HSI-AGENT: BRIDGING MOTION DYNAMICS AND SCENE SEMANTICS FOR UNIFIED HUMAN-SCENE INTERACTION GENERATION AND UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Human-scene interaction (HSI) generation and understanding rely on the same knowledge of human motion, scene geometry, and object interactions. A unified model can share this knowledge and support richer interaction than task-specific motion generators. However, current approaches still face following limitations. First, generation and understanding remain separate prediction tasks, even when both are included in one model. They are not explicitly connected through scene-grounded plans and persistent human-scene states. Second, a shared representation of complete motion in world coordinates mixes local body dynamics with global scene placement. This representation causes large encoding errors and makes it difficult to use motion priors learned from diverse motion data. We propose **HSI-Agent**, a unified framework for HSI planning, generation, and understanding. A language planner and a scene planner convert open-ended instructions into ordered interaction plans and maintain the current human-scene state. We then introduce a factorized representation based on root-origin anchoring. Motion Tokens encode root-relative body dynamics from pretrained motion models, while Affordance Tokens encode target regions and scene constraints. These tokens provide a shared interface across the framework. For generation, Trajectory-Flow predicts global root trajectories that place motion priors in the scene. For understanding, separate streams recover motion and interaction semantics from the shared token encoders. Joint training further allows generation supervision to improve HSI understanding. Experiments on HUMANISE and augmented HumanML3D show state-of-the-art performance in HSI generation and understanding. Evaluation on LINGO and TRUMANS further demonstrates improved physical generalization across datasets.

1 INTRODUCTION

Human motion generation has progressed from low-level controls and action labels (Holden et al., 2016; Petrovich et al., 2021) to natural-language instructions (Petrovich et al., 2022; Guo et al., 2022b; Zhang et al., 2024b). Large motion-language datasets (Plappert et al., 2016; Guo et al., 2022a) and recent diffusion, autoregressive, and flow-matching models (Tevet et al., 2023; Zhang et al., 2024a;b) now provide diverse motion priors. These advances support applications in animation (Lee et al., 2019; Shlizerman et al., 2018; Tang et al., 2018), virtual reality (Aliakbarian et al., 2023; 2022; Jiang et al., 2022), robotics (Ze et al., 2024), and embodied intelligence (Wang et al., 2021; Cao et al., 2020).

Human-scene interaction (HSI) adds a physical environment to this problem. A model must generate the requested action, select the correct object, and place the body at a valid location (Wang et al., 2022). HSI understanding asks the inverse question: given a motion and its scene, what action took place and which object was involved? The two tasks depend on the same motion dynamics, spatial relations, and contact cues. Their connection is useful for a system that must both act in a scene and interpret the current interaction. Similar links between generation and understanding have proved useful in other multimodal settings (Xie et al., 2025; Wu et al., 2025a).

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

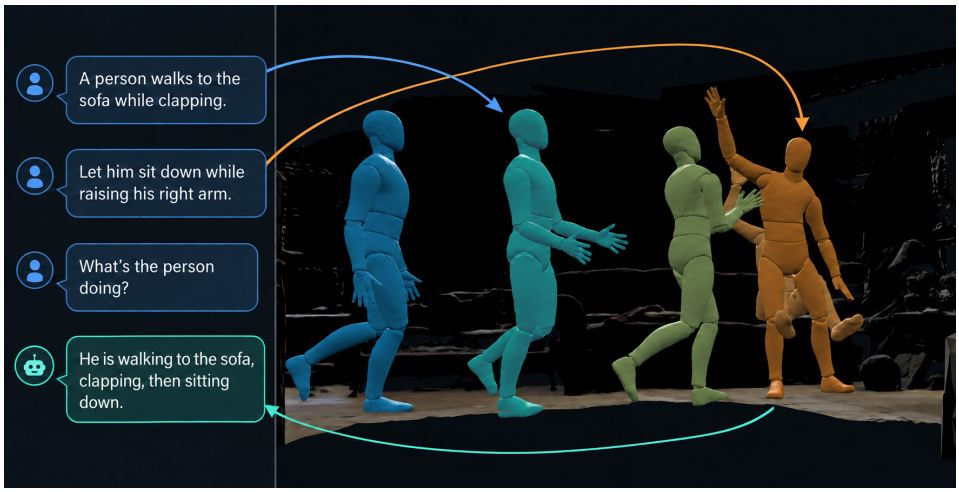


Figure 1: Overview of conversational human-scene interaction. The agent receives sequential natural-language instructions, generates the corresponding scene-grounded motion, and answers a follow-up question about the person's current action.

HSI research has moved toward this goal in stages. HUMANISE introduced paired language, motion, and scene data (Wang et al., 2022). GHOST improved target grounding (Milacski et al., 2025). MoST separated target selection, trajectory generation, and local motion (Cen et al., 2024), while Afford-Motion used scene affordances as an intermediate condition (Wang et al., 2024). These methods improve generation but do not connect it to interaction understanding. HSI-GPT takes the next step and places generation and understanding in a common scene-motion-language model (Wang et al., 2025). This shared prediction interface is flexible, but two questions remain. How should a system organize planning, execution, and interpretation across successive instructions? What representation can serve these abilities without mixing intrinsic body dynamics with arbitrary global placement?

We address both questions with **HSI-Agent**. We use the term agent in a specific operational sense. HSI-Agent receives user instructions, maintains a human-scene state, produces scene-grounded plans, and invokes generation or understanding modules to execute or interpret an interaction. It does not claim autonomous exploration or closed-loop control in the physical world. This definition places planning and state at the framework level, while the learned modules remain responsible for motion synthesis and interpretation.

The framework uses a common factorization throughout. A Language Planner and a Scene Planner translate an open-ended request into ordered actions, grounded objects, and spatial conditions. Root-origin anchoring separates body dynamics from global placement. Motion Tokens then encode root-relative motion from pretrained models, while Affordance Tokens encode target regions and scene constraints. For generation, Trajectory-Flow predicts the global root trajectory that places the motion in the scene. For understanding, separate streams recover motion semantics and scene-interaction semantics from the same token encoders. Thus, the planner organizes the interaction, and the factorized tokens provide a shared interface for execution and interpretation.

Our contributions are as follows: (i) We introduce a unified HSI framework that connects planning, generation, and understanding through persistent human-scene states. Compared with a common prediction interface, this framework organizes the three abilities as parts of the same interaction process. (ii) We develop a scene-aware planner that converts open-ended instructions into ordered, object-grounded interaction plans. It supplies the semantic and spatial conditions needed to generate and compose multiple motion segments. (iii) We introduce a factorized HSI representation based on root-origin anchoring, Motion Tokens, and Affordance Tokens. Trajectory-Flow combines pretrained root-relative motion with scene-conditioned global trajectories, while the understanding module reads the same factors through separate semantic streams. (iv) Experiments on HUMANISE, augmented HumanML3D, LINGO, and TRUMANS validate the framework across HSI generation, understanding, motion composition, and cross-dataset physical generalization.

2 Related Work

Human-Scene Interaction. Human-scene interaction modeling incorporates the geometry and semantics of the surrounding environment into motion synthesis. Early work mainly focused on human placement, learning where and in what poses humans should interact with objects in the scene. For example, POSA (Hassan et al., 2021b) models spatial and contact relations between the body and the scene to determine plausible body positions and poses. SAMP (Hassan et al., 2021a) and Wang et al. (2021) further study continuous motion generation conditioned on the scene, extending the focus from human placement to motion trajectories and body movements.

Language-conditioned HSI methods further use text to specify the intended interaction. HUMANISE (Wang et al., 2022) provides paired language, motion, and scanned-scene data. GHOST (Milacski et al., 2025) improves open-vocabulary target grounding, MoST (Cen et al., 2024) decomposes interaction targets, trajectories, and motions, and Afford-Motion (Wang et al., 2024) uses body-part-aware scene affordances to guide generation. TRUMANS (Jiang et al., 2024b) and LINGO (Jiang et al., 2024a) broaden the scale and diversity of captured HSI, while HSI-GPT (Wang et al., 2025) studies general-purpose scene-motion-language modeling for both generation and understanding.

Most existing approaches still learn task-specific scene-conditioned representations and optimize generation and understanding separately. We instead use Motion and Affordance Tokens as a shared interface between the two tasks. Trajectory-Flow adapts pretrained articulated motion to global scene constraints, while factorized understanding separates intrinsic motion semantics from scene-dependent interaction semantics.

