HARNESSING COUPLED STREAM COMPLETION FOR HUMAN–OBJECT INTERACTION MODELING

Anonymous authors
Paper under double-blind review

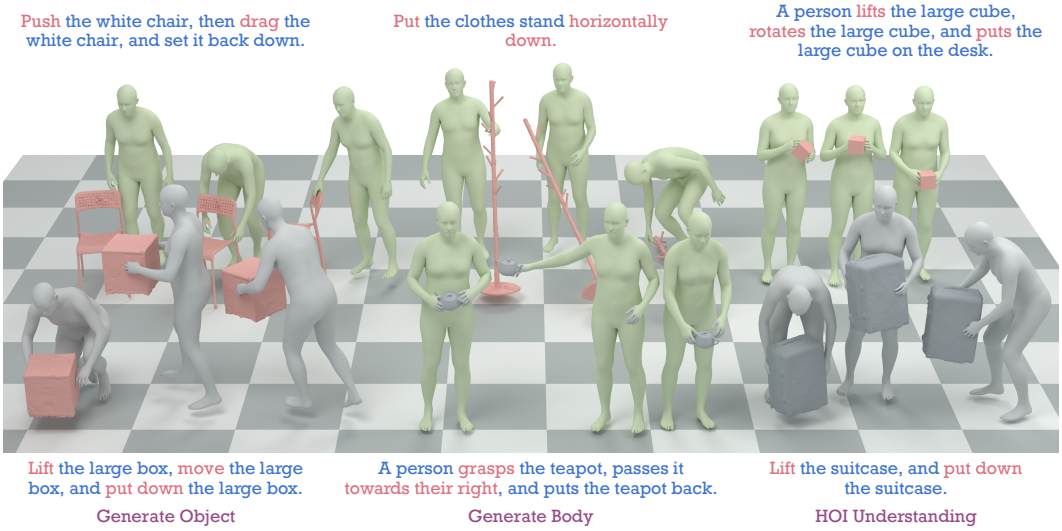

Figure 1: Given a text prompt, our TRACE generates coherent human-object interaction motions via a coupled latent flow, while its structured representation further supports stream completion and HOI understanding.

ABSTRACT

Text-conditioned human-object interaction (HOI) generation requires body motion, object trajectories & rotations, and hand articulation to remain coordinated. These components differ in scale and dynamics, but must agree on contact, relative pose, and timing. A shared representation may limit the distinct structure of each stream, while independent generation prevents each stream from responding to changes in the others. Latent supervision alone also does not directly constrain contact after decoding. We propose TRACE, a continuous latent framework that keeps stream states separate and couples their updates. TRACE encodes body, object, and hand motion into separate latents and predicts each stream velocity from the complete current interaction state. Geometric losses on decoded motion further constrain contact and object-relative motion over time. The same model supports completion of any single absent stream from the other two. Frozen flow features also serve as input to a language model for HOI understanding. Experiments on InterAct, OMOMO, and BEHAVE show that joint completion training improves generation and that frozen flow features improve understanding over raw-motion encoding. On InterAct, TRACE achieves the highest contact precision, recall, and F1 among the compared methods.

1 INTRODUCTION

Text-conditioned Human-Object Interaction (HOI) Generation synthesizes a 3D motion sequence in which a human interacts with a dynamic object according to a language description. Realistic HOI motion supports animation, virtual reality, gaming, and robotics simulation. Recent generative models have extended text-conditioned motion synthesis from isolated human actions to interactions with moving objects (Peng et al., 2025; Diller & Dai, 2024; Li et al., 2024a; Wu et al., 2025a; Song et al.,

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.

2024; Wang et al., 2026b). Unlike ordinary text-to-motion, the object is not merely a static condition: its generated trajectory changes the feasible human motion throughout the sequence. HOI comprises body motion, object trajectories, and fine-grained hand articulation, which differ substantially in scale, dimensionality, and dynamics. Generating coherent interaction therefore requires preserving these distinct structures while aligning contact locations, relative poses, and interaction timing.

These heterogeneity raise a central question: how can each stream retain its own structure while all streams stay coordinated? Recent generators coordinate motion through shared denoisers, separate branches, or auxiliary contact and relation guidance (Peng et al., 2025; Diller & Dai, 2024; Wu et al., 2025a; Song et al., 2024; Xue et al., 2025; Wang et al., 2026b). Two design choices matter here: how to represent each stream and how to exchange information across streams. A shared representation allows information exchange but may limit the distinct structure of each stream. Separate representations preserve this structure, but each update still needs access to the current states of the other streams. With independent dynamics, each update depends only on its own stream and the shared conditions. Even under the same text and geometry conditions, a body update cannot respond to the current object and hand states, and vice versa. The streams may therefore appear plausible individually while producing missed contact, object drift, or inconsistent interaction timing when combined. Compressed latent spaces reduce temporal redundancy and provide a smoother generation domain (Chen et al., 2023; Geng et al., 2025; Guan et al., 2026), but compression alone does not determine how contact information should pass among streams. Latent regression also leaves contact, relations in the object frame, and interaction timing unconstrained after decoding. More importantly, stream decomposition exposes natural conditional relations: any one component can be inferred from others. Conditional HOI methods show that one motion stream can be inferred from others (Li et al., 2023; Zhang et al., 2025). However, a multi-stream model supporting both complete HOI generation and completion of any missing stream, within a shared temporal model remains underexplored. Separate latent states need not imply independent generative dynamics. The distinction between *state separation* and *dynamic coupling* is essential. Body, object, and hand streams can retain specialized states and outputs, while each velocity depends on the complete current HOI state. Therefore, the model is joint in the information available to each update, but separate in what it represents and produces. When trajectories drift, every stream can respond to the current mismatch rather than only to static text and geometry conditions. Coordination is learned throughout generation instead of being imposed after motion is produced. Decomposing HOI into three streams naturally defines three completion tasks, each recovering one missing stream from the other two. These conditional relations define three single-missing-stream completion tasks within the same model. Features capable of predicting any missing stream are also encouraged to encode the action, manipulated object, and temporal progression, providing a natural semantic interface for understanding.

We introduce **TRACE**, short for **T**ri-stream **R**epresentation **A**nd Stream **C**ompletion for Human-Object Interaction Mod**E**ling, a coupled continuous latent framework for text-conditioned Human-Object Interaction Generation, stream completion, and HOI understanding. As illustrated in Figure 1, TRACE generates diverse human-object interactions from text and object geometry while supporting stream completion and HOI understanding. TRACE follows a separation-and-coupling principle. A split temporal VAE (Kingma & Welling, 2013) maps body, hand, and object motion into separate but temporally aligned latents, preserving their distinct structures while reducing sequence length. These streams are not generated independently: a coupled latent flow (Lipman et al., 2023; Liu et al., 2023) predicts every stream-specific velocity from the complete evolving HOI state, allowing their updates to respond to one another throughout generation. To ground latent transport in interaction geometry, clean latent estimates are decoded during training and supervised for contact, object-frame anchoring, and interaction timing. The same representation and flow backbone support Text-to-HOI generation and completion of any missing stream, while frozen coupled-flow states provide continuous memory tokens to a language model for HOI understanding.

Our contributions can be summarized into three parts:

- We introduce TRACE, a novel coupled continuous latent framework for text-conditioned Human-Object Interaction Generation. The coupled flow preserves separate body, object, and hand states, while each velocity depends on the complete current HOI state. Decoded geometric losses further constrain contact and object-relative motion over time.

- We reuse the coupled representation for completion and understanding. Completion of each missing stream is jointly trained with Text-to-HOI, while frozen flow features are bridged to a language model for HOI understanding.
- Experiments on three datasets evaluate generation, stream completion, and HOI understanding. Ablations quantify the benefits of stream coupling and decoded geometric supervision for contact consistency, along with the effect of the latent representation.

2 Related Work

Text-driven HOI Generation. Text-driven HOI generation has progressed from hand-object synthesis to dynamic whole-body interaction. Text2HOI generates detailed hand-object motion, while InterDiff predicts future human-object interactions from motion history (Cha et al., 2024; Xu et al., 2023). Recent diffusion-based methods extend language control to body motion and dynamic objects through initial-state and waypoint conditioning, physics-informed correction, contact or affordance guidance, and relation-aware generation (Li et al., 2024a; Peng et al., 2025; Diller & Dai, 2024; Wu et al., 2025a; Song et al., 2024). Other methods broaden the task toward zero-shot synthesis, long-horizon instruction following with physics tracking, and generation beyond the training domain (Xu et al., 2024; Wu et al., 2025b; Zhang et al., 2024). More recent work further targets geometric fidelity, long-sequence consistency, and modality-aware generation (Xue et al., 2025; Zeng et al., 2025; Geng et al., 2025; Wang et al., 2026b), reflecting growing attention to both interaction realism and controllability under increasingly complex human-object dynamics. These advances increasingly emphasize accurate coordination among articulated human motion, object trajectories, and contact events rather than treating object motion as a secondary condition. Despite this progress, synthesizing text-aligned body, object, and hand motion that remains spatially and temporally coherent throughout the interaction remains challenging.

Motion Representation Learning. Motion generation has explored representations for preserving temporal structure and dependencies among motion components. Full-sequence representations jointly encode synchronized human-object trajectories (Xu et al., 2023; Li et al., 2024a), whereas factorized representations separate body, object, hand, or contact components (Cha et al., 2024; Peng et al., 2025; Song et al., 2024; Wang et al., 2026b). Interaction representations introduce explicit relations through contact variables, geometric fields, or kinematic chains (Diller & Dai, 2024; Wu et al., 2025a; Xue et al., 2025; Zeng et al., 2025). Latent representations compress motion into compact generation spaces. MLD uses continuous VAE latents (Chen et al., 2023), ARDHOI predicts continuous HOI tokens (Geng et al., 2025), and FlowCoMotion combines discrete tokens with continuous latents for text-conditioned flow (Guan et al., 2026). Conditional modeling treats observed motion as input. OMOMO and DiffGrasp condition human motion on object trajectories, showing that partial observations constrain missing interaction dynamics (Li et al., 2023; Zhang et al., 2025). These formulations balance compactness, component specialization, and motion coordination. TriDi (Petrov et al., 2025) models static human-object distributions with one diffusion transformer, while Uni-HOI (Zhang et al., 2026) and HOIGPT (Huang et al., 2025) use discrete representations for joint or bidirectional generation. Unlike LIGHT's asynchronous denoising, TRACE couples temporal body, object, and hand streams through continuous latent flow, with each velocity depending on the complete current state.

