_{\text{transl}} \in \mathbb{R}^3$. For a dynamic object that moves over time with the person, its pose is represented by per-frame orientations $\mathbf{o}_{\text{orient}} \in \mathbb{R}^{T \times 6}$ and translations $\mathbf{o}_{\text{transl}} \in \mathbb{R}^{T \times 3}$ over T frames.

Taking IMU as input, the model sequentially generates motion, text, and objects. Specifically, we first predict the motions across all timesteps all at once, using learnable motion queries that attend to each other bidirectionally. Next, conditioned on the inputs and predicted motions, the model autoregressively decodes each text token until an <EOT> token is generated. Then the model predicts object class IDs one at a time, terminating when a designated stop token is produced, so the object count N_{obj} is determined dynamically. Finally, for each of the N_{obj} objects, the model predicts its initial 6-DoF pose and a per-frame trajectory in the body frame. A per-frame moving/static state gates the trajectory: static objects stay in place, and moving objects follow the person.

3.2 Tokenization and Embeddings

Our approach is a multimodal model that requires all modalities to be projected into a shared vector space. Therefore, effective encoding and tokenization are critical to our method. We first describe the tokenization strategies for each modality to balance compactness, expressiveness, and cross-modal consistency.

IMU Encoding. Given raw IMU signals of length T, we divide them into T/d_{win} non-overlapping windows ($d_{\text{win}} = 4$) and project each window into a continuous embedding of dimension d_{h} using an MLP to match the transformer embedding space. Each device is encoded independently, and the resulting embeddings are concatenated to form the full input sequence. Embeddings corresponding to inactive IMU devices are replaced with <MASK> embeddings.

Motion Tokens. Fig. 2 (left) illustrates our motion tokenization pipeline. We propose a novel motion tokenization scheme that preserves temporal coherence and achieves better reconstruction quality. Unlike FAST (Pertsch et al., 2025), which encodes motion tokens in the frequency domain, we follow prior works (Dwivedi et al., 2024; Xiao et al., 2025) and adopt a straightforward time-domain encoding. However, existing time-domain approaches fail to comprehensively tokenize

motion. TokenHMR (Dwivedi et al., 2024) does not explicitly encode temporal coherence, while Motion-VQ (Xiao et al., 2025) is limited in the reconstruction quality.

We identify that the low reconstruction quality of Motion-VQ stems from the pose normalization strategy. In Motion-VQ, poses are normalized using statistics computed over the entire dataset. Although normalized, the poses remain highly diverse, making it difficult for a fixed-size codebook to capture fine-grained motion details. To address this issue, we propose a per-window normalization and global shift-and-scale prediction scheme. For each N-frame window, poses are normalized using the statistics computed from those N frames, rather than the entire dataset. This significantly reduces the diversity within each normalized pose sequence, enabling more precise representation under quantization. The normalized N-frame poses are then tokenized and quantized into a small number of tokens using a VQ-VAE based on TOTEM (Talukder et al., 2024). Per-window normalization, however, removes the global shift and scale required to recover the original motion. To compensate for this loss, we additionally predict the mean and variance for each N-frame window. These predicted statistics enable reconstruction of the original pose values while maintaining temporal consistency across windows.

Finally, we describe our choice of pose representations. To ensure long-range stability, human poses are represented using absolute joint angles rather than frame-to-frame relative rotations. In contrast, global translation and rotation are expressed as relative motions to improve smoothness, reduce dynamic range, and remove bias toward absolute coordinates.

Object Tokens. We denote a collection of objects as $\mathbf{L} = \{\mathbf{O}_i\}_{i=1}^{N_{\text{obj}}}$, each described by a class label $\mathbf{o}_{\text{class}}$, static or dynamic $\mathbf{o}_{\text{type}}$, orientation $\mathbf{o}_{\text{orient}}$, and translation $\mathbf{o}_{\text{transl}}$. We tokenize the discrete $\mathbf{o}_{\text{class}}$ as one-hot labels and map to embeddings via a learned embedding layer. We generate objects sequentially in an autoregressive manner. At each step, the model predicts the class label $\mathbf{o}_{\text{class}}$ of the next object conditioned on all previously generated tokens, including the object IDs of preceding objects. We employ a specialized object detokenization process. Once an object class $\mathbf{o}_{\text{class}}$ is generated, the embedding of the generated object class is fed into a lightweight MLP decoder. The decoder predicts the object's static or dynamic label $\mathbf{o}_{\text{type}}$ together with its pose: a single orientation and translation for a static object, or a sequence of orientations and translations forming the object trajectory for a dynamic object.

Text Tokens. We adopt the tokenizer and detokenizer from the pretrained Phi-1.5 model (Li et al., 2023d) for texts. To support our multimodal setup, We extend the vocabulary with special tokens that mark the start and end of each IMU token, as well as the start and end of each output modality (text, motion, and object).

3.3 Architecture

Our architecture is illustrated in Fig. 2. We use Phi-1.5 (Li et al., 2023d) as our backbone and use pretrained checkpoints from Show-o (Xie et al., 2024). We adopt the 1.5B-parameter variant to balance efficiency and performance.

Multimodal Integration. To handle multiple modalities, we map all inputs into a unified continuous embedding space for joint sequential processing, and employ modality-specific decoders to detokenize the resulting representations, as detailed in Sec. 3.2. Modality-specific projection layers encode IMU readings, human motions, and 3D object attributes into embeddings of the same dimension as the language model's token embeddings.

3.4 Training

The model is trained with teacher forcing. We adopt a two-stage training strategy to leverage joint reasoning across modalities while adapting the model to individual downstream tasks:

Stage 1: Mixed dataset pretraining. Starting from the pretrained weights of Show-o, we jointly train the model to predict human motions, text descriptions, and 3D objects from IMU signals using a mixture of available datasets. For datasets that provide motion sequences but no IMU measurements, we synthesize the corresponding IMU signals from the motions. We apply supervision only to the modalities available in each sample. This mixed-dataset pretraining enables the model to leverage

Table 2: **Quantitative comparison of IMU-to-Motion on real-world IMU datasets DIP-IMU (Huang et al., 2018) and IMUPoser (Mollyn et al., 2023).** We omit MTE following MobilePoser (Xu et al., 2024c). Evaluated on whole sequences using a 3pt setup (ear, wrist, thigh).

	DIP-IMU				IMUPoser			
Method	MPJPE ↓	PA-MPJPE ↓	MPJRE ↓	MPJVE ↓	MPJPE ↓	PA-MPJPE ↓	MPJRE ↓	MPJVE ↓
IMUPoser	64.25	53.13	9.06	85.83	66.86	52.62	9.81	84.24
MobilePoser	62.91	51.33	7.95	81.34	61.96	50.07	8.44	75.59
Ours	**47.31**	**39.97**	**5.31**	**58.71**	**42.45**	**37.43**	**5.70**	**54.71**



Figure 3: **Real-world capture in a meeting room.** From three consumer IMUs, MobilePoser drifts away from the subject and leaves the frame, whereas our prediction stays registered to the person throughout the sequence. Numbers denote frame indices at 30 fps; the RGB frames are shown only for visualization and are not used as input.

complementary supervision from diverse data sources and provides a strong initialization for the subsequent training stage.

Stage 2: Task-specific finetuning. We further finetune the pretrained model on each specific downstream task and its corresponding dataset to adapt the learned representations to task-specific data distributions and supervision, further improving performance.

4 Experiments

In this section, we start with the implementation details and evaluation settings. Then we evaluate IMU4D on 4D human motion, text description, and 3D object predictions across four datasets. Our experiments are designed to answer three key questions: (1) How accurately can IMU4D infer everyday human activities and their surrounding objects? (2) Does large-scale pre-training on a foundation model provide advantages over task-specific specialized models? (3) Does joint reasoning outperform a pipelined approach that cascades state-of-the-art modules?