Motion Agents and Embodied Interaction Planning. LLM-based agents use language models to interpret user instructions, plan tasks, and select tools. ReAct (Yao et al., 2023) combines reasoning with actions, allowing agents to gather information from external tools and environments and update their plans based on the results. This connects language-based planning with task execution, enabling agents to handle requests that require multiple steps and adjust their actions based on feedback.

In human motion modeling, AvatarGPT (Zhou et al., 2024) uses language as a common interface for motion understanding, planning, and generation. Motion-Agent (Wu et al., 2025b) combines MotionLLM with an external LLM to support conversational motion generation, editing, and understanding. CoMA (Sun et al., 2026) uses multiple agents for task planning and motion generation.

However, existing motion agents mainly operate on scene-agnostic motions and do not explicitly reason about target objects, 3D spatial constraints, or persistent human-scene interaction states. HSI-Agent extends this agent framework to scene-grounded interaction: a Language Planner works with a SpatialLM-based Scene Planner (Mao et al., 2025) to create executable HSI plans, which are then executed and interpreted by generation and understanding branches sharing Motion and Affordance Tokens.

3 Method

3.1 Overview

As shown in Figure 2, HSI-Agent connects planning, generation, and understanding through shared motion and affordance representations.

The generation and scene-interaction understanding branches share **one Motion Encoder and one Affordance Encoder**. These encoders produce Motion Tokens $\mathbf{Z}_m$ and Affordance Tokens $\mathbf{Z}_a$ from motion, scene geometry, and contact cues $\mathbf{C}$. During joint fine-tuning, generation and captioning losses update the shared encoders, while the task-specific modules remain fixed. This allows generation supervision to improve the representations used for understanding. The shared tokens and training objective are defined in Section 3.4. Background on pretrained motion priors and scene affordance is provided in Appendix A.

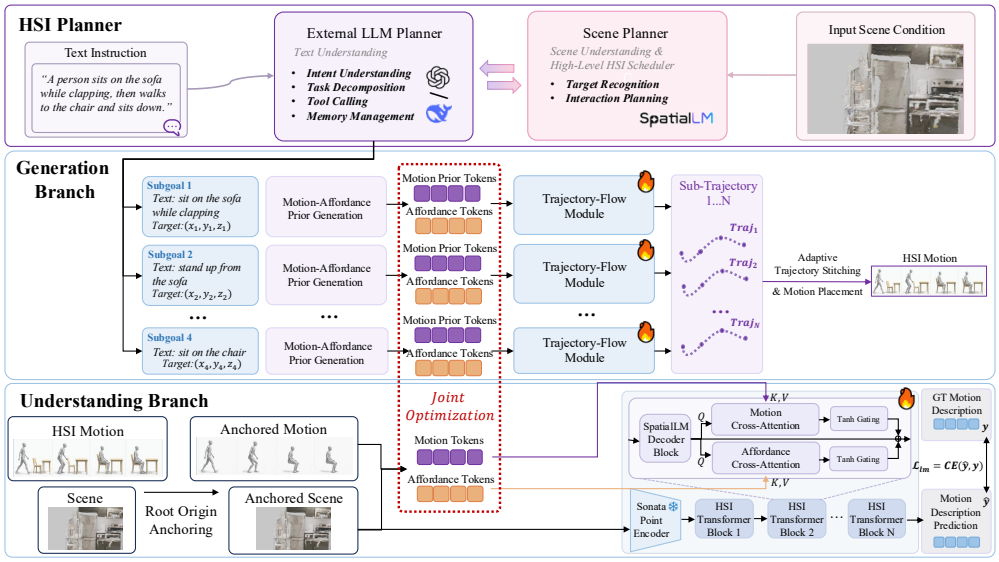


Figure 2: **Overview of HSI-Agent. i) Scene-aware planning.** The Language Planner and Scene Planner use the instruction, scene, and planning context to produce ordered interactions with target objects and spatial constraints. **ii) Factorized HSI generation.** Trajectory-Flow uses Motion and Affordance Tokens to predict global root trajectories. These trajectories are combined with pretrained root-relative motions to generate scene-grounded interactions. **iii) HSI understanding.** Two streams extract motion semantics and scene-interaction semantics, which the Planner combines into a description. Generation and scene-interaction understanding share **one Motion Encoder and one Affordance Encoder**. Both training objectives update these shared encoders.

3.2 Scene-Aware Planning and State Tracking

Our interaction planner consists of two collaborative components: a *Language Planner* implemented with an external large language model and a *Scene Planner* implemented with SpatialLM. Given a natural-language instruction $\mathbf{Q}$ and a 3D scene $\mathbf{S}$, they jointly construct a high-level interaction plan $\mathbf{\Pi} = \{\boldsymbol{\pi}_i\}_{i=1}^{N}$, with $\boldsymbol{\pi}_i = (\mathbf{a}_i, e_i, \mathbf{c}_i)$, where $\mathbf{a}_i$ describes the i-th interaction, e_i denotes its scene-grounded target instance, and $\mathbf{c}_i$ contains the associated spatial, temporal, and contextual conditions.

The two planners operate collaboratively at each planning step:

$$\mathbf{a}_i = \Phi_{\text{language}}(\mathbf{Q}, \mathbf{h}_i), \quad (e_i, \mathbf{c}_i) = \Phi_{\text{scene}}(\mathbf{S}, \mathbf{a}_i), \quad \mathbf{h}_{i+1} = \text{Update}(\mathbf{h}_i, \boldsymbol{\pi}_i), \tag{1}$$

where $\mathbf{h}_i$ denotes the planning context inherited from previous interactions. The Language Planner interprets the action semantics, target objects, temporal relations, and contextual dependencies in the instruction, and organizes them into an ordered sequence of interaction intents. Conditioned on each intent, the Scene Planner analyzes the point cloud, recognizes the relevant object instances, and extracts their spatial properties. The grounded result is then incorporated into the planning context before the next interaction is constructed. In this way, the Language Planner provides semantic guidance for scene recognition, while the Scene Planner supplies object-level grounding and spatial constraints. Their collaboration determines what interactions should be performed, which scene objects they should involve, where they should occur, and how they should be temporally organized, producing executable high-level conditions for the downstream generation module. The role of the Planner in long-horizon generation is detailed in Appendix G. Videos of generated long-sequence HSI are provided in the supplementary material.

3.3 Factorized Scene-Grounded Motion Generation

Given a scene point cloud $\mathbf{S}$ and an interaction plan $\Pi = \{\pi_i\}_{i=1}^{K}$ produced by the high-level planner, the generation module realizes each plan unit as a scene-grounded motion segment. As shown in Figure 3, the module combines pretrained motion priors with predicted global root trajectories.

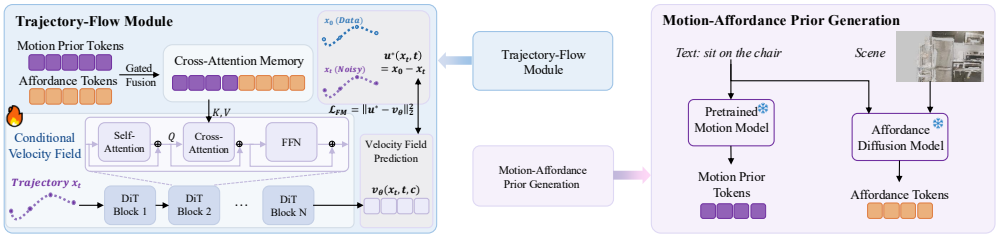


Figure 3: **Motion-Affordance Prior Generation and Trajectory-Flow.** Motion Tokens encode pretrained motion dynamics, while Affordance Tokens encode scene interaction constraints. Trajectory-Flow uses these tokens to predict global root trajectories, which are combined with root-relative motion priors to generate scene-grounded motion.

Motion prior and interaction conditions. Directly learning scene-grounded full-body motion requires a model to simultaneously capture articulated motion dynamics and global interaction trajectories. HUMANISE contains only four major interaction categories - walking, sitting, standing up, and lying down - providing limited supervision for diverse body-motion patterns (Wang et al., 2022).