Motion-Language Understanding. Motion-language understanding studies mappings between motion and text, including retrieval, understanding, and generation. Token-based methods discretize motion into language-like units for reciprocal translation and instruction following (Guo et al., 2022b; Jiang et al., 2023). Alignment-based methods learn shared motion-text spaces (Li et al., 2024b), while MotionGPT3 (Zhu et al., 2026) connects continuous motion latents to language parameters for bidirectional generation. Interaction-centered models extend this interface to human-scene and human-object interactions. HSI-GPT and HSI-GPT2 jointly model scene, motion, and language (Wang et al., 2025; 2026a), while Uni-HOI models a joint distribution over text, human, and object motion (Zhang et al., 2026). HOIGPT (Huang et al., 2025) represents hand-object sequences with discrete codes for bidirectional generation. These studies suggest that richer motion representations can provide language models with more explicit interaction semantics. In particular, representations learned from generation may encode contact evolution, object-relative motion, and coordinated body-hand dynamics that are directly useful for semantic description. Whether frozen human-object generation features retain reusable interaction semantics therefore remains underexplored.

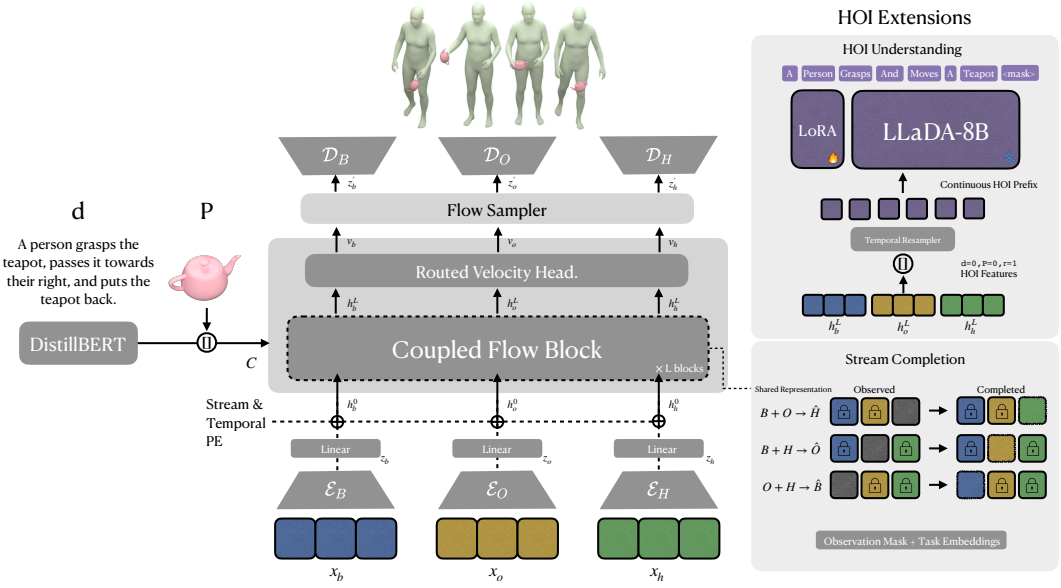

Figure 2: **Overview of TRACE.** *Left: HOI Generation.* We form different modalities, e.g., body, hand, and object, each compressed into latent vectors by its own encoder. After adding stream and temporal embeddings, a coupled flow transformer predicts their velocities, which are integrated by a flow sampler. Then, different modalities' decoders reconstruct them into clean motion. *Right: HOI Extensions.* Leveraging its HOI representation, we enable HOI understanding via language model fine-tuning. Joint training with three completion paths, each targeting one stream, adds completion capability while improving HOI Generation.

3 METHODOLOGY

Overview. We formulate HOI generation as synthesizing a T-frame 3D motion sequence in which a human interacts with an object. As shown in Figure 2, the input consists of a text description $\boldsymbol{d}$ and canonical object geometry $\boldsymbol{p}$ (Prokudin et al., 2019), and the output is a synchronized sequence $\boldsymbol{x} = (\boldsymbol{x}_b, \boldsymbol{x}_h, \boldsymbol{x}_o)$ of body, hand, and object streams. Following SMPL-H (Loper et al., 2015; Romero et al., 2017), $\boldsymbol{x}_b \in \mathbb{R}^{T\times 22\times 3}$ contains 22 body joint positions, while $\boldsymbol{x}_h \in \mathbb{R}^{T\times 120}$ contains 30 hand joint positions and rotations. The object stream $\boldsymbol{x}_o \in \mathbb{R}^{T\times 9}$ contains a 6D rotation (Zhou et al., 2019) and 3D translation. TRACE encodes the three streams into separate continuous latents, jointly evolves them through a coupled flow, and decodes them into motion. It also completes any missing stream $\hat{x}_m$, $m \in \{b, o, h\}$, from $\boldsymbol{d}$, $\boldsymbol{p}$, and the observed streams $\boldsymbol{x}_{-m}$. For HOI understanding, frozen flow hidden states are used to fine-tune a language model for interaction modeling.

HOI Representation. Following Cha et al. (2024), we retain body, hand, and object motion as three explicit streams with distinct properties: articulated body motion, rigid object trajectories, and fine-grained hand movements. Separate encoders avoid forcing heterogeneous streams into one feature distribution, while temporal alignment preserves the frame-wise correspondence needed to recover contact and relative motion after decoding. The three streams share the same temporally aligned latent grid. The temporal VAE reduces its length by a factor s. We therefore construct a split temporal variational autoencoder (VAE) (Kingma & Welling, 2013), with a separate encoder and decoder $(\mathcal{E}_m, \mathcal{D}_m)$ for each stream $m \in \{b, o, h\}$. The encoder compresses the original motion $\boldsymbol{x}_m \in \mathbb{R}^{t\times f_m}$ into a latent vector $\boldsymbol{z}_m \in \mathbb{R}^{l\times c_m}$, where f_m and c_m denote the input and latent dimensions of the stream, respectively, and $l = \lceil t/s \rceil$ with temporal downsampling factor s. The latent encoding and reconstruction are defined as:

$$q_m(\boldsymbol{z}_m \mid \boldsymbol{x}_m) = \mathcal{N}(\boldsymbol{\mu}_m(\boldsymbol{x}_m), \boldsymbol{\Sigma}_m(\boldsymbol{x}_m)), \qquad \widehat{\boldsymbol{x}}_m = \mathcal{D}_m(\boldsymbol{z}_m). \tag{1}$$

Here, $\boldsymbol{\mu}_m(\boldsymbol{x}_m)$ and $\boldsymbol{\Sigma}_m(\boldsymbol{x}_m)$ denote the mean and diagonal covariance predicted by $\mathcal{E}_m$, and $\widehat{\boldsymbol{x}}_m$ denotes the reconstructed motion produced by $\mathcal{D}_m$. To preserve the motion characteristics of each stream after temporal compression, we train the split VAE with the following objective: $\mathcal{L}_{\text{vae}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{reg}}$. $\mathcal{L}_{\text{rec}}$ denotes masked reconstruction over valid frames. The regularization term $\mathcal{L}_{\text{reg}}$ aims to promote reasonable Human-Object motion reconstruction, and it consists of four components. Specifically, $\mathcal{L}_{\text{reg}}$ combines object pose reconstruction, temporal velocity consistency, body kinematic consistency, and weak KL regularization, complementing masked reconstruction

without overwhelming the continuous latent space. Full details are provided in Appendix A.1. After training, we freeze the VAE and the latent space produced by $\mathcal{E}_m$. Each mean is standardized independently for each channel as $\widetilde{\boldsymbol{z}}_m = (\boldsymbol{\mu}_m(\boldsymbol{x}_m) - \overline{\boldsymbol{\mu}}_m)/\overline{\boldsymbol{\sigma}}_m$, where $\overline{\boldsymbol{\mu}}_m$ and $\overline{\boldsymbol{\sigma}}_m$ denote the channel normalization parameters of the latent space. The resulting $\widetilde{\boldsymbol{z}}_m$ is then used as the target latent vector for flow matching modeling.

Latent Flow Modeling. Given the latent vectors $\widetilde{\boldsymbol{z}}_m$ produced by the frozen VAE, we learn a direct flow from independent Gaussian sources to the conditional distribution of HOI latents of the three streams, following flow matching (Lipman et al., 2023; Liu et al., 2023). For each stream $m \in \{b, o, h\}$, we independently sample $\boldsymbol{z}_{m,0} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, while sharing the flow time $\tau \sim \mathcal{U}(0, 1)$ across the three streams. The intermediate state and its target velocity along a linear path are defined as:

$$\boldsymbol{z}_{m,\tau} = (1-\tau)\boldsymbol{z}_{m,0} + \tau\widetilde{\boldsymbol{z}}_m, \qquad \boldsymbol{u}_{m,\tau} = \widetilde{\boldsymbol{z}}_m - \boldsymbol{z}_{m,0}. \tag{2}$$

Let $\boldsymbol{z}_\tau = (\boldsymbol{z}_{b,\tau}, \boldsymbol{z}_{o,\tau}, \boldsymbol{z}_{h,\tau})$ denote the joint state; conditioned on $(\boldsymbol{z}_\tau, \tau, \boldsymbol{d}, \boldsymbol{p})$, the model predicts one velocity for each stream and trains them with:

$$\mathcal{L}_{\text{fm}} = \sum_{m\in\{b,o,h\}} \frac{\|\boldsymbol{M} \odot (\boldsymbol{v}_{\theta,m} - \boldsymbol{u}_{m,\tau})\|_2^2}{c_m \|\boldsymbol{M}\|_1}. \tag{3}$$

Here, $\boldsymbol{M}$ is the temporal validity mask, broadcast across latent channels, and the denominator normalizes each stream by its valid latent elements.

Flow matching supervises latent velocity but not decoded interaction geometry. At time τ, the predicted velocity yields a clean latent estimate $\widehat{\boldsymbol{z}}_{m,1} = \boldsymbol{z}_{m,\tau} + (1-\tau)\boldsymbol{v}_{\theta,m}$. The frozen VAE decodes the three latents while propagating gradients from decoded motion to the flow model. We further impose two interaction objectives: $\mathcal{L}_{\text{c}}$ supervises human-object contact with BCE and Dice losses (Cha et al., 2024; Dwivedi et al., 2025), while $\mathcal{L}_{\text{A/E}}$ preserves object anchoring and temporal consistency. The contact term supervises whether interaction occurs, whereas the anchoring and episode terms constrain where it occurs in the object coordinate frame and how the relation evolves over time. The final objective is:

$$\mathcal{L}_{\text{flow}} = \mathcal{L}_{\text{fm}} + \lambda_{\text{c}}\mathcal{L}_{\text{c}} + \mathcal{L}_{\text{A/E}}. \tag{4}$$

It jointly supervises latent transport and decoded interaction geometry. Full details are provided in Appendix A.1. These constraints improve motion quality and interaction consistency (Figure 3 and Table 1).