4.1 Experimental Settings

Tasks and Evaluation Datasets. We evaluate three tasks: (1) **IMU-to-Motion (I2M)**, which predicts human motion from IMU signals. We use HumanML (Guo et al., 2022), DIP-IMU (Huang et al., 2018), IMUPoser (Mollyn et al., 2023) as our evaluation datasets, among which DIP-IMU and IMUPoser provide IMU signals captured in *real-world* settings. We additionally collect a real-world commodity-device IMU dataset, **IMU4D-Real**, using earbuds, smartphones, and smartwatches to evaluate performance under practical sensing conditions. We evaluate under two configurations: a 3-point (3pt) setup with IMUs mounted on one ear, one wrist, and one thigh, and a 5-point (5pt) setup with IMUs mounted on one ear, both wrists, and both thighs. We report results under 3pt setup on DIP-IMU and IMUPoser, and under 5pt setup on HumanML. (2) **IMU-to-MotionText (I2MT)**, which jointly predicts motion sequences and text descriptions from IMU inputs. We conduct our evaluation on HumanML (Guo et al., 2022), HUMOTO (Lu et al., 2025a), OMOMO (Li et al., 2023b), and HiPhi (Ji et al., 2026). Note that HumanML is a superset of AMASS (Mahmood et al., 2019)—a dataset widely adopted for IMU pretraining in prior work (Xu et al., 2024c; Mollyn et al., 2023). We use a 5pt setup for I2MT. (3) **IMU-to-MotionTextObject (I2MTO)**, which requires holistic reasoning over motion, text descriptions, and 3D objects in the interaction. We evaluate the

Table 3: **Quantitative comparison of IMU-to-MotionText using 480-frame IMUs.** We use the 5pt setup (ear, both wrists, both thighs).

Benchmark	Method	Motion Metrics					Text Metrics				
		MPJPE ↓	PA-MPJPE ↓	MPJRE ↓	MPJVE ↓	MTE ↓	ROUGE-L ↑	CIDEr ↑	BLEU@1 ↑	BLEU@4 ↑	BERT ↑
HumanML	Mod-I2MTO	43.70	34.59	11.32	56.05	275.39	36.80	133.15	32.55	4.85	26.50
	Ours	**14.69**	**11.32**	**3.06**	**19.31**	**48.50**	**43.35**	**197.96**	**38.08**	**11.00**	**35.17**
HUMOTO	Mod-I2MTO	50.08	38.77	11.59	60.38	277.02	33.21	88.27	30.17	4.43	19.43
	Ours	**28.94**	**21.80**	**7.16**	**34.47**	**86.69**	**34.96**	**116.24**	**35.87**	**7.12**	**23.00**
OMOMO	Mod-I2MTO	43.06	33.74	9.97	59.97	235.57	74.21	537.99	72.71	50.22	53.66
	Ours	**25.45**	**18.55**	**5.96**	**34.13**	**58.88**	**79.70**	**624.56**	**78.24**	**60.00**	**61.16**
HiPhi	Mod-I2MTO	43.52	32.81	11.73	58.32	228.42	36.16	115.94	34.16	6.68	20.67
	Ours	**22.53**	**16.66**	**6.87**	**29.86**	**66.59**	**40.82**	**172.55**	**40.12**	**12.45**	**23.63**

Table 4: **Real-world commodity-device results on IMU4D-real dataset.** Whole sequences, 3pt setup (left wrist, right thigh, head). MobilePoser's text is obtained by captioning its motion with MotionGPT3. N/A: no global translation; –: no text output.

	IMU-to-Motion					IMU-to-Text				
Method	MPJPE ↓	PA-MPJPE ↓	MPJRE ↓	MPJVE ↓	MTE ↓	B@1 ↑	B@4 ↑	R-L ↑	CIDEr ↑	BERT ↑
IMUPoser	94.83	66.89	13.53	121.59	N/A	–	–	–	–	–
MobilePoser (+MotionGPT3)	87.46	62.77	14.46	113.94	979.77	31.2	1.0	31.5	89.8	16.3
Ours	**37.12**	**31.30**	**7.46**	**50.18**	**302.41**	**31.5**	**3.4**	**35.5**	**100.4**	**17.2**

scene predictions on HUMOTO (Lu et al., 2025a), OMOMO (Li et al., 2023b) and HiPhi (Ji et al., 2026).

Evaluation Metrics. We evaluate three output modalities: 4D human motion, text description, and 3D objects. For motion, we follow prior work (Xu et al., 2024c; Hong et al., 2025) and report joint accuracy (MPJPE, PA-MPJPE), mesh accuracy (MPJVE), and trajectory error (MTE) for long-horizon prediction, all in millimeters, and joint rotation error (MPJRE) in degrees. For text, we adopt the linguistic metrics from MotionGPT3 (Zhu et al., 2025a) and report BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), CIDEr (Vedantam et al., 2015), and BERT-Score (Zhang et al., 2019). For 3D objects, we report 3D bounding box IoU (3D-IoU) and identity-level precision and recall (ID-P, ID-R) of predicted objects. For more details on metrics, please refer to the supplementary material.

Baselines. For IMU-to-Motion, we compare against IMUPoser (Mollyn et al., 2023) and MobilePoser (Xu et al., 2024c), both using same IMU placement as ours. As no prior work jointly predicts motion, text, and 3D layouts from IMU signals, we construct a modular baseline, *Mod-I2MTO*, combining MobilePoser for motion estimation, MotionGPT3 (Zhu et al., 2025a) for motion captioning, and Summon (Ye et al., 2022), Pose2Room (Nie et al., 2022), Obj-popup (Petrov et al., 2023), or TriDi (Petrov et al., 2025) for object-layout prediction. All baselines are retrained on same data.

4.2 Experimental Results

IMU-to-Motion. We first evaluate IMU-to-4D's human activity inference on the IMU-to-Motion task. As shown in Tab. 2, our model outperforms all baselines in motion prediction. In Tab. 4, we evaluate on IMU4D-real dataset, matching our target sensing scenarios. The results demonstrate robustness with commodity devices. Qualitative results in Fig. 3 show that our method better tracks the ground-truth motion over time; video comparisons with predicted captions are on the supplementary webpage, despite not using ground-truth text when fine-tuning on IMU4D-real.

IMU-to-MotionText and IMU-to-MotionTextObject. We evaluate the I2MT and I2MTO tasks to demonstrate the benefits of jointly reasoning across modalities. As shown in Tab. 3, our model jointly reasons about motion and text from IMU signals, achieving superior performance across all motion and text evaluation metrics compared to a modular baseline Mod-I2MTO that processes IMU

Table 6: **Quantitative comparison of scene generations in IMU-to-MotionTextObject on HUMOTO (Lu et al., 2025a).** Our method consistently outperform modular baseline, evaluated on 480-frame IMUs using the 5pt setup (ear, both wrists, both thighs).

Method	3D-IoU ↑	ID-P ↑	ID-R ↑
Mod-I2MTO (Summon)	1.65	15.22	14.14
Mod-I2MTO (Pose2Room)	4.61	8.84	16.16
Ours	**8.01**	**25.40**	**32.32**

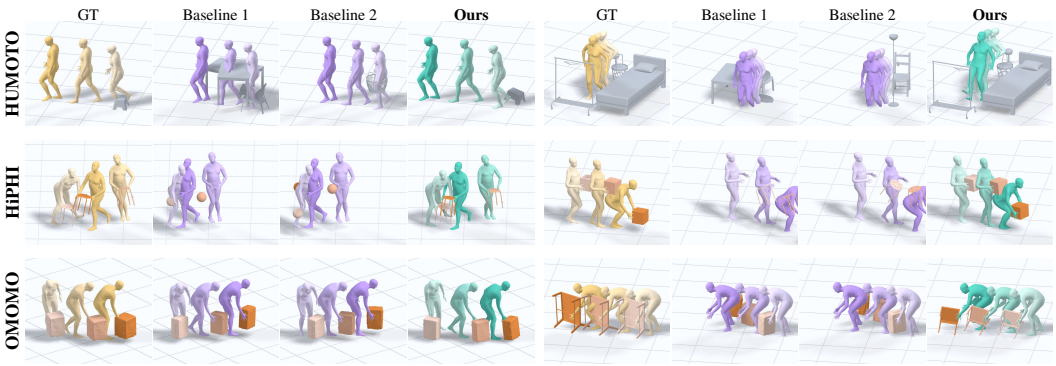

Figure 4: **Human–object recovery from IMUs.** Each image overlays three keyframes (lighter = earlier) under a shared camera. Bodies: GT (yellow), baselines (purple), Ours (green); objects: static (grey), moving (orange). Baselines lift MobilePoser motion to objects: Baseline 1 / 2 are Pose2Room / SUMMON on HUMOTO and Object Pop-up / TriDi on HiPHI and OMOMO.