Moreover, directly encoding complete body motion in world coordinates entangles local articulation with global scene placement. Our analysis in Appendix E shows that this representation substantially increases both local-pose and trajectory errors across different motion codecs. We therefore employ pretrained motion generators, such as HY-Motion (Wen et al., 2025) and MotionGPT3 (Zhu et al., 2026), to provide a diverse motion prior:

$$\widetilde{\mathbf{X}}_i = M_{\text{pre}}(\mathbf{y}_i), \quad \mathbf{Z}_{m,i} = E_m(\widetilde{\mathbf{X}}_i; \theta_m), \quad \mathbf{Z}_{a,i} = E_a(\mathbf{S}, \mathbf{C}_i; \theta_a). \tag{2}$$

where $\mathbf{y}_i$ is the textual condition derived from plan unit π_i, and $\widetilde{\mathbf{X}}_i \in \mathbb{R}^{T_i \times J \times 3}$ is the corresponding base motion. $\mathbf{C}_i$ denotes the interaction condition inferred from the scene and the current plan unit.

Motion Tokens $\mathbf{Z}_{m,i}$ preserve motion patterns learned from large-scale motion data, while Affordance Tokens $\mathbf{Z}_{a,i}$ encode scene geometry and interaction constraints. Trajectory-Flow places these scene-independent motion priors in the scene by predicting global root trajectories. This design preserves diverse motion patterns without learning them solely from limited HSI data.

Trajectory-Flow. To bridge the scene-agnostic motion prior and a specific 3D scene, we introduce Trajectory-Flow. Instead of regenerating or directly deforming the complete joint sequence, Trajectory-Flow retains its root-relative articulated motion and predicts only a scene-compatible global root trajectory:

$$\widehat{\boldsymbol{\tau}}_i = \mathcal{T}_\theta\left(\mathbf{Z}_{m,i}, \mathbf{Z}_{a,i}, \mathbf{e}_i\right), \qquad \boldsymbol{\tau}_i = \{\mathbf{p}_{i,t}, \boldsymbol{\rho}_{i,t}\}_{t=1}^{T_i}, \tag{3}$$

where $\mathbf{p}_{i,t} \in \mathbb{R}^3$ denotes the global root translation, $\boldsymbol{\rho}_{i,t} \in \mathbb{R}^6$ is a continuous root-rotation representation (Zhou et al., 2019), and $\mathbf{e}_i$ is the text embedding of the current plan unit.

We formulate $\mathcal{T}_\theta$ as a conditional flow-matching model (Lipman et al., 2023). Given a noise trajectory $\boldsymbol{\tau}_i^0$ and its corresponding ground-truth trajectory $\boldsymbol{\tau}_i^*$, we construct a linear probability path and regress its conditional velocity:

$$\mathcal{L}_{\text{TF}} = \mathbb{E}\Big[\|\mathbf{v}_\theta(\boldsymbol{\tau}_i^s, s \mid \mathbf{Z}_{m,i}, \mathbf{Z}_{a,i}, \mathbf{e}_i) - \mathbf{u}_i^s\|_2^2\Big]. \tag{4}$$

where $s \in [0, 1]$ denotes the flow time, $\boldsymbol{\tau}_i^s$ is the interpolated trajectory, and $\mathbf{u}_i^s$ is its target velocity. The interpolation path is defined in Appendix A.3. During inference, we sample an initial noise trajectory and integrate the learned velocity field to obtain $\widehat{\boldsymbol{\tau}}_i$.

The two token types provide complementary conditions for Trajectory-Flow. The motion tokens constrain the generated trajectory to remain compatible with the temporal dynamics of the base motion, while the affordance tokens guide its global translation and orientation according to the scene geometry. Trajectory-Flow therefore determines where and how the motion is executed in the scene without requiring the HSI model to relearn the full space of articulated human motion.

Scene-grounded motion reconstruction. After obtaining the root trajectory, we extract the root-relative pose from the base motion and place it along the predicted trajectory:

$$\widetilde{\mathbf{Q}}_{i,t} = \widetilde{\mathbf{R}}_{i,t}^{\mathsf{T}}(\widetilde{\mathbf{X}}_{i,t} - \widetilde{\mathbf{p}}_{i,t}), \qquad \widehat{\mathbf{X}}_{i,t} = \widehat{\mathbf{R}}_{i,t}\widetilde{\mathbf{Q}}_{i,t} + \widehat{\mathbf{p}}_{i,t}, \tag{5}$$

where $\widetilde{\mathbf{p}}_{i,t}$ and $\widetilde{\mathbf{R}}_{i,t}$ are the original root translation and rotation of the base motion, while $\widehat{\mathbf{p}}_{i,t}$ and $\widehat{\mathbf{R}}_{i,t}$ are generated by Trajectory-Flow. This reconstruction preserves the local articulated dynamics of the pretrained motion prior while adapting its global position and orientation to the input scene.

Long-sequence composition. For instructions containing multiple plan units, independently generated segments may exhibit discontinuities in both local body motion and global placement. Before applying final scene grounding, we encode the base motions into discrete VQ-VAE tokens (van den Oord et al., 2017), concatenate them according to the planned execution order, and jointly decode the resulting sequence: $\mathbf{q}_i = Q_{\mathrm{vq}}(E_{\mathrm{vq}}(\Phi(\widetilde{\mathbf{X}}_i)))$ and $\overline{\mathbf{F}} = D_{\mathrm{vq}}(\mathbf{q}_1 \| \cdots \| \mathbf{q}_K)$. Joint decoding provides temporal context across segment boundaries and regularizes local motion transitions.

Trajectory-Flow then generates a scene-grounded root trajectory for each refined segment. The trajectories are composed sequentially by using the terminal root state of one segment as the initial state of the next. Therefore, the VQ-VAE maintains continuity of the articulated body motion, while Trajectory-Flow and root alignment maintain global spatial consistency. The final sequence is $\widehat{\mathbf{X}} = \widehat{\mathbf{X}}_1 \oplus \cdots \oplus \widehat{\mathbf{X}}_K$, where $\oplus$ denotes temporal composition with boundary refinement. Additional details are provided in Appendix F.

3.4 Shared HSI Understanding and Joint Learning

HSI understanding requires both intrinsic motion semantics and scene-dependent interaction semantics. Directly predicting a description from the complete scene-grounded motion entangles articulated dynamics with object identity and spatial placement. Under the limited motion diversity of existing HSI datasets, this coupling can lead to action-object shortcuts and weaken compositional generalization. We therefore factorize understanding into two complementary description streams:

$$\widehat{\mathbf{Y}}_m = U_m(\mathcal{T}_0(\mathbf{X})), \quad \widehat{\mathbf{Y}}_s = U_s(\mathbf{X}, \mathbf{S}; \mathbf{Z}_m, \mathbf{Z}_a), \quad \widehat{\mathbf{Y}} = \mathcal{P}_{\mathrm{fuse}}(\widehat{\mathbf{Y}}_m, \widehat{\mathbf{Y}}_s). \tag{6}$$

Dual-stream description. The motion-semantic stream first canonicalizes the observed motion with $\mathcal{T}_0$ by removing its initial global translation while preserving its articulated poses and relative root trajectory. Without access to the scene, U_m focuses on intrinsic action dynamics and can benefit from motion-centric datasets with greater motion diversity. In parallel, the scene-interaction stream preserves the alignment between the motion $\mathbf{X}$ and scene $\mathbf{S}$. It uses a pretrained Sonata point encoder (Wu et al., 2025c) together with SpatialLM to encode the scene and introduces the shared Motion Tokens $\mathbf{Z}_m$ and Affordance Tokens $\mathbf{Z}_a$ through separate cross-attention bridges, enabling U_s to identify the interacted object and the corresponding spatial relation. The Planner then semantically fuses the two complementary descriptions into the final HSI caption $\widehat{\mathbf{Y}}$.

This factorization mirrors our generation design: generation composes scene-agnostic motion dynamics with scene-conditioned interaction constraints, whereas understanding decomposes a scene-grounded motion into motion and interaction semantics. It therefore reduces motion-scene entanglement while retaining the information required for fine-grained HSI description.