Coupled Flow. Separating latent states does not require factorized transport dynamics. For $m \in \{b, o, h\}$, an independent stream flow evolves each latent only from its own current state:

$$\frac{\mathrm{d}\boldsymbol{z}_{m,\tau}}{\mathrm{d}\tau} = \boldsymbol{v}_{\theta,m}(\boldsymbol{z}_{m,\tau}, \tau, \boldsymbol{d}, \boldsymbol{p}), \qquad \frac{\partial \boldsymbol{v}_{\theta,m}}{\partial \boldsymbol{z}_{n,\tau}} = \mathbf{0}, \qquad n \neq m. \tag{5}$$

Although shared textual and geometric conditions specify the interaction, they do not let one stream's velocity depend on other current states. Under independent Gaussian initialization, this induces conditionally factorized dynamics without a mechanism to reconcile body, object, and hand trajectories during transport. Thus, an object trajectory drifting from the hands cannot alter the body or hand updates, even though all streams share the same text and geometry conditions. We therefore couple the dynamics while preserving separate latent representations. TRACE evolves the separate stream latents through a single flow over the latent space $\mathcal{Z} = \mathcal{Z}_b \times \mathcal{Z}_o \times \mathcal{Z}_h$. Let $\boldsymbol{z}_\tau = (\boldsymbol{z}_{b,\tau}, \boldsymbol{z}_{o,\tau}, \boldsymbol{z}_{h,\tau})$ denote the complete HOI state. The coupled dynamics and responses between streams are:

$$\frac{\mathrm{d}\boldsymbol{z}_{m,\tau}}{\mathrm{d}\tau} = \boldsymbol{v}_{\theta,m}(\boldsymbol{z}_\tau, \tau, \boldsymbol{d}, \boldsymbol{p}), \qquad \boldsymbol{J}_{\theta,mn}(\tau) = \frac{\partial \boldsymbol{v}_{\theta,m}}{\partial \boldsymbol{z}_{n,\tau}}, \qquad m, n \in \{b, o, h\}. \tag{6}$$

Here, $\boldsymbol{J}_{\theta,mn}$ measures how stream m responds to stream n. Off-diagonal blocks need not vanish, enabling dependencies between streams. The vector field is joint in input but partitioned in output: each modality retains separate states and velocities while trajectories evolve jointly during transport. This joint input and separated output design couples transport without discarding the latent coordinates of each stream, allowing every frozen decoder to retain the motion statistics learned during VAE pretraining. Empirically, coupled flow improves motion quality and interaction consistency (Figure 5 and Table 3).

Architecture. The left of Figure 2 illustrates the coupled flow transformer parameterizing $\boldsymbol{v}_\theta$. (i) *Latent vectors.* Each modal latent is linearly projected into a shared hidden dimension, then augmented with a learned stream embedding and sinusoidal temporal encoding: $\boldsymbol{h}_m^0 = \phi_m(\boldsymbol{z}_{m,\tau}) + \boldsymbol{e}_m^{\mathrm{s}} + \boldsymbol{e}^{\mathrm{t}}$, $m \in \{b, o, h\}$. (ii) *Flow conditioning.* The shared flow time τ is sinusoidally embedded and combined with the text and geometry condition to modulate attention and feed-forward layers via AdaLN (Peebles & Xie, 2023). (iii) *Text and geometry conditions.* The description $\boldsymbol{d}$ is encoded by frozen DistilBERT (Sanh et al., 2019), while object geometry $\boldsymbol{p}$ is represented by BPS (Prokudin et al., 2019) and projected into a geometry token. They are concatenated and injected into each stream via cross-attention in every transformer block. (iv) *Joint attention.* Joint attention realizes the responses between streams $\boldsymbol{J}_{\theta,mn}$ required by the coupled flow. In each block r, the body, object, and hand streams form queries, keys, and values through separate projections. These features are concatenated for joint attention, split into individual streams, and processed by separate output projections and feed-forward networks:

$$\boldsymbol{h}_m^{r+1} = \mathcal{B}_m^r\left(\boldsymbol{h}_b^r, \boldsymbol{h}_o^r, \boldsymbol{h}_h^r; \tau, \boldsymbol{d}, \boldsymbol{p}\right), \qquad m \in \{b, o, h\}. \tag{7}$$

Here, $\mathcal{B}_m^r$ denotes the update of stream m. Since each state depends on all streams, joint attention establishes a differentiable path from $\boldsymbol{z}_{n,\tau}$ to $\boldsymbol{v}_{\theta,m}$, enabling nonzero off-diagonal responses $\boldsymbol{J}_{\theta,mn}$. (v) *Routed velocity head.* The velocity head maps coupled states to body, object, and hand velocities using a top-k mixture-of-experts (Shazeer et al., 2017). We use a shared top-k router to produce a common mixture of stream-specific output projections for each interaction. A shared router predicts weights from the joint state and flow condition, while each expert uses stream-specific projections:

$$\boldsymbol{v}_{\theta,m} = \sum_{e \in \mathcal{K}(\boldsymbol{\pi})} \pi_e \, \mathcal{O}_{e,m}(\boldsymbol{h}_m), \qquad \mathcal{L}_{\mathrm{r}} = E \sum_{e=1}^{E} \left(\overline{\pi}_e - \frac{1}{E}\right)^2. \tag{8}$$

Here, $\mathcal{K}(\boldsymbol{\pi})$ denotes the selected experts, π_e the normalized routing weight, and $\overline{\pi}_e$ its batch average. Shared routing weights mix all streams, while $\mathcal{O}_{e,m}$ remains stream-specific. The routing loss prevents expert concentration, giving $\mathcal{L} = \mathcal{L}_{\text{flow}} + \lambda_{\mathrm{r}}\mathcal{L}_{\mathrm{r}}$.

Stream Completion. The coupled tri-stream representation allows TRACE to recover any missing stream. Unified conditional generators similarly share parameters across multimodal mappings (Li et al., 2025; Petrov et al., 2025), while partial HOI sequences have been used to condition motion completion (Huang et al., 2025). We jointly train standard HOI generation and three single-stream completion tasks. For a missing stream $m \in \{b, o, h\}$, $(\boldsymbol{x}_{-m}, \boldsymbol{d}, \boldsymbol{p}) \mapsto \widehat{\boldsymbol{x}}_m$. All four tasks share the same VAE, coupled flow backbone, and sampling head. Joint training adds completion capability without requiring separate generators.

For each sample, observation and task embeddings identify the available streams and target m. Observed latents use $\tau_n = 1$ for $n \neq m$, while the missing stream follows $\tau_m = \tau$. Joint attention infers the missing motion from the observed trajectories. To preserve the conditions, observed hidden states are restored after each block and their latents remain fixed throughout sampling. Joint training on all three completion tasks also improves HOI generation through stronger coupled flow coordination (Figure 4, Table 1, E, A and B).

HOI Understanding. For HOI understanding, we connect frozen coupled flow states to LLaDA and adapt it through LoRA (Hu et al., 2022) to describe a complete interaction. Given an interaction $\boldsymbol{x}$, the frozen VAE first maps its body, object, and hand streams into clean normalized latents. We read the frozen flow at $\tau = 1$ under zero conditioning, which is included during flow training through 0.1 condition dropout. The final hidden states before the velocity head form the HOI representation $\boldsymbol{h} = [\boldsymbol{h}_b, \boldsymbol{h}_o, \boldsymbol{h}_h]$. A learned readout embedding is added to each stream before the valid tokens are concatenated. A temporal slot resampler then compresses them into fixed memory tokens, followed by a projection into LLaDA's embedding space as a continuous prefix (Nie et al., 2025; You et al., 2026):

$$\widehat{\boldsymbol{d}} = \mathcal{F}_{\mathrm{LM}}\left(\boldsymbol{q}, \mathcal{P}(\mathcal{R}(\boldsymbol{h}))\right). \tag{9}$$

Here, $\mathcal{R}$ and $\mathcal{P}$ denote the temporal resampler and prefix projection, while $\boldsymbol{q}$ contains the object category and learning query. The language branch reads frozen flow features without altering generation or completion. Full details on fine-tuning the language model are provided in Section A.2 of the Appendix. Empirically, harnessing the interaction relations learned during generation improves HOI understanding over directly encoding raw motion (Table 2).

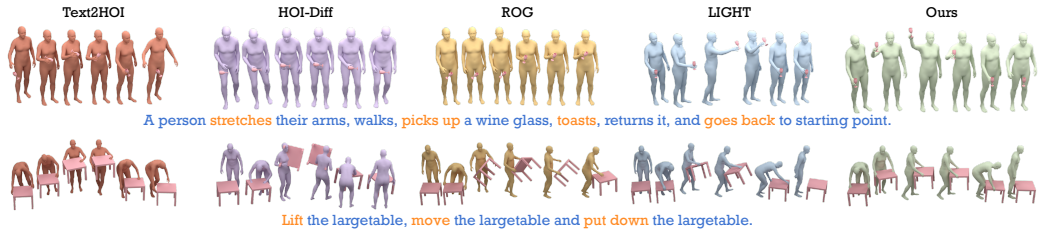

Figure 3: **Qualitative comparison** with baselines. Our method yields more realistic human-object interactions, fewer contact/penetration artifacts, more accurate finger positioning, and better text-motion alignment.

Table 1: **Quantitative comparisons on the InterAct dataset between our method and baseline approaches.** We report R-Precision with a batch size of 256.