Table 5: **Dynamic-object results.** Evaluated with the 5pt setup in the 480-frame setting. Obj-popup and TriDi require an object as input, thus we train a motion-to-object predictor for both. Thus, it shares same recall (%).

Benchmark	Method	Recall (%) ↑	Moving F1@0.25 (%) ↑	Velocity error (mm/s) ↓
OMOMO	Mod-I2MTO (Obj-popup)	40.42	9.73	**543.87**
	Mod-I2MTO (TriDi)	40.42	13.55	1580.35
	Ours	**48.96**	**15.44**	550.76
HiPhi	Mod-I2MTO (Obj-popup)	42.70	15.33	349.59
	Mod-I2MTO (TriDi)	42.70	16.08	1621.91
	Ours	**49.14**	**18.11**	**332.32**

signals sequentially through separate single-purpose models (MobilePoser for IMU-to-Motion and MotionGPT3 for motion-to-text). The modular pipeline suffers from error propagation, where inaccuracies in motion prediction cascade into downstream text and object predictions. In contrast, our unified framework learns all modalities within a single model, enabling cross-modal understanding through shared representations. We further evaluate the quality of object predictions produced by our I2MTO model in Tab. 6 (static objects), Tab. 5 (dynamic objects) and Fig. 4. To complete the I2MTO baseline, we use either Summon (Ye et al., 2022) or Pose2Room (Nie et al., 2022) on HUMOTO, and Obj-popup (Petrov et al., 2023) or TriDi (Petrov et al., 2025) on OMOMO and HiPhi. Our approach achieves stronger cross-modal consistency compared to baselines, demonstrating the advantage of unified multimodal learning over sequential pipelines.

Application: Re-Localization in 3D scene. When fine-tuned on sequences captured in a single fixed environment (ParaHome), our model can additionally predict the person's initial global position and heading from IMU signals alone, which by themselves carry no absolute positional information. We hypothesize that the model learns an environment-specific prior associating activities with plausible locations in that scene. This capability is scene-specific and requires in-scene training data; Please refer to Sec. C for more details.

4.3 Analysis

We summarize our key conclusions from the results as follows: **(1) Accuracy of IMU4D.** Our model improves human motion reconstruction and activity captioning over the evaluated modular baselines. For object prediction, it improves recognition recall across the evaluated datasets, while geometric and motion accuracy remain challenging and show mixed results on dynamic objects. **(2) Benefits of large pre-training.** Large pre-training provides a strong foundation model that consistently improves generalization and performance compared to task-specific models. As shown in Tab. 7, both I2M and I2MT scale effectively with increasing model capacity. **(3) Joint reasoning vs cascaded**

Table 7: **Backbone size comparison of our model on HumanML using 5pt inputs and 480 frames.** All variants are trained for 157k iterations. Scaling the base model from 355M to 1.5B parameters consistently reduces motion errors (e.g., MPJPE from 36.30 to 23.60), and the largest model also achieves the best text metrics, indicating that our framework benefits from increased model capacity.

Base model	#Params	Motion Metrics					Text Metrics				
		MPJPE ↓	PA-MPJPE ↓	MPJRE ↓	MPJVE ↓	MTE ↓	ROUGE-L ↑	CIDEr ↑	BLEU@1 ↑	BLEU@4 ↑	BERT ↑
GPT-2 medium	355M	36.30	27.15	7.13	47.16	115.20	41.11	169.25	35.32	7.68	31.76
Qwen3-0.6B	0.6B	26.52	19.40	5.52	34.62	67.49	39.88	165.54	34.01	7.49	30.34
Phi-1.5B	1.5B	**23.60**	**17.66**	**4.99**	**30.59**	**63.55**	**42.15**	**180.09**	**36.63**	**8.78**	**33.71**

Table 8: **Joint vs. single-task and cascaded formulations.** All variants use the same architecture and pretraining. I2T and I2O predict text and object directly from IMUs, while M2T and M2O predict text and scene directly from motion. We use 480-frame IMU sequences under the 5pt setup (ear, both wrists, and both thighs). Our joint formulation (I2MTO) consistently outperforms the cascaded (I2M + M2T / M2O) and single-task (I2M / I2T / I2O) approaches across all text and object metrics and most motion metrics, demonstrating that joint reasoning not only mitigates error propagation in cascaded pipelines but also improves individual task performance.

(a) IMU-to-Motion on HumanML

Method	MPJPE ↓	PA-MPJPE ↓	MPJRE ↓	MPJVE ↓	MTE ↓
I2M	16.76	12.59	3.65	21.69	**44.28**
I2MTO (Ours)	**14.69**	**11.32**	**3.06**	**19.31**	48.50

(b) IMU-to-Text on HumanML

Method	ROUGE-L ↑	CIDEr ↑	BLEU@1 ↑	BLEU@4 ↑	BERT ↑
I2T	37.28	137.20	30.67	3.98	25.79
I2M + M2T	35.28	127.70	28.95	3.37	21.23
I2MTO (Ours)	**43.35**	**197.96**	**38.08**	**11.00**	**35.17**

(c) IMU-to-Object on HUMOTO

Method	3D-IoU ↑	ID-P ↑	ID-R ↑
I2O	1.76	14.71	15.15
I2M + M2O	1.09	20.93	18.18
I2MTO (Ours)	**8.01**	**25.40**	**32.32**

approaches. Jointly reasoning over motion, text, and 3D objects consistently improves performance over single-task training. The gain is particularly large for 3D object prediction, where interaction data are limited. In Tab. 8, we compare single-task training, a cascaded pipeline, and joint training using the same architecture and pretraining. These results show that complementary modalities benefit all tasks. **(4) Challenges and Limitations.** Attention alone cannot fully prevent human–object penetration; explicit contact losses are left for future work. We recover interaction-relevant objects rather than the full 3D scene, whose reconstruction from sparse IMUs would require substantial hallucination. Object prediction is limited to a closed set of categories seen in training and does not recover object shape; absolute localization accuracy (e.g., 3D-IoU) remains low, reflecting the inherent ambiguity of inferring object placement from body motion alone.

5 Conclusion

We introduced IMU4D, a framework for 4D human–object understanding without relying on cameras. By repurposing large language models for non-visual spatiotemporal reasoning, our approach recovers human motion, activity descriptions, and interacted objects using only a few everyday wearable IMUs. Experiments across diverse real-world datasets demonstrate that IMU4D produces more coherent and temporally stable estimates than state-of-the-art cascaded pipelines. These results highlight the strong potential of wearable sensors as a privacy-preserving and scalable alternative to vision-based perception.

AI USE STATEMENT

We used Large Language Models (LLMs) in two capacities in this work.

Data preprocessing. We employed Qwen-0.6B to re-caption the existing text annotations of Motionmillions, Humoto, Omomo, and HiPhi datasets in order to align their caption styles. The model was used only to rewrite existing annotations; it did not introduce new semantic content beyond what was present in the original captions. The re-captioned annotations will be released to support reproducibility.

Writing assistance. LLMs were also used to polish the manuscript, including refining wording, correcting grammar, and improving readability.

Apart from the uses described above, LLMs were not involved in the formulation of research ideas, the design of methods and experiments, or the analysis and interpretation of results. All LLM-assisted outputs, including the re-captioned annotations and the revised text, were reviewed by the authors, who take full responsibility for the content of this paper.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. IMU4D is motivated in part by privacy: it infers human motion and interactions from wearable inertial sensors rather than cameras, and therefore never captures faces, bystanders, or the appearance of private spaces at inference time.

Human subjects. This study involved human participants but no animal experimentation. IMU4D-Real was recorded from 2 participants wearing commodity devices while being filmed by calibrated RGB cameras, which were used solely to obtain ground-truth motion. All participants consented and well-informed about the capture.

Privacy. To protect participant privacy, the raw videos will not be released. We release only IMU signals, SMPL-X parameters, and text annotations, which contain no faces or other directly identifying information.

Datasets and licenses. All other motion datasets used are publicly available and were used in compliance with their respective licenses and usage guidelines.

Potential misuse. We recognize that the same capability carries risks. Commodity earbuds, watches, and phones already record IMU streams, often without the wearer's attention, and a model like ours could be misused to track individuals covertly, infer sensitive activities or health conditions, or, through scene re-localization, estimate where a person is within their home. Deployments should require the wearer's informed consent, process data on-device or under strict access control, and be transparent about what is inferred.