Shared representations. The generation and scene-interaction understanding branches share the same motion and affordance representations. Given a human motion $\mathbf{X}$, scene geometry $\mathbf{S}$, and contact observations $\mathbf{C}$, the shared encoders produce $\mathbf{Z}_m = E_m(\mathbf{X}; \theta_m)$ and $\mathbf{Z}_a = E_a(\mathbf{S}, \mathbf{C}; \theta_a)$, where $\mathbf{Z}_m$ and $\mathbf{Z}_a$ denote the Motion and Affordance Tokens, respectively. These tokens condition both task-specific branches: $\mathcal{F}_{\mathrm{gen}} = G(\mathbf{Z}_m, \mathbf{Z}_a)$ and $\widehat{\mathbf{Y}}_s = U_s(\mathbf{X}, \mathbf{S}; \mathbf{Z}_m, \mathbf{Z}_a)$, where G and U_s denote the generation and scene-interaction understanding modules.

Training the two branches independently does not fully exploit their shared representations. In particular, captioning supervision mainly promotes semantic discrimination, whereas motion generation additionally constrains temporal dynamics, spatial placement, and physical interaction. We therefore use the generation task as auxiliary supervision for the shared tokens, encouraging $\mathbf{Z}_m$ to preserve coherent motion dynamics and $\mathbf{Z}_a$ to encode interaction-relevant geometric information.

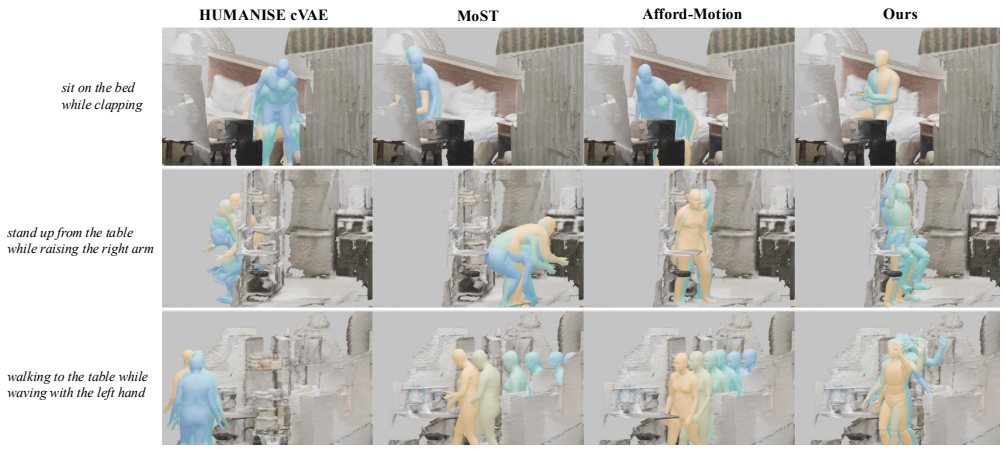


Figure 4: **Qualitative comparison of scene-grounded HSI generation.** Each row corresponds to one text instruction, and each column shows temporal motion snapshots produced by the same method. Compared with previous approaches, our method preserves the fine-grained action specified by the instruction while maintaining a scene-compatible interaction with the target object.

Joint optimization. We retain the original objectives of both branches. The generation branch uses the conditional flow-matching objective $\mathcal{L}_{\text{gen}}$, while the understanding branch uses the autoregressive captioning objective $\mathcal{L}_{\text{cap}}$. The joint objective is

$$\mathcal{L}_{\text{joint}} = \lambda_{\text{gen}}\mathcal{L}_{\text{gen}} + \lambda_{\text{cap}}\mathcal{L}_{\text{cap}}. \tag{7}$$

During joint fine-tuning, the task-specific generation and language modules remain fixed, and gradients from both objectives are accumulated only at the shared motion and affordance encoders. Consequently, generation supervision enhances the representations used for HSI understanding without introducing additional inference-time modules.

4 Experiments

4.1 Experimental Setup

We evaluate HSI generation and understanding on HUMANISE (Wang et al., 2022), motion captioning on augmented HumanML3D (Guo et al., 2022a), and cross-dataset physical generalization on LINGO (Jiang et al., 2024a) and TRUMANS (Jiang et al., 2024b).

For generation, we compare with HUMANISE cVAE (Wang et al., 2022), Afford-Motion (Wang et al., 2024), MoST (Cen et al., 2024), and HSI-GPT (Wang et al., 2025). For understanding, we compare with TM2T (Guo et al., 2022b), MotionGPT (Zhang et al., 2024b), and HSI-GPT. We use HY-Motion (Wen et al., 2025) and MotionGPT3 (Zhu et al., 2026) as pretrained motion priors. Dataset splits, baseline settings, and evaluation metrics are detailed in Appendix D; joint fine-tuning settings are provided in Appendix B.

4.2 Scene-Grounded HSI Generation

Table 1 compares our method with HUMANISE cVAE (Wang et al., 2022), Afford-Motion (Wang et al., 2024), MoST (Cen et al., 2024), and HSI-GPT (Wang et al., 2025). HSI-GPT generates motion through autoregressive motion-token prediction. Our method preserves pretrained root-relative motion dynamics and uses Trajectory-Flow to predict global root trajectories. With HY-Motion, our method reduces *Goal Distance* from 0.182 for HSI-GPT to $\mathbf{0.011}$. Using MotionGPT3 also reduces this metric to $\underline{0.033}$. These results show more accurate target placement with both motion priors. With MotionGPT3, *Contact* increases from 92.31 to $\mathbf{96.35}$, while *Non-collision* matches HSI-GPT at $\mathbf{99.82}$. This comparison shows that our framework improves target placement and contact while maintaining a high *Non-collision* score.

Table 1: **Quantitative results of text-conditioned HSI generation on HUMANISE. Bold** and underlined values indicate the best and second-best results.

Method	Goal Dist. ↓	APD ↑	Contact ↑	N-collision ↑
HUMANISE cVAE	0.422	4.094	84.06	99.77
MoST	0.384	-	75.56	99.80
Afford 1-stage @ Enc	0.326	**5.510**	76.11	99.71
Afford 1-stage @ Dec	0.185	4.063	86.43	99.76
Afford 2-stage @ Enc	0.156	2.597	95.86	99.69
Afford 2-stage @ Dec	0.156	2.459	96.04	99.70
HSI-GPT	0.182	3.492	92.31	**99.82**
Ours @ HY-Motion	**0.011**	2.684	95.96	99.71
Ours @ MotionGPT3	0.033	2.743	**96.35**	**99.82**

Table 2: **HSI captioning results on the augmented HumanML3D benchmark. Bold** and underlined values indicate the best and second-best results.

Method	R-Precision ↑			MultiModal Dist. ↓	BLEU-1 ↑	BLEU-4 ↑	ROUGE ↑	CIDEr ↑	BERTScore ↑
	Top-1	Top-2	Top-3						
TM2T (Guo et al., 2022b)	0.516	0.720	0.823	2.935	**48.9**	7.0	38.1	16.8	32.2
MotionGPT (Zhang et al., 2024b)	0.543	-	0.827	2.821	48.2	12.5	37.4	29.2	32.4
HSI-GPT (Wang et al., 2025)	0.551	**0.735**	0.838	2.746	48.6	14.2	38.4	30.8	31.8
Ours	**0.578**	0.711	**0.846**	**2.744**	37.9	**18.0**	**39.4**	**31.5**	**34.7**

Figure 4 further shows that this formulation preserves fine-grained actions such as clapping, raising an arm, and waving while placing the motion at the intended object. More qualitative results are provided in Appendix H.

4.3 HSI UNDERSTANDING

Motion and interaction captioning. We evaluate captioning on augmented HumanML3D and compare with HSI-GPT, which also supports both generation and understanding. Table 2 shows that our method improves *Top-1* and *Top-3 R-Precision* from 0.551 and 0.838 to $\mathbf{0.578}$ and $\mathbf{0.846}$, respectively. These gains indicate better motion-text matching. *MultiModal Distance* also decreases slightly from 2.746 to $\mathbf{2.744}$.

For caption quality, *BLEU-4* increases from 14.2 to $\mathbf{18.0}$, indicating better agreement with reference phrases. *ROUGE* and *CIDEr* increase from 38.4 and 30.8 to $\mathbf{39.4}$ and $\mathbf{31.5}$, respectively. *BERTScore* rises from 31.8 to $\mathbf{34.7}$, suggesting greater semantic similarity to the reference descriptions. Together with the generation results, these comparisons show improvements over HSI-GPT in both target placement and motion description within a unified framework.