Method	R-Precision ↑			FID ↓	MM Dist. ↓	Diversity →	FSR ↓	Pene. ↓	Contact →	Interaction ↑		
	Top 1	Top 2	Top 3							C_{prec}	C_{rec}	C_{F1}
Ground Truth	$0.600^{\pm0.004}$	$0.834^{\pm0.001}$	$0.909^{\pm0.001}$	$0.000^{\pm0.000}$	$1.475^{\pm0.003}$	$7.781^{\pm0.140}$	$0.083^{\pm0.204}$	$0.076^{\pm0.000}$	$0.208^{\pm0.000}$	$1.000^{\pm0.000}$	$1.000^{\pm0.000}$	$1.000^{\pm0.000}$
HOI-Diff	$0.413^{\pm0.010}$	$0.624^{\pm0.009}$	$0.740^{\pm0.011}$	$0.689^{\pm0.031}$	$3.029^{\pm0.007}$	$7.620^{\pm0.085}$	$0.072^{\pm0.160}$	$\mathbf{0.103}^{\pm0.010}$	$0.084^{\pm0.006}$	$0.722^{\pm0.002}$	$0.447^{\pm0.026}$	$0.501^{\pm0.022}$
CHOIS	$\underline{0.439}^{\pm0.005}$	$\underline{0.660}^{\pm0.002}$	$\underline{0.766}^{\pm0.003}$	$0.572^{\pm0.008}$	$2.781^{\pm0.017}$	$\underline{7.717}^{\pm0.115}$	$0.119^{\pm0.234}$	$0.131^{\pm0.013}$	$0.115^{\pm0.002}$	$0.710^{\pm0.012}$	$0.520^{\pm0.004}$	$0.541^{\pm0.007}$
InterDiff	$\mathbf{0.501}^{\pm0.009}$	$\mathbf{0.722}^{\pm0.003}$	$\mathbf{0.824}^{\pm0.007}$	$0.215^{\pm0.001}$	$2.461^{\pm0.013}$	$\underline{7.717}^{\pm0.115}$	$0.092^{\pm0.178}$	$0.116^{\pm0.011}$	$0.124^{\pm0.003}$	$0.715^{\pm0.000}$	$0.562^{\pm0.013}$	$0.584^{\pm0.003}$
Text2HOI	$0.428^{\pm0.003}$	$0.637^{\pm0.000}$	$0.745^{\pm0.000}$	$0.331^{\pm0.003}$	$2.665^{\pm0.002}$	$7.631^{\pm0.000}$	$\underline{0.055}^{\pm0.123}$	$\underline{0.105}^{\pm0.003}$	$0.102^{\pm0.001}$	$0.711^{\pm0.001}$	$0.488^{\pm0.000}$	$0.532^{\pm0.002}$
ROG	$0.402^{\pm0.007}$	$0.613^{\pm0.008}$	$0.739^{\pm0.009}$	$0.326^{\pm0.021}$	$2.584^{\pm0.018}$	$\mathbf{7.758}^{\pm0.076}$	$0.058^{\pm0.147}$	$0.107^{\pm0.005}$	$0.168^{\pm0.003}$	$0.786^{\pm0.005}$	$0.646^{\pm0.009}$	$0.667^{\pm0.007}$
LIGHT	$0.421^{\pm0.014}$	$0.637^{\pm0.016}$	$0.754^{\pm0.016}$	$\mathbf{0.148}^{\pm0.014}$	$2.756^{\pm0.018}$	$7.712^{\pm0.050}$	$0.078^{\pm0.197}$	$0.132^{\pm0.001}$	$0.132^{\pm0.001}$	$0.731^{\pm0.002}$	$0.615^{\pm0.016}$	$0.627^{\pm0.010}$
TRACE	$0.404^{\pm0.003}$	$0.615^{\pm0.007}$	$0.744^{\pm0.005}$	$0.199^{\pm0.015}$	$\mathbf{2.256}^{\pm0.006}$	$7.927^{\pm0.051}$	$\mathbf{0.047}^{\pm0.146}$	$0.117^{\pm0.001}$	$\underline{0.196}^{\pm0.001}$	$\underline{0.830}^{\pm0.001}$	$\underline{0.810}^{\pm0.002}$	$\underline{0.799}^{\pm0.004}$
TRACE^{joint}	$0.407^{\pm0.003}$	$0.619^{\pm0.008}$	$0.755^{\pm0.001}$	$\underline{0.188}^{\pm0.007}$	$\underline{2.262}^{\pm0.008}$	$7.911^{\pm0.063}$	$\mathbf{0.047}^{\pm0.148}$	$0.111^{\pm0.001}$	$\mathbf{0.198}^{\pm0.001}$	$\mathbf{0.835}^{\pm0.001}$	$\mathbf{0.811}^{\pm0.003}$	$\mathbf{0.806}^{\pm0.003}$

4 EXPERIMENTS

Dataset. We mainly experiment on InterAct (Xu et al., 2025), with ablations on its major subsets BEHAVE (Bhatnagar et al., 2022) and OMOMO (Li et al., 2023). InterAct provides fine-grained textual annotations, and we follow its official train-test split. Since the data use both SMPL-H (Romero et al., 2017) and SMPL-X (Pavlakos et al., 2019), we convert all sequences to SMPL-H using the official conversion and consistently use SMPL-H joints.

Metrics. Following Guo et al. (2022a), we repeat evaluation 20 times and report mean ± 95% confidence intervals. We evaluate HOI realism, diversity, text alignment, and physical plausibility. **FID** measures distribution similarity, **Diversity** measures motion variation, and **R-Precision** and **MM Dist** assess motion-text alignment. Following LIGHT (Wang et al., 2026b), we report Foot Skating Ratio (**FSR**), Penetration Ratio (**Pene**), Contact Ratio (**Contact**), and frame-wise contact precision (C_{prec}), recall (C_{rec}), and F1 (C_{F1}). These complementary criteria capture distribution-level fidelity and interaction-level correctness, which is essential because visually plausible motions may still exhibit weak semantic alignment or physically implausible contact. We therefore report all metrics under the same evaluator and test split, enabling consistent comparisons across generation, completion, and understanding settings in our experiments. For HOI understanding, following Jiang et al. (2023), we use **Bleu@1/4**, **Rouge**, **Cider**, and **BertScore** to evaluate generated descriptions, together with **R-Precision** and **MM Dist** for motion-text alignment.

Implementation Details. Object geometry is represented by a 1024-point BPS (Prokudin et al., 2019). The split VAE uses temporal downsampling $s = 4$ and latent dimensions $(c_b, c_o, c_h) = (128, 64, 128)$. The coupled-flow backbone is a 12 layer transformer with hidden dimension 1024 and 16 heads. The velocity head uses 4 experts with top-2 routing. Models are trained on one NVIDIA A800 GPU with BF16. The language branch uses a two-layer temporal resampler with 32 memory tokens and rank 8 LoRA. Training uses AdamW with cosine learning-rate decay and gradient clipping for stable optimization of the coupled-flow backbone. We keep the VAE frozen during flow training and optimize the flow model jointly. For understanding, only the resampler, LoRA parameters, and output projection are updated, leaving the generator unchanged. Additional loss weights and schedules are provided in the Appendix A.1.

Baselines. We compare against six recent HOI generation methods. HOI-Diff (Peng et al., 2025) uses transformer-based diffusion with affordance-guided sampling. CHOIS (Li et al., 2024a) is adapted to text-only conditioning by removing state and waypoint inputs. InterDiff (Xu et al., 2023) replaces its motion-history encoder with a text encoder. Text2HOI (Cha et al., 2024) is extended to full-body HOI while retaining its static contact-map estimator. ROG (Xue et al., 2025) combines motion diffusion with relation guidance and is retrained on InterAct. LIGHT (Wang et al., 2026b) follows its original setting. All methods use the same preprocessing, split, and evaluator, while adapted baselines retain their original objectives and training schedules. For HOI understanding, we compare with HOIGPT (Huang et al., 2025), TM2T (Guo et al., 2022b), MotionGPT (Jiang et al., 2023), and

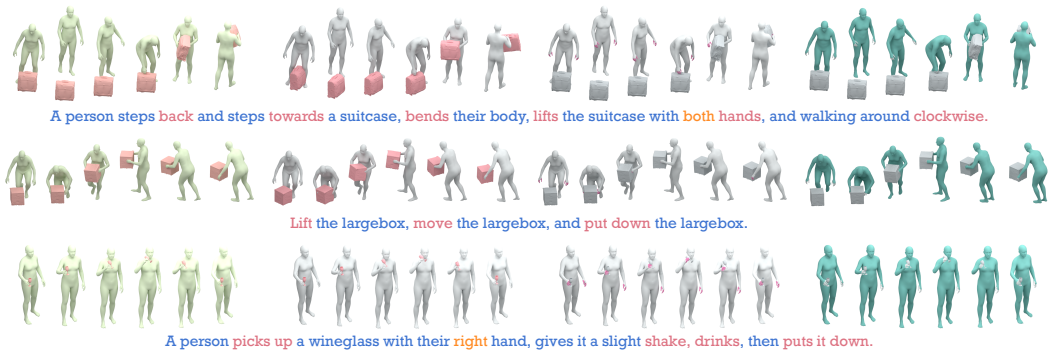

Figure 4: **Qualitative results of stream completion.** From left to right, we show the ground-truth interaction and the completion of object, hand, and body streams. Gray denotes observed motion, while pink, magenta, and teal denote the generated object, hands, and body, respectively.

Table 2: **Comparison of HOI understanding on InterAct.** Our model is trained for 100 epochs.

Methods	R-Precision ↑			MM Dist. ↓	Bleu@1 ↑	Bleu@4 ↑	Rouge ↑	Cider ↑	BertScore ↑
	Top 1	Top 2	Top 3						
Real	0.592	0.757	0.853	1.486	–	–	–	–	–
TM2T	0.206	0.351	0.466	3.342	35.18	24.12	56.17	2.61	52.16
MotionGPT	0.224	0.391	0.504	3.086	39.86	29.63	60.84	3.42	57.83
LaMPM2T	0.243	0.415	0.532	2.894	44.37	34.42	66.14	4.09	63.18
HOIGPT	0.256	**0.452**	**0.570**	2.756	**50.31**	39.22	**71.08**	4.89	**69.05**
TRACE ^{Raw}	0.232	0.408	0.512	3.020	41.57	30.82	62.24	3.55	58.81
TRACE	**0.260**	0.449	0.567	**2.742**	49.87	**39.54**	70.18	**4.95**	67.47

LaMPM2T (Li et al., 2024b). HOIGPT is extended with a body stream, while other baselines only adapt their motion encoders to our datasets. All models are retrained on InterAct with their original understanding architectures and objectives. TRACE Raw replaces the frozen VAE-flow encoder with a frame projection of normalized motion, followed by the same temporal resampler and prefix projection. Both variants share object prompts, LLaDA/LoRA configurations, language objectives, and training schedules, isolating frozen coupled-flow features from raw-motion encoding.

Quantitative Evaluation. As shown in Table 1, TRACE achieves the strongest interaction accuracy: joint training produces the highest contact precision, recall, and F1, while our variants yield the lowest MM Dist and FSR. Table 2 shows that TRACE performs best on R@1, MM Dist, BLEU@4, and CIDEr for HOI understanding, with coupled-flow features outperforming raw-motion encoding on all metrics. Tables A, B, C, and D report generation and understanding on OMOMO and BEHAVE, confirming the benefits of joint training and flow-based HOI representations.

Qualitative Evaluation. As shown in Figure 3, TRACE better follows descriptions while maintaining coordinated human-object interaction and stable contact, whereas baselines often omit actions or exhibit object drift, penetration, and floating artifacts. Figure A in the Appendix further shows the advantage of the complete tri-stream representation for HOI understanding.