We are committed to maintaining transparency and integrity throughout the research process, and will document the intended use and limitations of IMU4D alongside the released code and data.

REPRODUCIBILITY STATEMENT

Sec. 3 describes the tokenization of each modality, the model architecture, and the two-stage training procedure. Appendix D gives the training hyperparameters, hardware and compute budget, data augmentation, and inference procedure. Appendix E lists every dataset, the filtering applied to MotionMillion, the train/test splits, and the IMU4D-Real capture setup, and Appendix F defines all metrics and evaluation protocols. Baseline configurations are described in Sec. 4 and Appendix A. The model is initialized from the public Show-o (Xie et al., 2024) weights, and all training and evaluation data except IMU4D-Real are publicly available. We will release the code, model weights, and the IMU4D-Real dataset upon acceptance.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15935–15946, 2022.

Jona Braun, Sammy Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis. In *2024 International Conference on 3D Vision (3DV)*, pp. 464–473. IEEE, 2024.

Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. Motionllm: Understanding human behaviors from human motions and videos. *arXiv preprint arXiv:2405.20340*, 2024a.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Jihoon Chung, Cheng-hsin Wuu, Hsuan-ru Yang, Yu-Wing Tai, and Chi-Keung Tang. Haa500: Human-centric atomic action dataset with curated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13465–13474, 2021.

Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. Posescript: Linking 3d human poses and natural language. *IEEE transactions on pattern analysis and machine intelligence*, 2024.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

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

Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J Black. Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1323–1333, 2024.

Ke Fan, Shunlin Lu, Minyue Dai, Runyi Yu, Lixing Xiao, Zhiyang Dou, Junting Dong, Lizhuang Ma, and Jingbo Wang. Go to zero: Towards zero-shot motion generation with million-scale data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13336–13348, 2025.

Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J Black. Chatpose: Chatting about 3d human pose. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2093–2103, 2024.

Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Learning complex 3d human self-contact. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1343–1351, 2021a.

Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9919–9928, 2021b.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Computer Graphics Forum*, volume 42, pp. 1–12. Wiley Online Library, 2023.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, November 2024.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. 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*, pp. 19383–19400, 2024.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5152–5161, 2022.

Chuan Guo, Inwoo Hwang, Jian Wang, and Bing Zhou. Snapmogen: Human motion generation from expressive texts. *arXiv preprint arXiv:2507.09122*, 2025.

Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11374–11384, 2021.

Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. In *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–9, 2023.

Fangzhou Hong, Vladimir Guzov, Hyo Jin Kim, Yuting Ye, Richard Newcombe, Ziwei Liu, and Lingni Ma. Egolm: Multi-modal language model of egocentric motions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5344–5354, 2025.

Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)*, 37(6):1–15, 2018.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. π0. 5: a vision-language-action model with open-world generalization, 2025. *URL https://arxiv. org/abs/2504.16054*, 2025.

Jiahao Ji, Ji Ma, Runhan Zhang, Runyi Yu, Wenjia Wang, Weiheng Chi, Qianqian Peng, Weichao Yan, Yongfei Gu, Ye Tian, et al. Hiphi: A large-scale benchmark for high-precision human motion and object-interaction. *arXiv preprint arXiv:2608.16222*, 2026.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023.

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

Nan Jiang, Zimo He, Zi Wang, Hongjie Li, Yixin Chen, Siyuan Huang, and Yixin Zhu. Autonomous character-scene interaction synthesis from text instruction. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024a.

Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1737–1747, 2024b.

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

Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. Parahome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1816–1828, 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. Nifty: Neural object interaction fields for guided human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 947–957, 2024.

Jiye Lee and Hanbyul Joo. Locomotion-action-manipulation: Synthesizing human-scene interactions in complex 3d environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9663–9674, 2023.

Jiye Lee and Hanbyul Joo. Mocap everyone everywhere: Lightweight motion capture with smartwatches and a head-mounted camera. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1091–1100, 2024.

Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 1272–1279, 2022.

Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023a.

Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023b.

Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. In *European Conference on Computer Vision*, pp. 54–72. Springer, 2024a.

Jianan Li, Tao Huang, Qingxu Zhu, and Tien-Tsin Wong. Physics-based scene layout generation from human motion. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–10, 2024b.

Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. Genmo: A generalist model for human motion. *arXiv preprint arXiv:2505.01425*, 2025.

Ronghui Li, Junfan Zhao, Yachao Zhang, Mingyang Su, Zeping Ren, Han Zhang, Yansong Tang, and Xiu Li. Finedance: A fine-grained choreography dataset for 3d full body dance generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10234–10243, 2023c.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023d.

Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, 132(9):3463–3483, 2024.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36:25268–25280, 2023.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention. *arXiv preprint arXiv:2402.08268*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Jiaxin Lu, Chun-Hao Paul Huang, Uttaran Bhattacharya, Qixing Huang, and Yi Zhou. Humoto: A 4d dataset of mocap human object interactions. *arXiv preprint arXiv:2504.10414*, 2025a.

Shunlin Lu, Jingbo Wang, Zeyu Lu, Ling-Hao Chen, Wenxun Dai, Junting Dong, Zhiyang Dou, Bo Dai, and Ruimao Zhang. Scamo: Exploring the scaling law in autoregressive motion generation model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27872–27882, 2025b.

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

Josh Merel, Saran Tunyasuvunakool, Arun Ahuja, Yuval Tassa, Leonard Hasenclever, Vu Pham, Tom Erez, Greg Wayne, and Nicolas Heess. Catch & carry: reusable neural controllers for vision-guided whole-body tasks. *ACM Transactions on Graphics (TOG)*, 39(4):39–1, 2020.

Thomas B Moeslund and Erik Granum. A survey of computer vision-based human motion capture. *Computer vision and image understanding*, 81(3):231–268, 2001.

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

Yinyu Nie, Angela Dai, Xiaoguang Han, and Matthias Nießner. Pose2room: understanding 3d scenes from human activities. In *European Conference on Computer Vision*, pp. 425–443. Springer, 2022.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

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*, pp. 10318–10329, 2025.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10975–10985, 2019.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Object pop-up: Can we infer 3d objects and their poses from human interactions alone? In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4726–4736. IEEE, 2023.

Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Tridi: Trilateral diffusion of 3d humans, objects, and interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5523–5535, 2025.

Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 722–731, 2021.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2545–2555, 2025.

Martin Schepers, Matteo Giuberti, Giovanni Bellusci, et al. Xsens mvn: Consistent tracking of human motion using inertial sensing. *Xsens Technol*, 1(8), 2018.

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

Junyu Shi, Lijiang Liu, Yong Sun, Zhiyuan Zhang, Jinni Zhou, and Qiang Nie. Genm^3: Generative pretrained multi-path motion model for text conditional human motion generation. *arXiv preprint arXiv:2503.14919*, 2025.

Sony Corporation. mocopi - mobile motion capture system, 2025. URL `https://www.sony.net/Products/mocopi-dev/en/`. Accessed: November 2025.

Sabera Talukder, Yisong Yue, and Georgia Gkioxari. Totem: Tokenized time series embeddings for general time series analysis. *arXiv preprint arXiv:2402.16412*, 2024.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024. *URL https://arxiv. org/abs/2405.09818*, 9(8), 2024.

Qwen Team. Qwen3-vl: The most capable vision-language model in the qwen series. *arXiv preprint arXiv:2511.21631*, 2025.

Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *ISMIR*, volume 1, pp. 6, 2019.

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

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Timo Von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer graphics forum*, volume 36, pp. 349–360. Wiley Online Library, 2017.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024a.

Ye Wang, Sipeng Zheng, Bin Cao, Qianshan Wei, Weishuai Zeng, Qin Jin, and Zongqing Lu. Scaling large motion models with million-level human motions. *arXiv preprint arXiv:2410.03311*, 2024b.

Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say interact as you can: Language-guided human motion generation with scene affordance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 433–444, 2024c.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*, 2024a.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024b.

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. *arXiv preprint arXiv:2503.15451*, 2025.

Xuan Xiao, Jianjian Wang, Pingfa Feng, Ao Gong, Xiangyu Zhang, and Jianfu Zhang. Fast human motion reconstruction from sparse inertial measurement units considering the human shape. *Nature Communications*, 15(1):2423, 2024.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-x: Towards versatile human-human interaction analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22260–22271, 2024a.