Generation-enhanced HSI understanding. We next study whether the generation objective improves the shared motion and affordance representations used for scene-interaction understanding. Table 5 reports verb and object recognition on HUMANISE. Joint optimization improves *Verb Accuracy* from 95.02% to $\mathbf{97.58}\%$ and *Object Accuracy* from 68.32% to $\mathbf{71.21}\%$, corresponding to gains of 2.56 and 2.89 percentage points, respectively. These results show improvements in both action and object recognition, with a larger gain in object accuracy. They suggest that generation supervision helps the shared encoders capture motion dynamics and scene-interaction information. Since only the shared encoders are updated, these gains introduce no additional inference-time component.

4.4 CROSS-DATASET PHYSICAL GENERALIZATION

We evaluate whether the generated motions remain physically plausible outside the HUMANISE training distribution. The results are reported in Table 8 in Appendix C.4. Our method reduces the proportion of *penetrated frames* from 64.58% to $\mathbf{38.82}\%$ on LINGO and from 68.20% to $\mathbf{46.96}\%$

Table 3: Effect of directly encoding complete motion in world coordinates. Absolute XYZ entangles local articulated motion with global scene placement, whereas t_0-anchored XYZ removes the initial global translation. Results are averaged over three downstream seeds. **Bold** values indicate the best results within each codec family. Lower is better for all metrics.

Codec	Coordinates	Future MPJPE (mm) ↓	Local MPJPE (mm) ↓	Pelvis ADE (mm) ↓	Pelvis FDE (mm) ↓	Boundary-step error (mm) ↓	Acceleration error (mm/frame^2) ↓
MLD-VAE	Absolute XYZ	47.97	17.28	45.31	60.34	35.86	1.29
	t_0-anchored XYZ	**29.30**	**12.42**	**25.69**	**36.88**	**20.63**	**0.97**
VQ-VAE	Absolute XYZ	331.58	116.88	316.75	370.63	253.57	15.87
	t_0-anchored XYZ	**138.00**	**60.69**	**122.79**	**165.60**	**90.98**	**8.51**
RVQ-VAE	Absolute XYZ	371.13	128.42	354.34	423.64	293.74	10.91
	t_0-anchored XYZ	**190.01**	**88.37**	**168.42**	**235.67**	**123.66**	**8.60**

Table 4: **Ablation of Motion and Affordance Tokens for HSI understanding.** All values are reported on the same HUMANISE validation split. **Bold** and underlined values indicate the best and second-best results.

Setting	BLEU-4 ↑	ROUGE-L ↑	Verb Acc. ↑	Object Acc. ↑
w/o Motion Tokens	25.06	54.56	61.21	61.52
w/o Affordance Tokens	42.83	66.85	88.18	46.97
Motion + Affordance Tokens	**46.58**	**86.15**	**97.58**	**71.21**

Table 5: **Verb and object recognition on HUMANISE.** † denotes our local reproduction results. **Bold** and underlined values indicate the best and second-best results.

Method	Verb Acc. ↑	Object Acc. ↑
MotionGPT†	90.91	32.42
Ours (Baseline)	95.02	68.32
Ours (Joint Optimization)	**97.58**	**71.21**

on TRUMANS. It also slightly improves *Non-collision* on both datasets. On LINGO, the *mean penetration depth* decreases from 1.54 to $\mathbf{1.34}$ cm, whereas the official TRUMANS baseline retains a lower depth when penetration occurs. Similarly, the released baselines obtain higher *Contact* scores. The results therefore indicate that our method more consistently avoids pervasive scene penetration, although accurate contact and penetration severity remain challenging under cross-dataset transfer. Full contact, support, and clip-level diagnostics are reported in Appendix C.4.

4.5 Representation Analysis and Ablations

We test the two design choices that connect the framework. First, we compare absolute world coordinates with root-origin anchoring across three motion codecs. Table 3 shows lower local-pose, trajectory, boundary, and acceleration errors for the anchored representation. The result supports the separation of body dynamics from scene placement.

Second, we test how the shared representations contribute to understanding. Table 4 shows that Motion Tokens are most important for verb recognition and caption quality, while Affordance Tokens provide the strongest evidence for the target object. Their complementary effects support the shared interface used by generation and understanding.

Finally, Table 5 compares independent and joint training. Both verb and object accuracy improve when generation and captioning update the shared encoders. This result verifies that the common interface can benefit from both tasks, although cross-task transfer is not the sole focus of the framework. Fusion variants and full training details are provided in Appendices C.1 and B.

5 Conclusion

HSI-Agent unifies planning, generation, and understanding through shared Motion and Affordance Tokens. The collaborative planner turns multi-step instructions into scene-grounded plans for long-sequence generation. Trajectory-Flow predicts global root trajectories to adapt pretrained motions to scenes, while joint training improves understanding through generation supervision. Experiments show improved generation, understanding, and cross-dataset physical generalization. Future work will explore diverse trajectory data and stronger physical constraints to improve interaction quality.

REFERENCES

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

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

Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *European Conference on Computer Vision (ECCV)*, 2020.

Zhi Cen, Huaijin Pi, Sida Peng, Zehong Shen, Minghui Yang, Shuai Zhu, Hujun Bao, and Xiaowei Zhou. Generating human motion in 3d scenes from text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1855–1866, 2024.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18000–18010, 2023.

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*, 2022a.

Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. TM2T: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *Proceedings of the European Conference on Computer Vision*, 2022b.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. MoMask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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*, 2021a.

Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3d scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021b.

Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics*, 35(4), 2016.

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 (ECCV)*. Springer, 2022.

Nan Jiang, Zimo He, Zi Wang, Hongjie Li, Yixin Chen, Siyuan Huang, and Yixin Zhu. Autonomous character-scene interaction synthesis from text instruction. In *ACM SIGGRAPH Asia 2024 Conference Papers*, 2024a. doi: 10.1145/3680528.3687595.

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.

Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. In *Advances in Neural Information Processing Systems*, 2019.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Yongsen Mao, Junhao Zhong, Chuan Fang, Jia Zheng, Rui Tang, Hao Zhu, Ping Tan, and Zihan Zhou. SpatialLM: Training large language models for structured indoor modeling. In *Advances in Neural Information Processing Systems*, 2025.

Zoltán Á. Milacski, Koichiro Niinuma, Ryosuke Kawamura, Fernando De la Torre, and László A. Jeni. GHOST: Grounded human motion generation with open vocabulary scene-and-text contexts. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Mathis Petrovich, Michael J. Black, and G"ul Varol. Action-conditioned 3d human motion synthesis with transformer VAE. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Mathis Petrovich, Michael J. Black, and G"ul Varol. TEMOS: Generating diverse human motions from textual descriptions. In *Proceedings of the European Conference on Computer Vision*, 2022.

Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*, 4(4):236–252, 2016.

Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. Audio to body dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7574–7583, 2018.

Shanlin Sun, Jiaqi Xu, Gabriel de Araujo, Shenghan Zhou, Hanwen Zhang, Ziheng Huang, Chenyu You, and Xiaohui Xie. CoMA: Compositional human motion generation with multi-modal agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9206–9214, 2026. doi: 10.1609/aaai.v40i11.37878.

Taoran Tang, Jia Jia, and Hanyang Mao. Dance with melody: An LSTM-autoencoder approach to music-oriented dance synthesis. In *Proceedings of the 26th ACM International Conference on Multimedia*, 2018.

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

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, 2017.

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*, 2015.

Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Yuan Wang, Yali Li, Xiang Li, and Shengjin Wang. HSI-GPT: A general-purpose large scene-motion-language model for human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7147–7157, 2025.

Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. HUMANISE: Language-conditioned human motion generation in 3d scenes. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

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*, 2024.

Yuxin Wen, Qing Shuai, Di Kang, Jing Li, Cheng Wen, Yue Qian, Ningxin Jiao, Changhai Chen, Weijie Chen, Yiran Wang, et al. HY-Motion 1.0: Scaling flow matching models for text-to-motion generation. *arXiv preprint arXiv:2512.23464*, 2025.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

Qi Wu, Yubo Zhao, Yifan Wang, Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. Motion-agent: A conversational framework for human motion generation with LLMs. In *International Conference on Learning Representations*, 2025b.

Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22193–22204, 2025c.