Stream Completion. As shown in Figure 4, TRACE completes a missing body, object, or hand stream from the other two observed streams. For body completion, the generated posture follows the fixed hand-object trajectories and preserves their spatial constraints over time. Object completion recovers a rigid trajectory consistent with the observed human motion, while hand completion aligns fine-grained movements with the body and object. In all three settings, the missing component is inferred from cross-stream dependencies rather than generated independently. The completed component remains coordinated with the observed trajectories, while the observed motion is preserved. Table E in the appendix reports stream-specific reconstruction and physical metrics on OMOMO, BEHAVE, and InterAct.

Impact of Stream Modeling. We compare unified, independent, and coupled stream modeling. As shown in Table 3, the coupled model performs best in R-Precision, FID, MM Dist, FSR, Contact, and all three interaction metrics. These results show that neither merging the streams nor evolving them independently captures their interaction dependencies as effectively as coupled flow. Figure 5 further shows that the alternative strategies often produce object drift or misaligned hand-object contact,

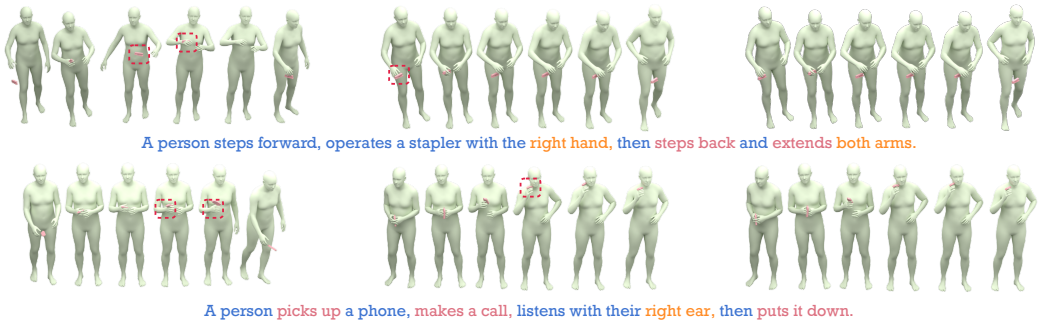

Figure 5: **Qualitative comparison of stream modeling.** We compare unified, independent, and coupled stream modeling. Red boxes highlight object drift and hand-object misalignment in the first two variants, while coupled modeling yields coherent interactions.

whereas coupled modeling maintains coordinated trajectories throughout the interaction. Tables F and G in the Appendix exhibit the same overall trend on OMOMO and BEHAVE.

Table 3: **Ablation study of stream modeling on InterAct.** We report R-Precision with a batch size of 256.

Strategy	R-Precision ↑			FID ↓	MM Dist. ↓	Diversity →	FSR ↓	Pene. ↓	Contact →	Interaction ↑		
	Top 1	Top 2	Top 3							C_{prec}	C_{rec}	C_{F1}
GT	$0.600^{\pm 0.004}$	$0.834^{\pm 0.001}$	$0.909^{\pm 0.001}$	$0.000^{\pm 0.000}$	$1.475^{\pm 0.003}$	$7.781^{\pm 0.140}$	$0.083^{\pm 0.204}$	$0.076^{\pm 0.000}$	$0.208^{\pm 0.000}$	$1.000^{\pm 0.000}$	$1.000^{\pm 0.000}$	$1.000^{\pm 0.000}$
Independent	$0.383^{\pm 0.005}$	$0.584^{\pm 0.007}$	$0.710^{\pm 0.005}$	$0.261^{\pm 0.017}$	$2.447^{\pm 0.008}$	$8.003^{\pm 0.056}$	$0.052^{\pm 0.109}$	$\mathbf{0.103}^{\pm 0.001}$	$0.146^{\pm 0.002}$	$0.756^{\pm 0.003}$	$0.579^{\pm 0.003}$	$0.606^{\pm 0.003}$
Unified	$0.400^{\pm 0.004}$	$0.609^{\pm 0.006}$	$0.737^{\pm 0.004}$	$0.213^{\pm 0.016}$	$2.306^{\pm 0.007}$	$7.896^{\pm 0.054}$	$0.059^{\pm 0.133}$	$0.130^{\pm 0.001}$	$0.186^{\pm 0.002}$	$0.818^{\pm 0.002}$	$0.764^{\pm 0.002}$	$0.760^{\pm 0.002}$
Coupled	$\mathbf{0.404}^{\pm 0.003}$	$\mathbf{0.615}^{\pm 0.007}$	$\mathbf{0.744}^{\pm 0.005}$	$\mathbf{0.199}^{\pm 0.015}$	$\mathbf{2.256}^{\pm 0.006}$	$7.927^{\pm 0.051}$	$\mathbf{0.047}^{\pm 0.146}$	$0.117^{\pm 0.001}$	$\mathbf{0.196}^{\pm 0.001}$	$\mathbf{0.830}^{\pm 0.001}$	$\mathbf{0.810}^{\pm 0.002}$	$\mathbf{0.799}^{\pm 0.004}$

Table 4: **Ablation study of latent representation on InterAct.** We report R-Precision with a batch size of 256.

Space	R-Precision ↑			FID ↓	MM Dist. ↓	Diversity →	FSR ↓	Pene. ↓	Contact →	Interaction ↑		
	Top 1	Top 2	Top 3							C_{prec}	C_{rec}	C_{F1}
GT	$0.600^{\pm 0.004}$	$0.834^{\pm 0.001}$	$0.909^{\pm 0.001}$	$0.000^{\pm 0.000}$	$1.475^{\pm 0.003}$	$7.781^{\pm 0.140}$	$0.083^{\pm 0.204}$	$0.076^{\pm 0.000}$	$0.208^{\pm 0.000}$	$1.000^{\pm 0.000}$	$1.000^{\pm 0.000}$	$1.000^{\pm 0.000}$
Raw	$0.399^{\pm 0.003}$	$0.607^{\pm 0.006}$	$0.735^{\pm 0.004}$	$0.228^{\pm 0.014}$	$2.314^{\pm 0.007}$	$7.918^{\pm 0.052}$	$0.055^{\pm 0.132}$	$0.135^{\pm 0.001}$	$0.182^{\pm 0.002}$	$0.809^{\pm 0.002}$	$0.750^{\pm 0.003}$	$0.748^{\pm 0.003}$
VQ	$0.384^{\pm 0.004}$	$0.586^{\pm 0.006}$	$0.711^{\pm 0.004}$	$0.269^{\pm 0.016}$	$2.406^{\pm 0.007}$	$7.884^{\pm 0.053}$	$0.063^{\pm 0.001}$	$0.146^{\pm 0.184}$	$0.158^{\pm 0.002}$	$0.771^{\pm 0.002}$	$0.681^{\pm 0.003}$	$0.685^{\pm 0.003}$
VAE	$\mathbf{0.404}^{\pm 0.003}$	$\mathbf{0.615}^{\pm 0.007}$	$\mathbf{0.744}^{\pm 0.005}$	$\mathbf{0.199}^{\pm 0.015}$	$\mathbf{2.256}^{\pm 0.006}$	$7.927^{\pm 0.051}$	$\mathbf{0.047}^{\pm 0.146}$	$\mathbf{0.117}^{\pm 0.001}$	$\mathbf{0.196}^{\pm 0.001}$	$\mathbf{0.830}^{\pm 0.001}$	$\mathbf{0.810}^{\pm 0.002}$	$\mathbf{0.799}^{\pm 0.004}$

Impact of Latent Representation. We compare raw motion, discrete VQ latents, and continuous VAE latents under the same coupled-flow architecture. Table 4 shows that the VAE achieves the best R-Precision, FID, MM Dist, FSR, penetration, contact, and interaction metrics. Raw motion also outperforms VQ on most metrics, suggesting that discrete quantization discards fine-grained interaction cues. By preserving continuous stream-specific information while reducing temporal redundancy, the VAE provides a more effective space for coupled flow modeling.

Impact of Geometric Constraints. Table 5 evaluates the geometric constraints applied to the decoded interaction. Adding $\mathcal{L}_{\text{c}}$ to the flow objective consistently improves R@3, FID, FSR, penetration, and contact accuracy. Further introducing $\mathcal{L}_{\text{A/E}}$ achieves the best result on every reported metric. This progression shows that contact supervision improves local alignment, while object-local anchoring and temporal consistency further stabilize interaction geometry over time.

Table 5: **Ablation study of geometric constraints on InterAct.** We report R-Precision with a batch size of 256.

$\mathcal{L}_{\text{c}}$	$\mathcal{L}_{\text{A/E}}$	R@3 ↑	FID ↓	FSR ↓	Pene. ↓	Contact →	Interaction ↑		
							C_{prec}	C_{rec}	C_{F1}
×	×	$0.736^{\pm 0.004}$	$0.220^{\pm 0.016}$	$0.053^{\pm 0.138}$	$0.137^{\pm 0.002}$	$0.173^{\pm 0.002}$	$0.796^{\pm 0.003}$	$0.708^{\pm 0.004}$	$0.715^{\pm 0.004}$
✓	×	$0.740^{\pm 0.003}$	$0.208^{\pm 0.015}$	$0.050^{\pm 0.104}$	$0.126^{\pm 0.001}$	$0.187^{\pm 0.002}$	$0.814^{\pm 0.002}$	$0.770^{\pm 0.003}$	$0.765^{\pm 0.003}$
✓	✓	$\mathbf{0.744}^{\pm 0.005}$	$\mathbf{0.199}^{\pm 0.015}$	$\mathbf{0.047}^{\pm 0.146}$	$\mathbf{0.117}^{\pm 0.001}$	$\mathbf{0.196}^{\pm 0.001}$	$\mathbf{0.830}^{\pm 0.001}$	$\mathbf{0.810}^{\pm 0.002}$	$\mathbf{0.799}^{\pm 0.004}$

5 Conclusion

We introduced TRACE, a structured continuous-latent framework for Text-to-HOI generation. TRACE preserves separate body, object, and hand states while conditioning each stream-specific velocity on the complete evolving HOI state; decoded-space geometric constraints further ground latent generation in contact locations, object-local relations, and interaction timing. The same generator handles Text-to-HOI and three stream completion tasks, while frozen interaction features support HOI understanding through a language bridge. Experiments across three datasets demonstrate generation and interaction performance, completion capability, and consistent understanding gains over raw-motion encoding. Together, these results show that structured cross-stream coupling provides a unified representation for coordinated generation, conditional reconstruction, and semantic interpretation across diverse human-object interaction tasks.

REFERENCES

Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2022.

Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1577–1585, 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 (CVPR)*, pp. 18000–18010, June 2023.

Christian Diller and Angela Dai. CG-HOI: Contact-guided 3d human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Sai Kumar Dwivedi, Dimitrije Antić, Shashank Tripathi, Omid Taheri, Cordelia Schmid, Michael J. Black, and Dimitrios Tzionas. Interactvlm: 3d interaction reasoning from 2d foundational models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22605–22615, June 2025.