Sirui Xu, Yu-Xiong Wang, Liangyan Gui, et al. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. *Advances in Neural Information Processing Systems*, 37:52858–52890, 2024b.

Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, et al. Interact: Advancing large-scale versatile 3d human-object interaction generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7048–7060, 2025.

Vasco Xu, Chenfeng Gao, Henry Hoffmann, and Karan Ahuja. Mobileposer: Real-time full-body pose estimation and 3d human translation from imus in mobile consumer devices. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–11, 2024c.

Sifan Ye, Yixing Wang, Jiaman Li, Dennis Park, C Karen Liu, Huazhe Xu, and Jiajun Wu. Scene synthesis from human motion. In *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–9, 2022.

Zhankai Ye, Bofan Li, Yukai Jin, Shuoqiu Li, Wei Wang, Yanfu Zhang, Shangqian Gao, and Xin Liu. Geomotiongpt: Geometry-aligned motion understanding with large language models. *arXiv preprint arXiv:2601.07632*, 2026.

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*, pp. 7072–7084, 2025a.

Hongwei Yi, Chun-Hao P Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, and Michael J Black. Human-aware object placement for visual environment reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3959–3970, 2022a.

Hongwei Yi, Chun-Hao P Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J Black. Mime: Human-aware 3d scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12965–12976, 2023.

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

Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13167–13178, 2022b.

Xinyu Yi, Yuxiao Zhou, and Feng Xu. Physical non-inertial poser (pnp): modeling non-inertial effects in sparse-inertial human motion capture. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Xinyu Yi, Shaohua Pan, and Feng Xu. Improving global motion estimation in sparse imu-based motion capture with physics. *ACM Transactions on Graphics (TOG)*, 44(4):1–16, 2025b.

Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17016–17027, 2023.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities, 2023.

Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Hoi-mˆ 3: Capture multiple humans and objects interaction within contextual environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 516–526, 2024a.

Libo Zhang, Xinyu Yi, and Feng Xu. Baroposer: Real-time human motion tracking from imus and barometers in everyday devices. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–9, 2025.

Mingfang Zhang, Yifei Huang, Ruicong Liu, and Yoichi Sato. Masked video and body-worn imu autoencoder for egocentric action recognition. In *European Conference on Computer Vision*, pp. 312–330. Springer, 2024b.

Mingyuan Zhang, Daisheng Jin, Chenyang Gu, Fangzhou Hong, Zhongang Cai, Jingfang Huang, Chongzhi Zhang, Xinying Guo, Lei Yang, Ying He, et al. Large motion model for unified multi-modal motion generation. In *European Conference on Computer Vision*, pp. 397–421. Springer, 2024c.

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*, pp. 180–200. Springer, 2022.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7368–7376, 2024d.

Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A Ali Heydari, Girish Narayanswamy, Max Xu, Ahmed Metwally, Jinhua Xu, Jake Garrison, Xuhai Xu, et al. Sensorlm: Learning the language of wearable sensors. *Advances in Neural Information Processing Systems*, 38:46813–46849, 2026.

Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, and Siyu Tang. Synthesizing diverse human motions in 3d indoor scenes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14738–14749, 2023.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5745–5753, 2019.

Bingfan Zhu, Biao Jiang, Sunyi Wang, Shixiang Tang, Tao Chen, Linjie Luo, Youyi Zheng, and Xin Chen. Motiongpt3: Human motion as a second modality. *arXiv preprint arXiv:2506.24086*, 2025a.

Zunjie Zhu, Yan Zhao, Yihan Hu, Guoxiang Wang, Hai Qiu, Bolun Zheng, Chenggang Yan, and Feng Xu. Progressive inertial poser: Progressive real-time kinematic chain estimation for 3-d full-body pose from three imu sensors. *IEEE Transactions on Instrumentation and Measurement*, 74:1–13, 2025b.

Chengxu Zuo, Yiming Wang, Lishuang Zhan, Shihui Guo, Xinyu Yi, Feng Xu, and Yipeng Qin. Loose inertial poser: Motion capture with imu-attached loose-wear jacket. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2209–2219, 2024.

Chengxu Zuo, Jiawei Huang, Xiao Jiang, Yuan Yao, Xiangren Shi, Rui Cao, Xinyu Yi, Feng Xu, Shihui Guo, and Yipeng Qin. Transformer imu calibrator: Dynamic on-body imu calibration for inertial motion capture. *ACM Transactions on Graphics (TOG)*, 44(4):1–14, 2025.

We provide supplementary videos, including comparisons with baselines and failure cases, in the supplementary file `index.html`. In this supplementary material, we provide additional results and implementation details. It is organized as follows:

- Sec. A provides additional baseline comparisons for the IMU-to-Motion task, as well as alternative modular pipeline configurations used as baselines for the IMU-to-MotionText task.
- Sec. B includes ablation studies on input configurations and model designs.
- Sec. C demonstrates diverse outputs through qualitative results.
- Sec. D provides implementation details for both the training and testing phases.
- Sec. E describes the datasets used for IMU-to-MotionText pretraining and IMU-to-MotionTextScene fine-tuning.
- Sec. F details the evaluations.
- Sec. G illustrates failure cases and limitations.

A More Baseline Comparisons

In this section, we provide comparisons to additional baselines of IMU-to-Motion and IMU-to-MotionText.

A.1 IMU-to-Motion

In Tab. 2 and Tab. 4 of the main paper, we evaluate our method under the 3-sensor setting (3pt) on HumanML and our self-collected data, both in the real-world scenarios. We additionally compare to ProgIP (Zhu et al., 2025b), IMUPoser and MobilePoser on LINGO in the 3pt and 5pt settings in Tab. 9.

Table 9: **Quantitative comparison of IMU-to-Motion on LINGO (Jiang et al., 2024a).** Evaluated on 480 frames. *ProgIP uses a 3-point head-and-wrists configuration. N/A: the method does not predict global translation.

Method	#IMUs	MPJPE ↓	PA-MPJPE ↓	MPJRE ↓	MPJVE ↓	MTE ↓
ProgIP	3pt*	64.89	45.66	15.52	68.77	N/A
IMUPoser	3pt	55.46	43.86	14.70	74.17	N/A
MobilePoser	3pt	59.75	49.67	18.53	78.61	208.06
Ours	3pt	**40.67**	**28.54**	**10.21**	**52.28**	**28.87**
IMUPoser	5pt	46.69	35.16	13.28	62.49	N/A
MobilePoser	5pt	49.61	39.63	17.23	64.95	195.95
Ours	5pt	**31.53**	**18.55**	**8.13**	**40.56**	**24.39**

In Tab. 10, we further evaluate our method against PNP (Yi et al., 2024) and GlobalPose (Yi et al., 2025b) under a privileged 6-sensor setting (6pt), which adds a torso-mounted IMU to the 5pt configuration and typically requires users to wear full motion-tracking suits. Our model significantly outperforms PNP and GlobalPose on MTE while achieving comparable performance on the other evaluation metrics. Note that this privileged 6pt setting is generally impractical for real-world deployment. We include them as upper-bound references.

A.2 Alternative Modular Pipeline Configurations in IMU-to-MotionText

In the main paper, we use a cascaded pipeline of MobilePoser (IMU→motion) and MotionGPT3 (motion→text) as the baseline for IMU-to-MotionText. To show the effectiveness of joint modeling over cascaded approaches, we further compare additional combinations of state-of-the-art IMU-to-motion methods (IMUPoser (Mollyn et al., 2023), MobilePoser (Xu et al., 2024c)) and motion-to-text methods (MotionGPT (Jiang et al., 2023), MotionGPT3 (Zhu et al., 2025a), GeoMotionGPT(Ye et al., 2026)) as modular pipelines.

Table 10: **Quantitative comparison of IMU-to-motion on LINGO (Jiang et al., 2024a) under a privileged 6-sensor setting (6pt).** Evaluated on 60 frames. Units are in millimeters (mm). Note that this setting includes a torso-mounted IMU, which helps reduce drift but is impractical for real-world deployment. We include them as upper-bound references and show that our model adapted to the same 6-point setting achieves the best result on MTE and remains competitive in the other evaluation metrics.