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. In *International Conference on Learning Representations*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In *Robotics: Science and Systems*, 2024.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. MotionDiffuse: Text-driven human motion generation with diffusion models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024a.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020.

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, 2024b.

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*, 2019.

Zixiang Zhou, Yu Wan, and Baoyuan Wang. AvatarGPT: All-in-one framework for motion understanding, planning, generation and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1357–1366, 2024.

Bingfan Zhu, Biao Jiang, Sunyi Wang, Shixiang Tang, Tao Chen, Linjie Luo, Youyi Zheng, and Xin Chen. MotionGPT3: Human motion as a second modality. In *International Conference on Learning Representations*, 2026.

APPENDIX

A PRELIMINARIES

A.1 PRETRAINED MOTION PRIORS

Recent text-to-motion models learn rich human motion distributions from large-scale motion-language datasets. Given a textual description $\mathbf{y}$, a pretrained motion generator produces a motion sequence

$$\widetilde{\mathbf{X}} = M_{\text{pre}}(\mathbf{y}), \tag{8}$$

where $\widetilde{\mathbf{X}} \in \mathbb{R}^{T\times J\times 3}$ contains the positions of J body joints over T frames. By training on diverse motion data, these models provide a strong prior for natural articulated poses, temporal dynamics, and fine-grained action semantics.

However, most pretrained motion generators operate independently of the target 3D scene. Consequently, their outputs specify how a person moves but do not determine where the motion should be performed or how it should be spatially aligned with an object. In this work, we retain the diverse local motion dynamics provided by pretrained motion generators and adapt their global trajectories to scene-specific interaction constraints.

A.2 SCENE AFFORDANCE

Scene affordance describes the action possibilities that the environment provides to a human. In language-guided HSI generation, Afford-Motion Wang et al. (2024) introduces scene affordance as an explicit intermediate representation between scene grounding and motion generation. It further encodes the affordance representation into latent features for conditioning the motion generator. We follow this established formulation rather than treating affordance encoding itself as a contribution of our work.

Specifically, affordance can be represented as a body-part-aware interaction field over the scene point cloud. For a scene point $\mathbf{s}_n$ and the trajectory of the j-th body joint, their interaction distance is defined as

$$d_{n,j} = \min_t \|\mathbf{s}_n - \mathbf{x}_{t,j}\|_2, \qquad a_{n,j} = \exp\left(-\frac{d_{n,j}^2}{2\sigma^2}\right), \tag{9}$$

where $\mathbf{x}_{t,j}$ denotes the position of joint j at frame t and σ controls the spatial extent of the interaction response. A larger $a_{n,j}$ indicates that scene point $\mathbf{s}_n$ is more likely to be associated with the corresponding body part.

The resulting affordance map $\mathbf{A} = [a_{n,j}] \in \mathbb{R}^{N_s\times J_c}$ jointly describes interaction-relevant scene locations and their associations with J_c selected body joints. An affordance encoder maps the scene geometry and affordance responses to a latent representation:

$$\mathbf{Z}_a = E_a(\mathbf{S}, \mathbf{A}), \tag{10}$$

where $\mathbf{Z}_a$ denotes the encoded affordance features. In our framework, this established affordance representation serves as an interface between the scene and the downstream generation and understanding modules. Together with the pretrained motion prior, it provides the motion dynamics and scene-interaction constraints required for scene-grounded HSI modeling.

A.3 CONDITIONAL FLOW-MATCHING PATH

Trajectory-Flow uses a linear path between a noise trajectory $\boldsymbol{\tau}_i^0$ and the ground-truth root trajectory $\boldsymbol{\tau}_i^*$. At flow time $s \in [0, 1]$, the interpolated trajectory and its target velocity are

$$\begin{aligned} \boldsymbol{\tau}_i^s &= (1-s)\boldsymbol{\tau}_i^0 + s\boldsymbol{\tau}_i^*, \\ \mathbf{u}_i^s &= \boldsymbol{\tau}_i^* - \boldsymbol{\tau}_i^0. \end{aligned} \tag{11}$$

The velocity network takes the interpolated trajectory and flow time as inputs, together with Motion Tokens, Affordance Tokens, and the text embedding. It learns to predict $\mathbf{u}_i^s$ using the loss in Equation 4. At inference time, integration of the learned velocity field transforms an initial noise trajectory into the predicted global root trajectory.

Table 6: **Generation-side token fusion comparison on HUMANISE. Bold** and underlined values indicate the best and second-best results.

Fusion	MPJPE (mm) ↓	Scene MPJPE (mm) ↓	APD ↑
Concat	17.30	78.70	**0.1461**
MLP	<u>17.21</u>	<u>78.63</u>	0.1456
Gated	**17.18**	**78.15**	<u>0.1460</u>

B Joint Fine-Tuning Details

We initialize the model from independently trained generation and understanding checkpoints. During joint fine-tuning, the task-specific generation and language heads are kept frozen, and only the shared motion and affordance encoders are updated. We set $\lambda_{\text{gen}} = 4$ and $\lambda_{\text{cap}} = 1$ in Equation 7. Approximately 17.93M motion-encoder parameters and 0.63M affordance-encoder parameters are optimized. We use AdamW (Loshchilov & Hutter, 2019) with learning rates of 2×10^{-8} and 5×10^{-8} for the motion and affordance encoders, respectively. The model is trained for three epochs with a batch size of 8, BF16 mixed precision, cosine learning-rate decay, and a gradient clipping threshold of 1.0. Since no additional module is introduced, joint fine-tuning does not increase the computational cost at inference time.

C Additional Experimental Results

C.1 Motion and Affordance Token Fusion

We compare three ways of combining Motion and Affordance Tokens on HUMANISE. Motion Tokens describe body poses and motion dynamics, while Affordance Tokens provide information about scene interactions. The comparison examines how the fusion method affects motion prediction and interaction understanding. Generation and understanding results are reported in Tables 6 and 7, respectively.

Fusion methods. **Concat** directly concatenates the two token streams. It retains both sources of information and leaves their interaction to the downstream module. **MLP** uses a learned nonlinear transformation to combine motion and affordance features before passing them to the downstream module. This introduces a learned fusion step beyond direct concatenation. **Gated** uses learned gates to regulate the contributions of motion and affordance features during fusion. The main distinction is therefore how the two sources are combined: Concat joins them directly, MLP transforms their features, and Gated controls their contributions through learned weights.

Effects on HSI generation. Table 6 shows that MLP slightly reduces *MPJPE* from 17.30 to <u>17.21</u> mm and *Scene MPJPE* from 78.70 to <u>78.63</u> mm compared with Concat. Gated fusion further reduces these errors to $\mathbf{17.18}$ and $\mathbf{78.15}$ mm, respectively. Compared with MLP, the reductions are 0.03 mm in *MPJPE* and 0.48 mm in *Scene MPJPE*. These are small improvements, with a larger absolute reduction in scene-space error. Meanwhile, Gated fusion obtains an *APD* of <u>0.1460</u>, close to the $\mathbf{0.1461}$ achieved by Concat. Thus, Gated fusion provides the lowest prediction errors while retaining comparable motion diversity.

Effects on HSI understanding. Table 7 shows larger differences in caption quality. Compared with Concat, MLP improves *BLEU-4* from 32.86 to <u>35.19</u>. Its *ROUGE-L*, *Verb Accuracy*, and *Object Accuracy* improve only slightly. Gated fusion provides a further improvement over MLP, raising *BLEU-4* from 35.19 to $\mathbf{44.38}$ and *ROUGE-L* from 81.27 to $\mathbf{85.35}$. The gains of 9.19 and 4.08 points indicate better agreement with reference phrases and descriptions. Gated fusion also improves *Verb Accuracy* from 94.59% to $\mathbf{95.02}\%$ and *Object Accuracy* from 67.57% to $\mathbf{68.32}\%$. These gains of 0.43 and 0.75 percentage points show improvements in both action and target-object recognition. The larger gain in object accuracy suggests that the fusion method helps the understanding branch use scene-interaction information.

Table 7: **Understanding-side token fusion comparison on HUMANISE. Bold** and underlined values indicate the best and second-best results.