Zichen Geng, Zeeshan Hayder, Wei Liu, and Ajmal Saeed Mian. Auto-regressive diffusion for generating 3d human-object interactions. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, 2025.

Dawei Guan, Di Yang, Chengjie Jin, and Jiangtao Wang. Flowcomotion: Text-to-motion generation via token-latent flow modeling. *arXiv preprint arXiv:2604.11083*, 2026.

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 (CVPR)*, pp. 5152–5161, June 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 *ECCV*, 2022b.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Mingzhen Huang, Fu-Jen Chu, Bugra Tekin, Kevin J Liang, Haoyu Ma, Weiyao Wang, Xingyu Chen, Pierre Gleize, Hongfei Xue, Siwei Lyu, Kris Kitani, Matt Feiszli, and Hao Tang. Hoigpt: Learning long sequence hand-object interaction with language models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, USA, 2025.

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*, 2023.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Trans. Graph.*, 42(6), 2023.

Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C. Karen Liu. Controllable human-object interaction synthesis. In *ECCV*, 2024a.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Zichun Liao, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. Omniflow: Any-to-any generation with multi-modal rectified flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13178–13188, June 2025.

Zhe Li, Weihao Yuan, Yisheng He, Lingteng Qiu, Shenhao Zhu, Xiaodong Gu, Weichao Shen, Yuan Dong, Zilong Dong, and Laurence T Yang. Lamp: Language-motion pretraining for motion generation, retrieval, and captioning. *arXiv preprint arXiv:2410.07093*, 2024b.

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

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 2015.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October 2023.

Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025.

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*, 2025.

Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4331–4340, 2019. doi: 10.1109/ICCV.2019.00443.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), November 2017.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Noam Shazeer, *Azalia Mirhoseini, *Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Wenfeng Song, Xinyu Zhang, Shuai Li, Yang Gao, Aimin Hao, Xia Hou, Chenglizhao Chen, Ning Li, and Hong Qin. Hoianimator: Generating text-prompt human-object animations using novel perceptive diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

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 (CVPR)*, 2025.

Yuan Wang, Xiang Li, Yali Li, Xuege Hou, and Shengjin Wang. Hsi-gpt2: A dual-granularity large motion reasoning model with diffusion refinement for human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026a.

Ziyin Wang, Sirui Xu, Chuan Guo, Bing Zhou, Jiangshan Gong, Jian Wang, Yu-Xiong Wang, and Liang-Yan Gui. Unleashing guidance without classifiers for human-object interaction animation. In *ICLR*, 2026b.

Qianyang Wu, Ye Shi, Xiaoshui Huang, Lan Xu, Jingyi Yu, and Jingya Wang. Thor: Text to human-object interaction diffusion via relation intervention. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, 2025a.

Zhen Wu, Jiaman Li, Pei Xu, and C. Karen Liu. Human-object interaction from human-level instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025b.

Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023.

Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liangyan Gui. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, Yu-Xiong Wang, and Liang-Yan Gui. Interact: Advancing large-scale versatile 3d human-object interaction generation. In *CVPR*, 2025.

Mengqing Xue, Yifei Liu, Ling Guo, Shaoli Huang, and Changxing Ding. Guiding human-object interactions with rich geometry and relations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22714–22723, 2025.

Zebin You, Shen Nie, Xiaolu Zhang, JUN ZHOU, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. Llada-v: Large language diffusion models with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10093–10105, 2026.

Ling-An Zeng, Guohong Huang, Yi-Lin Wei, Shengbo Gu, Yu-Ming Tang, Jingke Meng, and Wei-Shi Zheng. Chainhoi: Joint-based kinematic chain modeling for human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Mengfei Zhang, Jinlu Zhang, and Zhigang Tu. Uni-hoi: A unified framework for learning the joint distribution of text and human-object interaction. *arXiv preprint arXiv:2604.27491*, 2026.

Yixuan Zhang, Hui Yang, Chuanchen Luo, Junran Peng, Yuxi Wang, and Zhaoxiang Zhang. Ood-hoi: Text-driven 3d whole-body human-object interactions generation beyond training domains. *arXiv preprint arXiv:2411.18660*, 2024.

Yonghao Zhang, Qiang He, Yanguang Wan, Yinda Zhang, Xiaoming Deng, Cuixia Ma, and Hongan Wang. DiffGrasp: Whole-body grasping synthesis guided by object motion using a diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Yi Zhou, Connelly Barnes, Lu Jingwan, Yang Jimei, and Li Hao. On the continuity of rotation representations in neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 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. In *International Conference on Learning Representations*, 2026.

APPENDIX

In this Appendix, we provide additional details of the training objectives, completion training, and language adaptation in Sec. A, followed by additional experimental results in Sec. B.

LLM Usage. We employ large language models (LLMs), such as ChatGPT, to assist in polishing our paper. Specifically, LLMs are used to correct grammatical errors, refine word choice, and improve overall fluency. We do not use LLMs to formulate our methodology or run experiments.

A ADDITIONAL METHOD DETAILS

This section first presents the objectives and training procedure of the generation and completion model, followed by details of the language adaptation branch.

A.1 GENERATION AND COMPLETION TRAINING

Using the frozen VAE, we independently train the generation-only and joint generation-completion TRACE variants for 1000 epochs under the same optimization budget.

VAE Objectives. The split VAE is trained using masked reconstruction together with four motion regularizers. All losses are averaged over valid frames. The reconstruction term combines the complete motion and object-stream errors, while the regularization term supervises object pose, temporal velocity, body kinematics, and latent distribution:

$$\begin{aligned} \mathcal{L}_{\text{rec}} &= \mathcal{L}_1(\widehat{\boldsymbol{x}}, \boldsymbol{x}) + \lambda_{\text{o}}\mathcal{L}_1(\widehat{\boldsymbol{x}}_o, \boldsymbol{x}_o), \\ \mathcal{L}_{\text{reg}} &= \lambda_{\text{obj}}\mathcal{L}_{\text{obj}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{kin}}\mathcal{L}_{\text{kin}} + \beta\mathcal{L}_{\text{KL}}. \end{aligned} \tag{10}$$

Here, $\mathcal{L}_{\text{obj}}$ applies Smooth L1 loss to object translation and a rotation loss $1 - (\text{tr}(\widehat{\boldsymbol{R}}^\top \boldsymbol{R}) - 1)/2$. $\mathcal{L}_{\text{vel}}$ matches first-order differences of human joint positions and object translation, while $\mathcal{L}_{\text{kin}}$ matches the body bone vectors defined by the SMPL-H kinematic tree. $\mathcal{L}_{\text{KL}}$ is averaged over the valid latent positions of the three streams. We set $\lambda_{\text{o}} = 0.25$, $\lambda_{\text{obj}} = 0.25$, $\lambda_{\text{vel}} = 0.05$, $\lambda_{\text{kin}} = 0.1$, and $\beta = 10^{-4}$.

Decoded Interaction Constraints. Given the clean latent estimate $\widehat{\boldsymbol{z}}_{m,1}$, the frozen VAE reconstructs the complete interaction $\widehat{\boldsymbol{x}}$. Let $d_{t,j}$ and $\widehat{d}_{t,j}$ denote the ground-truth and predicted distances from human joint j to the object surface at frame t. We define the contact label and its differentiable prediction as $c_{t,j} = \mathbb{1}[d_{t,j} < \gamma]$ and $\widehat{c}_{t,j} = \text{sigmoid}((\gamma - \widehat{d}_{t,j})/\sigma_{\text{c}})$, respectively. The contact objective is

$$\mathcal{L}_{\text{c}} = \mathcal{L}_{\text{BCE}}(\widehat{\boldsymbol{c}}, \boldsymbol{c}) + 2\mathcal{L}_{\text{Dice}}(\widehat{\boldsymbol{c}}, \boldsymbol{c}), \tag{11}$$

where positive contacts receive a weight of 3. We use $\gamma = 0.05\,\text{m}$ and $\sigma_{\text{c}} = 0.01\,\text{m}$.

To supervise relative interaction geometry, let $\boldsymbol{a}_{t,j}$ be a human anchor and define its coordinates in the object frame as $\boldsymbol{r}_{t,j} = \boldsymbol{R}_t^\top(\boldsymbol{a}_{t,j} - \boldsymbol{t}_t)$. A soft contact weight $w_{t,j}$ selects anchors close to the ground-truth object. Using Smooth L1 loss $\rho(\cdot)$, the anchoring and temporal episode objectives are

$$\mathcal{L}_{\text{A}} = \frac{\sum_{t,j} w_{t,j}\, \rho((\widehat{\boldsymbol{r}}_{t,j} - \boldsymbol{r}_{t,j})/s_{\text{A}})}{\sum_{t,j} w_{t,j}}, \tag{12}$$

$$\mathcal{L}_{\text{E}} = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} \frac{\sum_{t,j} w_{t,j} w_{t+s,j}\, \rho((\Delta_s \widehat{\boldsymbol{r}}_{t,j} - \Delta_s \boldsymbol{r}_{t,j})/s_{\text{E}})}{\sum_{t,j} w_{t,j} w_{t+s,j}}, \tag{13}$$

where $\Delta_s \boldsymbol{r}_{t,j} = \boldsymbol{r}_{t+s,j} - \boldsymbol{r}_{t,j}$ and $\mathcal{S} = \{1, 4, 8\}$. We use $\mathcal{L}_{\text{A/E}} = 0.01\mathcal{L}_{\text{A}} + 0.005\mathcal{L}_{\text{E}}$, with $s_{\text{A}} = 0.05\,\text{m}$ and $s_{\text{E}} = 0.02\,\text{m}$. These terms constrain both the contact location in the object frame and its evolution over time.

Joint Generation and Completion. We represent each task using an observation mask $\boldsymbol{\omega} \in \{0,1\}^3$, where $\omega_m = 1$ indicates that stream m is observed. Text-to-HOI uses $\boldsymbol{\omega} = (0,0,0)$, whereas each completion task observes two streams and generates the remaining one. For an observed stream, we directly use its clean latent with flow time $\tau_m = 1$ and set its target velocity to zero. The missing stream follows the same flow used for generation. Let $\mathcal{G}(\boldsymbol{\omega}) = \{m \mid \omega_m = 0\}$ denote the generated streams. Its flow loss is normalized as

$$\mathcal{L}_{\mathrm{fm}}^{\boldsymbol{\omega}} = \frac{3}{|\mathcal{G}(\boldsymbol{\omega})|} \sum_{m \in \mathcal{G}(\boldsymbol{\omega})} \mathcal{L}_m, \tag{14}$$

which retains the three-stream loss scale for both generation and completion.