Method	#IMU Trained	MPJPE ↓	PA-MPJPE ↓	MPJRE ↓	MPJVE ↓	MTE ↓
PNP (Yi et al., 2024)	6pt	26.95	22.38	7.78	32.20	35.69
GlobalPose (Yi et al., 2025b)	6pt	24.88	**21.05**	**7.62**	**29.30**	39.88
Ours	6pt	**24.61**	23.70	8.19	36.83	**7.55**

Table 11: **Ablation on alternative IMU-to-motion and motion-to-text methods on HumanML.** Evaluated using 5pt inputs and 480 frames. All IMU-to-motion and motion-to-text baselines are trained on the same datasets as ours, and our unified model outperforms every combination of them across all text metrics.

IMU→Motion	Motion→Text	ROUGE-L ↑	CIDEr ↑	BLEU@1 ↑	BLEU@4 ↑	BERT ↑
IMUPoser	MotionGPT	28.94	86.40	24.72	2.09	13.89
IMUPoser	MotionGPT3	37.84	141.71	33.08	5.13	27.87
IMUPoser	GeoMotionGPT	27.11	76.16	22.51	1.33	11.88
MobilePoser	MotionGPT	29.69	85.26	25.21	2.39	15.56
MobilePoser	MotionGPT3	36.80	133.15	32.55	4.85	26.50
MobilePoser	GeoMotionGPT	30.68	98.89	25.99	2.90	17.05
Ours		**43.35**	**197.96**	**38.08**	**11.00**	**35.17**

As shown in Table 11, our unified model consistently outperforms all modular combinations across all text metrics on the HumanML dataset, validating the benefits of joint multimodal generation over cascaded approaches.

B Ablation Studies

B.1 LLM Initialization

We ablate the importance of initializing our model with pretrained LLM weights by comparing against training the same architecture from scratch in Tab. 12. LLM initialization consistently improves performance across all motion and text metrics. Notably, the benefits extend beyond text generation to motion reconstruction, reducing MPJPE from 26.66 to 24.62 and MTE from 75.81 to 65.88. These results suggest that the representations learned during pretraining provide a strong initialization not only for text understanding and generation, but also for learning the mapping between IMU signals and human motion.

B.2 Sensor Count and Placement

Sensor placement. We evaluate the robustness of our model to different sensor placements on HumanML in Tab. 13. Specifically, we consider all eight left/right combinations of the wrist, thigh, and ear sensors under the 3pt setup. Performance remains highly consistent across these configurations for both motion reconstruction and text generation. In particular, MPJPE varies by only ± 0.16,mm across placements, while similarly small variations are observed for the text metrics. These results demonstrate that our model is robust to which side of the body each sensor is worn, providing flexibility in sensor placement for practical deployment. The same holds with fewer sensors: moving the thigh sensor to the other side under the 2pt setup changes MPJPE by only 0.25,mm (Tab. 14).

Number of sensors. Tab. 14 shows how performance changes as sensors are removed one at a time, starting from the full 5pt setup. Motion error increases steadily (MPJPE 23.6 / 29.4 / 38.0 / $\sim$60,mm for 5/3/2/1pt), as expected from the loss of body constraints. Text generation is much less sensitive: 3pt keeps 99% of the 5pt CIDEr, and even a phone-and-watch 2pt setup keeps 91% (164.2

Table 12: **Ablation on initialization from an LLM.** Evaluated on HumanML using 5pt inputs and 480 frames, with both variants trained for 14k iterations.

Init. from LLM?	Motion Metrics					Text Metrics				
	MPJPE ↓	PA-MPJPE ↓	MPJRE ↓	MPJVE ↓	MTE ↓	ROUGE-L ↑	CIDEr ↑	BLEU@1 ↑	BLEU@4 ↑	BERT ↑
×	26.66	20.05	5.53	34.82	75.81	40.77	167.97	35.26	7.72	31.91
✓	**24.62**	**18.48**	**5.19**	**31.84**	**65.88**	**41.66**	**177.58**	**35.90**	**8.99**	**33.08**

Table 13: **Sensor-placement variations on HumanML.** Evaluated on 480 frames using the 3pt setup (one wrist, one thigh, and one ear); (L) and (R) denote the left and right side of the body. The last two rows report the mean and sample standard deviation across all eight placements. Performance is remarkably stable: the standard deviation across placements is only 0.16 mm for MPJPE and below 0.25 points for every text metric except CIDEr ($\pm$1.33), indicating that our model does not depend on which side of the body the sensors are worn.

IMU placement	Motion Metrics					Text Metrics				
	MPJPE ↓	PA-MPJPE ↓	MPJRE ↓	MPJVE ↓	MTE ↓	ROUGE-L ↑	CIDEr ↑	BLEU@1 ↑	BLEU@4 ↑	BERT ↑
(L) wrist, (L) thigh, (L) ear	29.35	24.16	5.82	38.78	**70.56**	**41.69**	178.27	36.15	8.65	32.96
(L) wrist, (R) thigh, (L) ear	29.44	24.24	5.82	38.72	71.46	41.51	175.89	35.92	8.55	32.99
(R) wrist, (L) thigh, (L) ear	**29.12**	**23.76**	**5.80**	**38.38**	70.94	41.39	175.68	35.91	8.40	32.62
(R) wrist, (R) thigh, (L) ear	29.21	23.90	5.82	38.43	71.81	41.50	177.66	36.04	8.68	32.94
(L) wrist, (L) thigh, (R) ear	29.52	24.29	5.85	39.09	71.30	41.62	177.09	35.99	8.56	32.87
(L) wrist, (R) thigh, (R) ear	29.61	24.36	5.86	39.02	72.11	41.64	**178.50**	**36.23**	**8.70**	**33.05**
(R) wrist, (L) thigh, (R) ear	29.26	23.86	5.83	38.62	71.10	41.58	177.11	36.02	8.58	32.83
(R) wrist, (R) thigh, (R) ear	29.36	24.01	5.84	38.68	72.18	41.14	174.73	35.61	8.37	32.31
All placements (average)	29.36	24.07	5.83	38.72	71.43	41.51	176.87	35.98	8.56	32.82
All placements (std.)	$\pm$0.16	$\pm$0.22	$\pm$0.02	$\pm$0.25	$\pm$0.57	$\pm$0.18	$\pm$1.33	$\pm$0.19	$\pm$0.12	$\pm$0.24

vs. 180.1), showing that a coarse sense of the activity survives when the motion itself is only partially observed. With a single sensor both tasks degrade clearly. Which sensor is kept also matters: a wrist-only input has no reading near the pelvis, so the global trajectory drifts (MTE 289,mm vs. 93,mm for thigh-only), and its captions are also worse (CIDEr 125 vs. 139). This suggests that a pocket-worn phone is the more informative single device.

B.3 Different portion of real-world data

As shown in Tab. 15, fine-tuning on only half of the real recordings already recovers most of the gain of full fine-tuning: on IMUPoser, 50% of the data closes 94% of the MPJPE gap between the zero-shot model and full fine-tuning (104.5 $\rightarrow$ 42.7 vs. 38.7 mm) and 97% of the joint-rotation gap; on DIP-IMU it closes 68% of the MPJPE gap and 89% of the MPJRE gap.

C More Qualitative Results

We provide extensive qualitative results on the supplementary webpage (`index.html`). We visualize predictions from both synthetic and real-world IMU signals, including the jointly inferred human motion, text descriptions, and interacting 3D objects. We further demonstrate generalization to out-of-domain IMU sequences and showcase diverse predictions generated from the same IMU observations. Finally, we present additional applications enabled by our unified formulation, including text-conditioned 3D object editing and re-localization of a person within a known 3D environment. These results complement our quantitative evaluations and demonstrate the versatility of our model across different inputs, outputs, and downstream applications.

Applications. The unified formulation enables several applications beyond the benchmarks (Sec. **??** and supplementary webpage): (i) *out-of-distribution object prediction* on HumanML and LINGO, which have no object annotations; (ii) *text-conditioned object editing*, where changing the object name in the predicted caption changes the predicted object while it still follows the motion; (iii) *sampling* diverse captions and object sets for the same IMU input; and (iv) *re-localization*: after fine-tuning in a fixed scene (ParaHome), the model predicts the person's global position from IMU alone.