Fusion	BLEU-4 ↑	ROUGE-L ↑	Verb Acc. ↑	Object Acc. ↑
Concat	32.86	81.22	94.43%	67.53%
MLP	<u>35.19</u>	<u>81.27</u>	<u>94.59%</u>	<u>67.57%</u>
Gated	**44.38**	**85.35**	**95.02%**	**68.32%**

Discussion. Gated fusion achieves the lowest generation errors and the best results across all four understanding metrics. The generation results show modest accuracy gains with similar motion diversity. The understanding results show clearer gains in caption quality, together with improved action and object recognition. These findings support learned gating as a useful way to combine the complementary information in Motion and Affordance Tokens.

C.2 Generation-Enhanced HSI Understanding

The baseline and joint variants use the same HUMANISE split and decoding protocol. Joint optimization updates only the shared encoders, while the task-specific generation and language modules remain frozen. The results are reported in Table 5 in the main paper.

C.3 Motion and Affordance Token Ablation

We remove one token stream at a time while retaining the scene input and all remaining components. Table 4 in the main paper reports caption quality, verb recognition, and object recognition under the same HUMANISE validation protocol.

C.4 Extended Cross-Dataset Physical Evaluation

Table 8 provides the complete physical diagnostics for the shared 64-clip cohorts on LINGO and TRUMANS. To ensure a controlled comparison, the official baseline and Ours-MotionGPT3 are evaluated on exactly the same clips using the same scene representation, body joints, distance thresholds, and evaluation implementation. We separate the results into primary contact and penetration measures, auxiliary support diagnostics, and a clip-level collision-free rate.

On the primary physical measures, our method reduces the proportion of *penetrated frames* from 64.58% to $\mathbf{38.82}\%$ on LINGO and from 68.20% to $\mathbf{46.96}\%$ on TRUMANS. *Non-collision* also improves slightly on both datasets. On LINGO, *mean and maximum penetration depths* decrease from 1.54 and 3.42 cm to $\mathbf{1.34}$ and $\mathbf{3.29}$ cm, respectively. On TRUMANS, however, the official baseline produces smaller penetration depths when a collision occurs. The official methods also obtain higher *Contact*, indicating that reducing scene penetration does not automatically guarantee more accurate interaction contact.

The auxiliary support diagnostics provide a more detailed view of this trade-off. Ours-MotionGPT3 increases the rate of detected foot support and reduces the proportion of floating-foot samples on both datasets. However, its foot-skating error is higher, showing that improved static support does not yet translate into stable temporal foot contact. *Foot-to-support gap* is computed only for clips in which local support is detected; because the number of valid clips differs substantially across methods, this value is reported for diagnosis rather than direct ranking.

At the clip level, our method produces more fully collision-free samples on both datasets. Overall, the extended evaluation shows that our model transfers more robustly in terms of avoiding persistent scene intersection, while precise contact placement, penetration severity, and foot stability remain important directions for future improvement. The TRUMANS comparison should additionally be interpreted in light of its stronger input condition: the released baseline receives two ground-truth initial frames and the ground-truth terminal pelvis position in the horizontal plane.

Table 8: **Physical comparison on LINGO and TRUMANS.** Only the official method and Ours-MotionGPT3 are compared; all values use the exact shared 64-clip cohort and the same evaluator. **Bold** values indicate the best results within each dataset.

		LINGO (exact-64)		TRUMANS (exact-64)	
Metric	Unit	Official	Ours-MotionGPT3	Official[‡]	Ours-MotionGPT3
Contact	%	**92.29**	84.87	**100.00**	90.24
Non-collision	%	99.964	**99.966**	99.927	**99.959**
Scene-point penetration	%	0.036	**0.034**	0.073	**0.041**
Penetrated frames	%	64.58	**38.82**	68.20	**46.96**
Mean penetration depth	cm	1.54	**1.34**	**1.29**	2.76
Max penetration depth	cm	3.42	**3.29**	**3.36**	8.76
Floating-foot samples[†]	%	99.89	**92.12**	76.71	**71.51**
Underground-foot samples[†]	%	**0.00**	**0.00**	**0.00**	**0.00**
Detected foot support[†]	%	0.11	**12.70**	23.55	**41.50**
Buried joint probes[†]	%	1.45	**1.33**	2.79	**0.68**
Foot skating[†]	cm/frame	**0.09**	1.57	**0.56**	1.30
Fully collision-free clips	%	12.50	**14.06**	7.81	**15.62**

[†] Support diagnostics use a sparse 8,192-point local-support probe and are auxiliary rather than primary ranking evidence. Foot-to-support gap is averaged only over clips with detected support and is intentionally not bold-ranked. [‡] The released TRUMANS baseline is a trajectory oracle: it receives two ground-truth initial frames and the ground-truth terminal pelvis position in the horizontal plane. Non-collision and scene-point penetration are exact complements and should not be counted as two independent improvements.

D Experimental Protocol

Datasets and evaluation cohorts. HUMANISE (Wang et al., 2022) is our primary benchmark for text-conditioned HSI generation and scene-interaction understanding. For motion captioning, we follow the augmented HumanML3D protocol used by prior HSI understanding methods. We additionally evaluate physical plausibility on LINGO (Jiang et al., 2024a) and TRUMANS (Jiang et al., 2024b) to measure cross-dataset generalization to unseen motions and scenes. For this comparison, all methods are evaluated using the same evaluator on the exact 64 clips available to both the released baselines and our method.

Baselines. For HSI generation, we compare against the HUMANISE cVAE baseline (Wang et al., 2022), the encoder and decoder variants of the one-stage and two-stage Afford-Motion model (Wang et al., 2024), MoST (Cen et al., 2024), and HSI-GPT (Wang et al., 2025). We instantiate our framework with HY-Motion (Wen et al., 2025) and MotionGPT3 (Zhu et al., 2026) to examine whether Trajectory-Flow operates independently of a particular pretrained motion generator. For HSI understanding, we compare with TM2T (Guo et al., 2022b), MotionGPT (Zhang et al., 2024b), and HSI-GPT (Wang et al., 2025). A † indicates a local reproduction using the official implementation and released checkpoint.

Evaluation metrics. For scene-grounded generation, *Goal Distance* measures the distance between the generated terminal location and the interaction target, and *APD* measures sample diversity. *Contact* and *Non-collision* evaluate physical compatibility with the scene. The cross-dataset evaluation additionally reports *penetrated frames*, penetration depth, local foot support, foot skating, and clip-level failure statistics. For captioning, we use *R-Precision* and *MultiModal Distance* following the HumanML3D protocol (Guo et al., 2022a), together with *BLEU* (Papineni et al., 2002), *ROUGE* (Lin, 2004), *CIDEr* (Vedantam et al., 2015), and *BERTScore* (Zhang et al., 2020). On HUMANISE, we further report *Verb Accuracy* and *Object Accuracy* to separately measure action and target-object recognition.

Reproduction protocol. We use released implementations and checkpoints whenever available. All locally reproduced methods use the same input clips and evaluation code as our method. Unless

otherwise stated, the reported values follow the standard test protocol of the corresponding benchmark. Complete physical diagnostics are reported in Appendix C.4, and long-sequence composition details are provided in Appendix F.

E Analysis of Full-Body Coordinate Representations

We investigate why directly encoding a complete scene-aligned motion is unsuitable for our generation module. A full-body sequence represented in absolute world coordinates simultaneously contains local articulation and global scene placement. Consequently, identical local motions performed at different locations may be assigned substantially different latent codes, making the representation sensitive to arbitrary global translations.

We compare absolute world-coordinate encoding with t_0-anchored encoding, where the initial global translation is removed before encoding. We evaluate three motion codecs: MLD-VAE (Chen et al., 2023), VQ-VAE (van den Oord et al., 2017), and RVQ-VAE (Guo et al., 2024). For each codec, a downstream model predicts 32 future frames from the same observed motion. Errors are measured between the predicted future motion and the corresponding canonical codec oracle. Results are averaged over three independently trained downstream models.

As shown in Table 3, directly encoding complete motion in absolute world coordinates consistently degrades both global trajectory prediction and local motion quality. Absolute encoding increases *future MPJPE* from $\mathbf{29.30}$ to $\underline{47.97}$ mm for MLD-VAE, from $\mathbf{138.00}$ to $\underline{331.58}$ mm for VQ-VAE, and from $\mathbf{190.01}$ to $\underline{371.13}$ mm for RVQ-VAE. The increase in *local MPJPE* further demonstrates that the degradation is not solely caused by global translation errors. Entangling global scene coordinates with articulated motion also makes local pose dynamics more difficult to represent.