The four tasks are sampled with a ratio of 2:1:1:1 for Text-to-HOI, body completion, object completion, and hand completion. Observation and task embeddings indicate the available streams and the stream to be generated. A zero-initialized completion adapter is activated only for the missing stream in each transformer block. Because joint attention also updates observed hidden states, we restore them after every block and clamp their latents throughout Euler sampling. The original observed streams are copied into the final motion after decoding, ensuring that only the missing component is modified.

Training Configuration. We optimize all models using AdamW with weight decay 10^{-4} and gradient clipping at 1.0. The split VAE is trained for $1{,}000$ epochs with batch size 64 and learning rate 2×10^{-4}. The Text-to-HOI flow is trained for 1000 epochs with batch size 128 and learning rate 10^{-4}. Its learning rate is warmed up over the first 2% of training and then reduced by cosine decay. The joint generation-completion model is also trained for 1000 epochs. We set the routing-balance weight to 0.01 and use a completion-adapter dimension of 128. Decoded interaction constraints are evaluated on at most 32 samples using contact-aware windows of 64 frames and flow times within $[0.2, 0.8]$; their weights are linearly warmed up over the first 10% of training. All models are trained with bf16 precision on NVIDIA A800 GPUs. At inference, we integrate the coupled flow using 32 Euler steps.

A.2 HOI Understanding and Language Adaptation

We detail how frozen coupled-flow states are converted into continuous memory tokens and used to adapt the language model for HOI understanding.

HOI Feature Extraction. Given a complete interaction $\boldsymbol{x}$, the frozen VAE encodes its body, object, and hand streams into normalized latent vectors $\{\widetilde{\boldsymbol{z}}_m\}_{m \in \{b,o,h\}}$. We pass these latents through the frozen coupled flow transformer at $\tau = 1$. Both the text and geometry conditions are replaced with zeros, ensuring that the extracted representation depends only on the interaction motion. Denote the frozen coupled flow transformer before its velocity head by $\mathcal{T}_\theta$. The extracted states are:

$$\left(\boldsymbol{h}_b^L, \boldsymbol{h}_o^L, \boldsymbol{h}_h^L\right) = \mathcal{T}_\theta\left(\widetilde{\boldsymbol{z}}_b, \widetilde{\boldsymbol{z}}_o, \widetilde{\boldsymbol{z}}_h; 1, \mathbf{0}, \mathbf{0}\right). \tag{15}$$

Since each state has exchanged information with the other streams through joint attention, hidden states encode both the motion of stream m and its relation to the complete interaction. We use $\boldsymbol{h} = [\boldsymbol{h}_b^L, \boldsymbol{h}_o^L, \boldsymbol{h}_h^L]$ as the motion representation for language adaptation.

HOI Language Bridge. The flow representation contains a number of tokens proportional to the motion duration, whereas the language model requires a compact input with a fixed size. We first add a learned readout stream embedding $\boldsymbol{e}_m^{\mathrm{r}}$ to distinguish body, object, and hand tokens:

$$\boldsymbol{h}_{\mathrm{r}} = \left[\boldsymbol{h}_b^L + \boldsymbol{e}_b^{\mathrm{r}}; \boldsymbol{h}_o^L + \boldsymbol{e}_o^{\mathrm{r}}; \boldsymbol{h}_h^L + \boldsymbol{e}_h^{\mathrm{r}}\right]. \tag{16}$$

A temporal resampler uses learned slots to attend to the valid tokens in $\boldsymbol{h}_{\mathrm{r}}$. Each resampler block applies cross-attention from the slots to the HOI tokens, followed by slot self-attention and a feed-forward network. We use two blocks, eight attention heads, and 32 learned slots. The resampled features are mapped into the embedding space of the language model using LayerNorm and a linear projection:

$$\boldsymbol{S} = \mathcal{R}\left(\boldsymbol{h}_{\mathrm{r}}, \boldsymbol{V}\right), \qquad \boldsymbol{E} = \mathcal{P}\left(\boldsymbol{S}\right). \tag{17}$$

Here, $\boldsymbol{V}$ masks padded motion tokens, $\boldsymbol{S}$ denotes the resampled slots, and $\boldsymbol{E}$ is the continuous prefix passed to the language model.

Table A: **Quantitative comparisons on the OMOMO dataset between our method and baseline approaches.** We report R-Precision with a batch size of 256.

Method	R-Precision ↑			FID ↓	MM Dist. ↓	Diversity →	FSR ↓	Pene. ↓	Contact →	Interaction ↑		
	Top 1	Top 2	Top 3							C_{prec}	C_{rec}	C_{F1}
Ground Truth	$0.318^{\pm 0.003}$	$0.560^{\pm 0.001}$	$0.731^{\pm 0.001}$	$0.000^{\pm 0.000}$	$1.499^{\pm 0.000}$	$7.400^{\pm 0.047}$	$0.012^{\pm 0.039}$	$0.067^{\pm 0.000}$	$0.262^{\pm 0.000}$	$1.000^{\pm 0.000}$	$1.000^{\pm 0.000}$	$1.000^{\pm 0.000}$
HOI-Diff	$0.221^{\pm 0.005}$	$0.403^{\pm 0.001}$	$0.542^{\pm 0.001}$	$0.987^{\pm 0.023}$	$2.520^{\pm 0.022}$	$7.248^{\pm 0.066}$	$0.021^{\pm 0.058}$	$0.131^{\pm 0.005}$	$0.135^{\pm 0.002}$	$0.862^{\pm 0.000}$	$0.696^{\pm 0.000}$	$0.739^{\pm 0.001}$
CHOIS	$0.291^{\pm 0.000}$	$0.515^{\pm 0.009}$	$0.676^{\pm 0.008}$	$0.132^{\pm 0.016}$	$1.746^{\pm 0.002}$	$7.417^{\pm 0.104}$	$0.019^{\pm 0.053}$	$0.096^{\pm 0.004}$	$0.230^{\pm 0.001}$	$0.937^{\pm 0.000}$	$0.900^{\pm 0.004}$	$\mathbf{0.910}^{\pm 0.002}$
InterDiff	$0.312^{\pm 0.003}$	$0.548^{\pm 0.004}$	$0.718^{\pm 0.004}$	$0.163^{\pm 0.009}$	$1.587^{\pm 0.006}$	$7.387^{\pm 0.045}$	$0.025^{\pm 0.062}$	$0.100^{\pm 0.002}$	$0.206^{\pm 0.000}$	$\mathbf{0.940}^{\pm 0.016}$	$0.863^{\pm 0.016}$	$0.890^{\pm 0.017}$
Text2HOI	$0.286^{\pm 0.001}$	$0.513^{\pm 0.001}$	$0.655^{\pm 0.001}$	$0.168^{\pm 0.000}$	$1.832^{\pm 0.006}$	$7.359^{\pm 0.034}$	$0.020^{\pm 0.056}$	$0.127^{\pm 0.002}$	$0.163^{\pm 0.000}$	$0.919^{\pm 0.001}$	$0.757^{\pm 0.002}$	$0.807^{\pm 0.002}$
ROG	$0.309^{\pm 0.006}$	$0.545^{\pm 0.007}$	$0.713^{\pm 0.006}$	$0.136^{\pm 0.018}$	$1.674^{\pm 0.012}$	$\mathbf{7.408}^{\pm 0.080}$	$0.016^{\pm 0.050}$	$0.093^{\pm 0.003}$	$0.221^{\pm 0.002}$	$0.928^{\pm 0.004}$	$0.846^{\pm 0.008}$	$0.876^{\pm 0.006}$
LIGHT	$0.302^{\pm 0.004}$	$0.538^{\pm 0.015}$	$0.707^{\pm 0.011}$	$0.099^{\pm 0.018}$	$1.657^{\pm 0.010}$	$7.443^{\pm 0.043}$	$\mathbf{0.013}^{\pm 0.045}$	$0.100^{\pm 0.001}$	$0.194^{\pm 0.000}$	$\mathbf{0.940}^{\pm 0.000}$	$0.851^{\pm 0.008}$	$0.881^{\pm 0.005}$
TRACE	$0.342^{\pm 0.003}$	$0.601^{\pm 0.006}$	$0.764^{\pm 0.002}$	$0.087^{\pm 0.009}$	$1.557^{\pm 0.017}$	$7.422^{\pm 0.122}$	$0.018^{\pm 0.054}$	$0.089^{\pm 0.005}$	$0.227^{\pm 0.002}$	$0.917^{\pm 0.009}$	$0.889^{\pm 0.009}$	$0.894^{\pm 0.010}$
TRACE ^{joint}	$\mathbf{0.344}^{\pm 0.004}$	$\mathbf{0.611}^{\pm 0.007}$	$\mathbf{0.768}^{\pm 0.004}$	$\mathbf{0.077}^{\pm 0.008}$	$\mathbf{1.541}^{\pm 0.012}$	$7.433^{\pm 0.099}$	$0.018^{\pm 0.042}$	$\mathbf{0.084}^{\pm 0.002}$	$\mathbf{0.232}^{\pm 0.001}$	$0.919^{\pm 0.005}$	$\mathbf{0.901}^{\pm 0.003}$	$0.904^{\pm 0.004}$

Table B: **Quantitative comparisons on the BEHAVE dataset between our method and baseline approaches.** We report R-Precision with a batch size of 256.