Table 14: **Number of sensors on HumanML** (480 frames, 157k iteration checkpoint). Rows remove one sensor at a time, starting from the full 5pt setup; (L) and (R) denote the left and right side of the body. Motion error grows as sensors are removed, while text generation holds up far better: 3pt keeps 99% of the 5pt CIDEr and 2pt keeps 91%. With a single wrist sensor there is no reading near the pelvis, so the global trajectory drifts (MTE 289 mm); a single thigh sensor avoids this.

#IMU	IMU placement	Motion Metrics					Text Metrics				
		MPJPE ↓	PA-MPJPE ↓	MPJRE ↓	MPJVE ↓	MTE ↓	ROUGE-L ↑	CIDEr ↑	BLEU@1 ↑	BLEU@4 ↑	BERT ↑
5pt	(L,R) wrist, (L,R) thigh, (L) ear	23.60	17.66	4.99	30.59	63.55	42.15	180.09	36.63	8.78	33.71
3pt	(L) wrist, (L) thigh, (L) ear	29.35	24.16	5.82	38.78	70.56	41.69	178.27	36.15	8.65	32.96
2pt	(L) wrist, (L) thigh	38.01	30.57	7.50	60.04	83.01	40.35	164.17	34.57	7.18	30.80
	(L) wrist, (R) thigh	38.26	30.77	7.52	60.33	82.65	40.01	163.72	34.12	7.22	30.24
1pt	(L) wrist	60.57	49.15	10.62	84.34	289.38	33.79	124.95	28.02	3.54	20.87
	(L) thigh	59.13	52.82	10.78	93.85	92.70	36.68	139.36	30.85	4.77	24.09

Table 15: Effect of real-world fine-tuning data (480 frames, 5-point layout). 0% is the synthetic-pretrained model without real-IMU fine-tuning.

Dataset	Train data	MPJPE↓	PA-MPJPE↓	MPJRE↓	MPJVE↓
DIP-IMU	0%	81.40	73.68	17.18	111.34
	50%	55.31	49.05	6.21	70.83
	100%	43.24	33.91	4.91	52.30
IMUPoser	0%	104.46	81.20	19.99	180.73
	50%	42.66	36.56	6.14	53.75
	100%	38.71	33.13	5.68	49.75

D Implementation Details

D.1 Training

Data augmentation. During training, we apply three augmentation strategies to improve robustness. First, we randomly sample a plausible IMU sensor configuration and mask unused sensors to simulate real-world scenarios. Second, each IMU–motion sequence is randomly cropped and re-aligned to the first frame to enhance temporal robustness. Third, we randomly drop text descriptions with probability $p = 0.3$ to accommodate datasets without text annotations. Since our sequence follows the order IMU→Motion→Text→Object, masking text during training also prevents erroneous tokens from propagating to the downstream 3D object prediction.

Training details. For IMU-to-MotionText pretraining, we train on 4 NVIDIA A100 GPUs (40GB) with an effective batch size of 96 using gradient accumulation for 500k iterations, which takes approximately 5 days. For IMU-to-MotionTextObject finetuning, we train for 30k iterations on HUMOTO (Lu et al., 2025a) and 50k iterations on ParaHome (Kim et al., 2025). For both stages, we adopt the same optimizer and learning rate scheduler as Show-o (Xie et al., 2024).

Training objective. All three outputs are trained jointly with a single sequence layout $[\mathbf{x}_{\mathrm{IMU}}, \mathbf{q}_{1:N}, \mathbf{c}, \mathbf{o}]$, where $\mathbf{c}$ is the caption and $\mathbf{o}$ the object list. *Motion.* Instead of teacher-forcing ground-truth motion tokens, the motion span consists of $N = T/4$ learnable queries $\mathbf{q}_n = \mathbf{e}_q + \mathrm{PE}(n)$, one per 4-frame window, that attend causally to the IMU prefix and bidirectionally to each other, so the full motion is decoded in one parallel pass and the training and inference inputs of this span are identical. Following TOTEM-style tokenization, each window of the root translation and orientation increments (3+6 channels) is represented by a mean, a standard-deviation and a normalized-shape token, predicted with cross-entropy, while the 21 local joint rotations (6D) are regressed directly:

$$\mathcal{L}_{\mathrm{motion}} = \frac{1}{3}\left(\mathcal{L}^{\mu}_{\mathrm{CE}} + \mathcal{L}^{\sigma}_{\mathrm{CE}} + \mathcal{L}^{s}_{\mathrm{CE}}\right) + \lambda_{\theta}\,\mathrm{SmoothL1}_{\beta=0.02}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}). \tag{1}$$

Text and scene. The caption and the object category tokens are generated autoregressively and trained with teacher forcing ($\mathcal{L}_{\mathrm{text}}$, $\mathcal{L}_{\mathrm{cat}}$). Since they attend to the motion-query hidden states rather



Figure 5: IMU-to-Motion and re-localization on ParaHome (Kim et al., 2025). White: ground truth; blue: our prediction. IMU signals carry no absolute position, and integrating relative motion alone accumulates drift. After fine-tuning on sequences from this fixed scene, our model predicts the person's global position and orientation from IMU input alone, keeping the predicted trajectory aligned with the ground truth over time.

than to ground-truth motion, the language and scene heads are conditioned on the model's own motion estimate both at training and at test time. For each object, the model predicts its first-frame pose, supervised with $\mathcal{L}_{\text{anc}}$, and its asset identity, supervised with $\mathcal{L}_{\text{id}}$; the ground-truth pose is embedded and fed as the next input token (the prediction is fed at inference). A temporal track decoder reads the whole motion-query sequence together with the root trajectory (ground truth during training, the decoded one at inference) and predicts per-frame object poses in the body frame, supervised on moving frames by $\mathcal{L}_{\text{track}} = \lambda_{\text{tr}}(\mathcal{L}_{\text{pos}} + 2\mathcal{L}_{\text{vel}} + 2\mathcal{L}_{\text{acc}})$, and a per-frame moving/static state by a binary cross-entropy $\mathcal{L}_{\text{state}}$. Object losses are averaged over the objects that carry supervision. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{motion}} + \lambda_{\text{text}}\mathcal{L}_{\text{text}} + \mathcal{L}_{\text{cat}} + \lambda_{\text{anc}}\mathcal{L}_{\text{anc}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \mathcal{L}_{\text{track}} + \lambda_{\text{st}}\mathcal{L}_{\text{state}}, \tag{2}$$

with $\lambda_\theta = 15.6$, $\lambda_{\text{text}} = \lambda_{\text{anc}} = \lambda_{\text{id}} = 1$, $\lambda_{\text{tr}} = 50$ and $\lambda_{\text{st}} = 5$. We first train only the modality adapters and output heads for 1k steps with the backbone frozen, and then train all parameters for 100k steps with AdamW (learning rate 10^{-4}, 1k warm-up steps, constant afterwards) on randomly sampled sensor layouts and crop lengths of 0.8–16 s; we report the EMA weights (decay 0.999).

D.2 Inference

Sampling stop condition. Given an IMU input sequence of length T, the model outputs human motion of the same length T. For text and object generation, decoding proceeds autoregressively until the model emits the special end tokens <EOT> and <EOOBJ>, where <EOT> and <EOOBJ> are the stop tokens of text and object respectively.

Shifted window with pose averaging. Given an IMU input of length T, we perform two overlapping inference passes: the first produces prediction m_1 over the full sequence, and the second produces a shifted prediction m_2 by offsetting the input by two frames. The final motion is obtained by averaging over the overlapping region as $m_{\text{final}} = (m_1 + m_2)/2$, which yields smoother and more accurate results in local coordinates.

E Dataset Details

E.1 Data splits.

All baselines and our model are trained on the same training split and evaluated on the same test clips. For the motion corpus we use the official MotionMillion split; no sample appears in more than one split. LINGO is split 80/10/10, and HiPHI, OMOMO, and HUMOTO are split separately from the motion corpus. Because MotionMillion assigns mirrored copies to splits independently, some HumanML3D test clips have a mirrored counterpart in the training split. This affects all methods equally; LINGO and the HOI test sets are free of such overlap.

E.2 Datasets for IMU-to-MotionText Pretraining.