These observations motivate our factorized generation design. We obtain diverse root-relative body dynamics from a pretrained motion generator and use Trajectory-Flow to model the scene-conditioned global root trajectory. Consequently, the generation module does not need to learn the complete human motion distribution from an HSI dataset with limited action diversity.

F Long-Sequence Motion Composition

Long-sequence generation must preserve both local articulated dynamics and global spatial consistency. We handle these two aspects separately: VQ-VAE composition regularizes the root-relative body motion, whereas root-trajectory alignment connects the globally positioned segments.

F.1 VQ-VAE-Based Motion Composition

Motion segments generated independently may have incompatible boundary poses and velocities. Direct concatenation can therefore introduce visible discontinuities. We convert each base motion into the representation used by the VQ-VAE, encode it into discrete motion tokens, concatenate the token sequences, and decode them jointly:

$$\begin{aligned} \mathbf{F}_i &= \Phi(\widetilde{\mathbf{X}}_i), \qquad \mathbf{q}_i = Q_{\mathrm{vq}}\left(E_{\mathrm{vq}}(\mathbf{F}_i)\right), \\ \overline{\mathbf{F}} &= D_{\mathrm{vq}}\left(\mathbf{q}_1 \| \cdots \| \mathbf{q}_K\right). \end{aligned} \tag{12}$$

Unlike independent decoding, joint decoding provides temporal context from both sides of each segment boundary. The learned motion prior can therefore regularize discontinuous poses and velocities at the transition.

The purpose of joint decoding is to improve boundary transitions rather than overwrite the internal dynamics of each segment. We divide $\overline{\mathbf{F}}$ back into segment-level features $\overline{\mathbf{F}}_i$ and blend them with the original features only within a short boundary window:

$$\mathbf{F}'_{i,t} = \mathbf{F}_{i,t} + \omega_{i,t}\left(\overline{\mathbf{F}}_{i,t} - \mathbf{F}_{i,t}\right), \qquad \omega_{i,t} = 0 \text{ if } t \notin \mathcal{W}_i, \tag{13}$$

where $\mathcal{W}_i$ is the temporal boundary window. Frames near the boundary receive a larger contribution from the jointly decoded sequence, while frames farther away preserve the original motion.

The refined joint sequence and its motion tokens are recovered as

$$\overline{\mathbf{X}}_i = \Phi^{-1}(\mathbf{F}'_i), \qquad \overline{\mathbf{Z}}_{m,i} = E_m(\overline{\mathbf{X}}_i; \theta_m). \tag{14}$$

Trajectory-Flow then uses $\overline{\mathbf{Z}}_{m,i}$ together with the affordance and text conditions to generate the scene-grounded trajectory.

F.2 Root-Trajectory Alignment

VQ-VAE composition improves local motion continuity, but independently generated segments may still have inconsistent global coordinate frames. We therefore align their root trajectories sequentially.

For segment i, the terminal root state of segment $i-1$ is used as its target initial state. The corresponding translation and rotation offsets are

$$\begin{aligned} \Delta\mathbf{p}_i &= \widehat{\mathbf{p}}_{i-1,T_{i-1}} - \widehat{\mathbf{p}}_{i,1}, \\ \Delta\mathbf{R}_i &= \widehat{\mathbf{R}}_{i-1,T_{i-1}} \widehat{\mathbf{R}}_{i,1}^{\mathsf{T}}. \end{aligned} \tag{15}$$

We apply the translation offset to the current root trajectory and gradually attenuate the rotation correction within a short transition window:

$$\begin{aligned} \widehat{\mathbf{p}}'_{i,t} &= \widehat{\mathbf{p}}_{i,t} + \Delta\mathbf{p}_i, \\ \widehat{\mathbf{R}}'_{i,t} &= \mathrm{Slerp}\left(\Delta\mathbf{R}_i, \mathbf{I}; \alpha_t\right) \widehat{\mathbf{R}}_{i,t}, \end{aligned} \tag{16}$$

where $\mathbf{I}$ is the identity rotation and α_t increases from zero to one over the transition window. The full correction is therefore applied at the boundary and gradually decays toward the original orientation of the current segment.

F.3 Boundary Pose Refinement

After VQ-VAE composition and root alignment, a small residual discrepancy may remain between the root-relative poses of adjacent segments. We remove this discrepancy through a short boundary refinement:

$$\begin{aligned} \mathbf{Q}'_{i,t} &= (1 - w_t)\mathbf{Q}_{i-1,T_{i-1}} + w_t \overline{\mathbf{Q}}_{i,t}, \\ w_t &= 3r_t^2 - 2r_t^3, \qquad r_t = \frac{t-1}{W-1}, \end{aligned} \tag{17}$$

where W is the length of the refinement window. The smoothstep weight reduces additional velocity discontinuities introduced by interpolation.

This refinement is applied only to the root-relative articulated pose. The scene-grounded root trajectory remains unchanged. The final global motion is reconstructed as

$$\widehat{\mathbf{X}}'_{i,t} = \widehat{\mathbf{R}}'_{i,t}\mathbf{Q}'_{i,t} + \widehat{\mathbf{p}}'_{i,t}. \tag{18}$$

After removing duplicated boundary frames, the complete sequence is

$$\widehat{\mathbf{X}} = \widehat{\mathbf{X}}_1 \oplus \widehat{\mathbf{X}}'_2[2{:}] \oplus \cdots \oplus \widehat{\mathbf{X}}'_K[2{:}]. \tag{19}$$

Overall, VQ-VAE joint decoding regularizes local motion transitions, Trajectory-Flow generates scene-compatible root trajectories, and root-trajectory alignment maintains global spatial continuity across successive interactions.

G Role of the Planner in Long-Horizon HSI Generation

Long-horizon HSI generation requires an ordered set of actions, target objects, and spatial constraints. Later interactions may also depend on earlier choices. For example, an instruction to walk to a chair, sit down, stand up, and then approach a table requires both a suitable action order and a target object for each interaction. The Planner organizes these requirements into the plan units defined in Section 3.2. It connects the overall instruction to the conditions used by the motion generator.

Instruction decomposition and action order. The Language Planner identifies the actions requested by the user and their temporal dependencies. It breaks a multi-step instruction into ordered interaction units. Each unit specifies an action, a target, and the relevant conditions. This gives the motion generator a clear task for each segment, while the full plan preserves the order required by the instruction. In the example above, sitting precedes standing up, and approaching the table follows the interaction with the chair.

Target grounding and planning context. The Scene Planner links each interaction intent to an object instance in the input scene and extracts its spatial properties. The resulting target and constraints are added to the planning context before the next unit is constructed, as described in Equation 1. This makes earlier target choices and interaction conditions available to later planning steps. The Language Planner handles the meaning and order of the requested actions, while the Scene Planner provides the object and spatial information needed to place them in the scene.

From interaction plans to continuous motion. The ordered plan supplies the action and scene constraints for each motion segment. Pretrained motion models provide the body-motion priors, and Trajectory-Flow predicts their global root trajectories. During sequence composition, the terminal root state of one segment is used to align the next segment. VQ-VAE composition and boundary refinement handle local pose transitions. These operations are detailed in Appendix F. Planning therefore organizes what should happen, which objects are involved, and in what order. Motion generation and composition turn these conditions into a continuous sequence. Both are needed to connect multi-step instructions with scene-grounded motion.

Supplementary videos. Videos of generated long-horizon HSI sequences are provided in the supplementary material. They allow the generated interactions and the transitions between successive motion segments to be viewed over time.

H Additional Qualitative Results

Figure 5 presents additional qualitative results of scene-grounded HSI generation. These examples cover different motion instructions, target objects, and scene layouts, providing a broader visualization of the generated articulated motions and their spatial alignment with the surrounding environments. qualitative comparisons with existing methods. Across diverse interaction scenarios, our method generates more realistic human motions that more faithfully follow the input text descriptions while maintaining plausible spatial relationships with the target objects and surrounding scenes.

Videos showing generated long-sequence HSI and the agent interaction interface are provided in the supplementary material.



Figure 5: **Additional qualitative comparisons of scene-grounded HSI generation.** Compared with existing methods, our approach produces more realistic human motions that better conform to the input text descriptions.