We use MotionMillion (Fan et al., 2025) as our primary data source for IMU-to-MotionText pretraining. This dataset aggregates large-scale motion sequences reconstructed from web videos

and existing motion datasets. To ensure data quality, we remove datasets with irregular joint angles, heavy drift, jittering, or self-collisions. After filtering, we retain the following high-quality datasets: BABEL (Punnakkal et al., 2021), PhantomDance (Li et al., 2022), AIST (Tsuchida et al., 2019), FineDance (Li et al., 2023c), Fit3D (Fieraru et al., 2021b), HI4D (Yin et al., 2023), HumanSC3D (Fieraru et al., 2021a), InterHuman (Liang et al., 2024), Inter-X (Xu et al., 2024a), TRUMANS (Jiang et al., 2024b), EgoBody (Zhang et al., 2022), HAA500 (Chung et al., 2021), HumanML (Guo et al., 2022), and Motion-X (Lin et al., 2023). The filtered dataset contains approximately 800k sequences. We additionally incorporate LINGO (Jiang et al., 2024a), which provides 20k high-quality motion-captured sequences. We follow the original train/val/test split for MotionMillion and apply an 80/10/10 split for LINGO.

E.3 Datasets of IMU-to-MotionTextObject finetuning

We adopt two human-object interaction datasets for IMU-to-MotionTextObject finetuning.

HUMOTO (Lu et al., 2025a) is a high-fidelity human-object interaction dataset with 735 sequences (7,875 seconds at 30 fps) capturing interactions with 63 objects and 72 articulated parts. We use 685 sequences for training and 50 for testing. We use the first-frame 3D bounding box annotations of 63 object classes as ground truth, excluding articulated parts. Since HUMOTO provides motion in MIXAMO format, we apply motion retargeting to recover SMPL-X parameters.

HiPhi (Ji et al., 2026) is a large-scale, high-precision human motion and object-interaction dataset captured with an optical motion capture system at 90 fps from 132 performers. It contains 617.5 hours of motion (including left-right mirrored sequences), comprising 371.8 hours of body-only motion and 245.7 hours of human-object interaction with synchronized object trajectories and meshes, covering 40 real-world objects from 12 categories. Since HiPHI provides motion in BVH format, we apply motion retargeting to recover SMPL-X parameters.

OMOMO (Li et al., 2023b) is a full-body human-object interaction dataset of approximately 10 hours, capturing 17 subjects manipulating 15 large everyday objects such as chairs, tables, boxes, a mop, and a vacuum. Human and object motions are recorded simultaneously with a 12-camera Vicon optical motion capture system at 120 fps, and object geometry is reconstructed from monocular videos. Since OMOMO provides SMPL-X parameters, no additional retargeting is required.

ParaHome (Kim et al., 2025) comprises 207 high-quality SMPL-X motion sequences (486 minutes total) from 38 participants, captured with 70 synchronized RGB cameras and wearable IMU-based motion capture suits. The dataset captures full-body and hand motion with multiple objects in home environments, paired with text descriptions. We use 166 sequences for training and 41 for testing. A key characteristic of ParaHome is that the major scene layout remains consistent across sequences while human activities and interactions vary. This property makes it particularly suitable for our 3D scene re-localization task, where the goal is to situate predicted motions within a fixed environment. Therefore, we do not utilize the 3D object annotations and instead focus on localizing human motion relative to the scene layout.

E.4 Datasets For Real-world IMU-to-motiontext FINETUNING

IMU4D-real As shown in 6, we wear the devices on the body, including AirPods on both ears (indicating head), Apple Watches on the left wrist (left wrist), and iPhones in the right front pocket (right thigh) for all captured sequences. We use the SenorLogger app for IMU logging and synchronization. Before the motions begin, we perform a T-pose calibration for 2-3 seconds to align the IMU global frames with the coordinate conventions used during training. To obtain the ground-truth motions for training, we set up multiple RGB camera with calibrations, we predict the body motion of each frame from each camera using GVHMR Shen et al. (2024), and perform post-processing to merge predictions from multiple views to get a stable GT motions. And we annotate the test sequences from the synchronized RGB videos with Qwen3-VL-32B (Team, 2025) into one HumanML3D-style reference sentence.



Figure 6: IMU collection device.

F Quantitative Evaluation Details

F.1 Metrics Definitions

Motion metrics. We evaluate motion quality using five metrics adopted from MobilePoser (Xu et al., 2024c). MPJPE (Mean Per Joint Position Error) measures the average Euclidean distance between predicted and ground truth joint positions. PA-MPJPE (Procrustes-Aligned MPJPE) first aligns the predicted joints to the ground truth using rigid transformation then computes MPJPE. MPJRE (Mean Per Joint Rotation Error) computes the angular differences between predicted and ground truth joint rotations. MPJVE (Mean Per Joint Vertices Error) measures the difference in human meshes between predictions and ground truth. MTE (Mean Trajectory Error) evaluates the accuracy of the root joint's global trajectory.

Text metrics. Following MotionGPT3 (Zhu et al., 2025a), we adopt five widely-used linguistic metrics to evaluate text description quality. BERTScore (Zhang et al., 2019) computes semantic similarity between predicted and reference texts using contextual embeddings from pre-trained BERT (Devlin et al., 2019) models. BLEU@1 and BLEU@4 (Papineni et al., 2002) measure n-gram precision with brevity penalty, where BLEU@1 focuses on unigram overlap (word-level accuracy) and BLEU@4 captures longer phrase structures up to 4-grams. ROUGE-L (Lin, 2004) computes the longest common subsequence between predicted and reference texts, emphasizing sequential word order and sentence structure. CIDEr (Vedantam et al., 2015) weighs n-gram matches by their informativeness using TF-IDF, prioritizing distinctive and descriptive terms over common words.

We clarify the negative BERTScore (-6.07) reported for IMUPoser (Mollyn et al., 2023) + MotionGPT (Jiang et al., 2023) in Tab. 11. IMUPoser predicts only local joint rotations without global translation; filling in the missing translation channels with zeros corrupts the 263-dimensional HumanML representation (Guo et al., 2022) that both MotionGPT and MotionGPT3 require as input. MotionGPT is more severely affected because its VQ-VAE quantizes motion into discrete tokens, so even small input corruptions can map to entirely wrong codebook entries. MotionGPT3, by contrast, encodes motion into a continuous latent space, which degrades much less under partial input corruption.

Static object metrics. For 3D object prediction, we evaluate both geometric accuracy and object recognition. We report 3D bounding box IoU (3D-IoU), which measures the volumetric overlap between predicted and ground-truth object bounding boxes. To evaluate object category recognition independently of localization quality, we additionally report identity-level precision and recall (ID-P, ID-R), which measure whether the predicted object classes match the ground-truth classes regardless of spatial accuracy. Together, these metrics capture both how well the model identifies which objects are present and how accurately it localizes them in 3D space.

Dynamic object metrics. We evaluate dynamic objects on OMOMO and HiPHI. Predicted objects are matched one-to-one to ground-truth objects of the same category by Hungarian matching on

first-frame oriented 3D box IoU. A ground-truth object is dynamic if it moves in at least one frame. A frame counts as moving when the object's linear speed exceeds 0.03 m/s or its angular speed exceeds 10°/s, smoothed with a 5-frame majority filter. Recall is the fraction of dynamic ground-truth objects that are matched. The other metrics are computed only on matched pairs. Pose error is computed as the mean L2 distance between corresponding vertices of the predicted and ground-truth object meshes after applying their predicted and ground-truth poses, and is reported separately over moving and static frames. Velocity error (mm/s) is the mean L2 difference between predicted and ground-truth per-point mesh velocities, computed by finite differences at 30 fps over consecutive moving frames. Each metric is averaged over the frames of a track and then over tracks.

F.2 Evaluation Details

For motion evaluation, we use the both 60 frames and 480 frames of each sequence across all benchmarks. For text evaluation, when multiple ground-truth captions are available for a single motion sequence, we compute the score between the predicted text and each reference caption individually, and take the maximum score as the final metric. For 3D object evaluation, we manually map each fitted object from SUMMON (Ye et al., 2022) to the corresponding object category in our class taxonomy. Both the baseline and our model are finetuned separately on each dataset before evaluation on the respective test set.

G Failure Cases and Limitations

In this section, we discuss the failure cases of our approach.

Drifting. Human motion may drift over time due to the accumulation of prediction errors. This issue is shared by all methods. We hope that incorporating loop-closure mechanisms could alleviate this problem. We leave it for future work.

Interpenetration. This occurs because our model does not explicitly enforce penetration constraints. Incorporating such penalties could help alleviate this issue.