Method	R-Precision ↑			FID ↓	MM Dist. ↓	Diversity →	FSR ↓	Pene. ↓	Contact →	Interaction ↑		
	Top 1	Top 2	Top 3							C_{prec}	C_{rec}	C_{F1}
Ground Truth	$0.559^{\pm 0.005}$	$0.910^{\pm 0.000}$	$0.926^{\pm 0.000}$	$0.000^{\pm 0.000}$	$1.483^{\pm 0.018}$	$6.940^{\pm 0.153}$	$0.181^{\pm 0.224}$	$0.094^{\pm 0.001}$	$0.204^{\pm 0.001}$	$1.000^{\pm 0.000}$	$1.000^{\pm 0.000}$	$1.000^{\pm 0.000}$
HOI-Diff	$0.167^{\pm 0.009}$	$0.298^{\pm 0.012}$	$0.358^{\pm 0.018}$	$1.463^{\pm 0.044}$	$4.152^{\pm 0.038}$	$6.723^{\pm 0.074}$	$0.179^{\pm 0.238}$	$0.171^{\pm 0.007}$	$0.100^{\pm 0.003}$	$0.646^{\pm 0.008}$	$0.515^{\pm 0.010}$	$0.519^{\pm 0.007}$
CHOIS	$0.219^{\pm 0.000}$	$0.396^{\pm 0.003}$	$0.492^{\pm 0.022}$	$0.705^{\pm 0.027}$	$3.423^{\pm 0.045}$	$6.863^{\pm 0.061}$	$0.191^{\pm 0.253}$	$0.128^{\pm 0.011}$	$0.170^{\pm 0.003}$	$0.717^{\pm 0.018}$	$0.662^{\pm 0.000}$	$0.650^{\pm 0.005}$
InterDiff	$0.258^{\pm 0.027}$	$0.443^{\pm 0.019}$	$0.531^{\pm 0.027}$	$0.630^{\pm 0.033}$	$3.219^{\pm 0.040}$	$6.862^{\pm 0.066}$	$0.208^{\pm 0.264}$	$0.143^{\pm 0.009}$	$\mathbf{0.173}^{\pm 0.005}$	$0.723^{\pm 0.001}$	$0.682^{\pm 0.013}$	$0.676^{\pm 0.004}$
Text2HOI	$0.135^{\pm 0.003}$	$0.242^{\pm 0.000}$	$0.322^{\pm 0.008}$	$1.291^{\pm 0.073}$	$3.815^{\pm 0.026}$	$6.544^{\pm 0.013}$	$0.177^{\pm 0.233}$	$0.179^{\pm 0.003}$	$0.128^{\pm 0.001}$	$0.703^{\pm 0.011}$	$0.579^{\pm 0.007}$	$0.587^{\pm 0.008}$
ROG	$0.351^{\pm 0.010}$	$0.528^{\pm 0.013}$	$0.620^{\pm 0.016}$	$0.612^{\pm 0.046}$	$3.169^{\pm 0.031}$	$\mathbf{6.948}^{\pm 0.084}$	$\mathbf{0.168}^{\pm 0.231}$	$\mathbf{0.106}^{\pm 0.006}$	$0.151^{\pm 0.003}$	$0.749^{\pm 0.006}$	$0.628^{\pm 0.010}$	$0.656^{\pm 0.008}$
LIGHT	$0.277^{\pm 0.011}$	$0.475^{\pm 0.014}$	$0.561^{\pm 0.041}$	$\mathbf{0.481}^{\pm 0.087}$	$3.277^{\pm 0.076}$	$6.961^{\pm 0.066}$	$0.171^{\pm 0.245}$	$0.140^{\pm 0.004}$	$0.170^{\pm 0.001}$	$0.731^{\pm 0.006}$	$\mathbf{0.720}^{\pm 0.011}$	$\mathbf{0.697}^{\pm 0.007}$
TRACE	$0.438^{\pm 0.009}$	$0.584^{\pm 0.004}$	$0.681^{\pm 0.009}$	$0.688^{\pm 0.001}$	$3.145^{\pm 0.011}$	$6.978^{\pm 0.447}$	$0.169^{\pm 0.001}$	$0.163^{\pm 0.006}$	$0.141^{\pm 0.003}$	$0.737^{\pm 0.002}$	$0.650^{\pm 0.002}$	$0.653^{\pm 0.003}$
TRACE ^{joint}	$\mathbf{0.466}^{\pm 0.004}$	$\mathbf{0.613}^{\pm 0.017}$	$\mathbf{0.688}^{\pm 0.017}$	$0.578^{\pm 0.013}$	$\mathbf{3.044}^{\pm 0.001}$	$6.906^{\pm 0.357}$	$0.175^{\pm 0.001}$	$0.122^{\pm 0.006}$	$0.161^{\pm 0.001}$	$\mathbf{0.758}^{\pm 0.009}$	$0.682^{\pm 0.002}$	$0.682^{\pm 0.006}$

Language Model Fine-tuning. We use the following prompt for all samples:

```
Object category: [CATEGORY]
Question: Describe the human-object interaction.
Answer:
```

The textual query contains the object category and the instruction *Describe the human-object interaction*. We freeze the VAE, coupled flow transformer, and LLaDA backbone, and optimize the HOI language bridge together with LoRA parameters (Hu et al., 2022). Training uses a masked diffusion objective computed only over answer tokens. For each caption, we mask either a random nonempty subset of its tokens or the complete answer with probability 0.8, and train the language model to recover the original description.

Raw-motion and flow representations use identical category prompts and contrast the correct HOI memory with a same-category mismatch. Let ℓ_i^+ denote the caption loss using the correct memory and ℓ_i^- the loss using another interaction from the same category. The complete language objective is:

$$\mathcal{L}_{\text{lang}} = \mathcal{L}_{\text{cap}} + \lambda_{\text{g}} \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \max\left(0, \delta + \ell_i^+ - \ell_i^-\right), \tag{18}$$

where $\mathcal{I}$ contains samples with a valid paired interaction. We set $\lambda_{\text{g}} = 0.5$ and $\delta = 0.1$. The bridge is first trained alone for 5 epochs, after which the LoRA parameters are enabled. We train for 100 epochs with learning rates of 5×10^{-5} for the bridge and 2×10^{-6} for LoRA. LoRA uses rank 8, scale 16, and dropout 0.05. During inference, LLaDA converts the continuous HOI memory into a textual description through 48 denoising steps.

B Additional Experimental Results

This section reports additional generation and understanding results on OMOMO and BEHAVE, evaluates all three stream completion tasks across datasets, and presents qualitative results for HOI understanding.

HOI Generation on Additional Datasets. Tables A and B report generation results on OMOMO and BEHAVE, respectively. Following prior methods, all models are trained separately on each dataset using the corresponding training split. TRACE performs favorably on most generation metrics, producing lower feature distribution distances and stronger alignment between generated motion and text. The improvements in contact and physical metrics further indicate that the generated human and object trajectories remain coordinated. Although OMOMO and BEHAVE contain substantially fewer interactions than InterAct, the coupled flow can still learn effective dependencies among body, object,

Table C: **Comparison of HOI understanding on OMOMO.** Our model is trained for 100 epochs.

Methods	R-Precision ↑			MM Dist. ↓	Bleu@1 ↑	Bleu@4 ↑	Rouge ↑	Cider ↑	BertScore ↑
	Top 1	Top 2	Top 3						
Real	0.345	0.609	0.778	1.499	–	–	–	–	–
TM2T	0.231	0.421	0.598	2.173	80.62	74.18	90.87	7.46	91.72
MotionGPT	0.254	0.459	0.632	2.046	84.51	79.36	91.92	8.16	92.31
LaMPM2T	0.276	0.490	0.661	1.928	88.74	85.61	92.96	**9.37**	93.21
HOIGPT	0.291	0.522	0.688	1.824	91.56	85.18	93.58	9.08	93.14
TRACE ^{Raw}	0.260	0.472	0.658	1.965	86.29	82.91	92.48	7.31	93.18
TRACE	**0.316**	**0.553**	**0.715**	**1.692**	**93.77**	**90.12**	**94.20**	8.80	**93.30**

Table D: **Comparison of HOI understanding on BEHAVE.** Our model is trained for 100 epochs.

Methods	R-Precision ↑			MM Dist. ↓	Bleu@1 ↑	Bleu@4 ↑	Rouge ↑	Cider ↑	BertScore ↑
	Top 1	Top 2	Top 3						
Real	0.913	0.944	0.956	1.498	–	–	–	–	–
TM2T	0.246	0.383	0.501	4.143	23.18	9.62	37.16	4.9	22.18
MotionGPT	0.276	0.421	0.544	3.846	26.43	11.57	40.13	5.9	24.16
LaMPM2T	0.306	0.459	0.586	3.596	29.14	13.92	43.09	6.7	26.12
HOIGPT	0.336	0.497	**0.632**	**3.301**	**32.64**	**16.52**	**46.57**	**8.32**	**30.61**
TRACE ^{Raw}	0.290	0.438	0.555	3.756	27.52	12.27	41.29	6.1	24.35
TRACE	**0.350**	**0.513**	0.619	3.484	30.58	13.83	44.34	7.4	25.95

and hand motion. These results show that TRACE remains effective under different dataset scales and interaction distributions.

HOI Understanding on Additional Datasets. Tables C and D compare understanding on OMOMO and BEHAVE. Across both datasets, frozen coupled-flow features improve every reported metric over raw-motion encoding. Against external baselines, performance varies by dataset: TRACE is competitive on OMOMO and improves R@1 and R@2 on BEHAVE, but does not lead most BEHAVE language metrics. These results support the internal representation comparison without implying uniform superiority over all baselines.

Table E: **Stream completion on InterAct, OMOMO, and BEHAVE.** For each dataset, one jointly trained checkpoint is evaluated under three conditioning settings.

Dataset	Setting	Target	Human reconstruction ↓			Object reconstruction ↓		Interaction ↑		
			B-MPJPE	Root	H-JPE	Center	V2V	C_{prec}	C_{rec}	C_{F1}
InterAct	Text/BPS only	Object	–	–	–	69.90	74.79	0.420	0.172	0.195
		Hands	–	–	726.46	–	–	0.390	0.148	0.164
		Body	170.74	574.55	–	–	–	–	–	–
	Streams only	Object	–	–	–	19.25	34.77	0.503	0.309	0.330
		Hands	–	–	**44.91**	–	–	0.669	0.613	0.610
		Body	**55.21**	**72.89**	–	–	–	–	–	–
	Full	Object	–	–	–	**15.34**	**26.51**	**0.584**	**0.411**	**0.434**
		Hands	–	–	47.19	–	–	**0.679**	**0.629**	**0.625**
		Body	60.74	80.01	–	–	–	–	–	–
OMOMO	Text/BPS only	Object	–	–	–	95.44	100.47	0.407	0.093	0.128
		Hands	–	–	969.38	–	–	0.379	0.075	0.107
		Body	200.85	781.40	–	–	–	–	–	–
	Streams only	Object	–	–	–	18.25	36.37	0.530	0.324	0.354
		Hands	–	–	**69.75**	–	–	**0.669**	**0.595**	**0.594**
		Body	**50.89**	**83.30**	–	–	–	–	–	–
	Full	Object	–	–	–	**14.16**	**25.30**	**0.612**	**0.414**	**0.445**
		Hands	–	–	71.34	–	–	0.665	0.592	0.592
		Body	51.80	83.81	–	–	–	–	–	–
BEHAVE	Text/BPS only	Object	–	–	–	52.70	60.37	0.382	0.135	0.155
		Hands	–	–	580.85	–	–	0.354	0.113	0.129
		Body	183.56	419.61	–	–	–	–	–	–
	Streams only	Object	–	–	–	21.99	37.55	0.407	0.261	0.275
		Hands	–	–	**61.76**	–	–	0.586	**0.540**	**0.543**
		Body	**68.75**	**80.87**	–	–	–	–	–	–
	Full	Object	–	–	–	**18.58**	**31.34**	**0.484**	**0.322**	**0.342**
		Hands	–	–	76.61	–	–	**0.589**	0.533	0.535
		Body	68.87	81.30	–	–	–	–	–	–

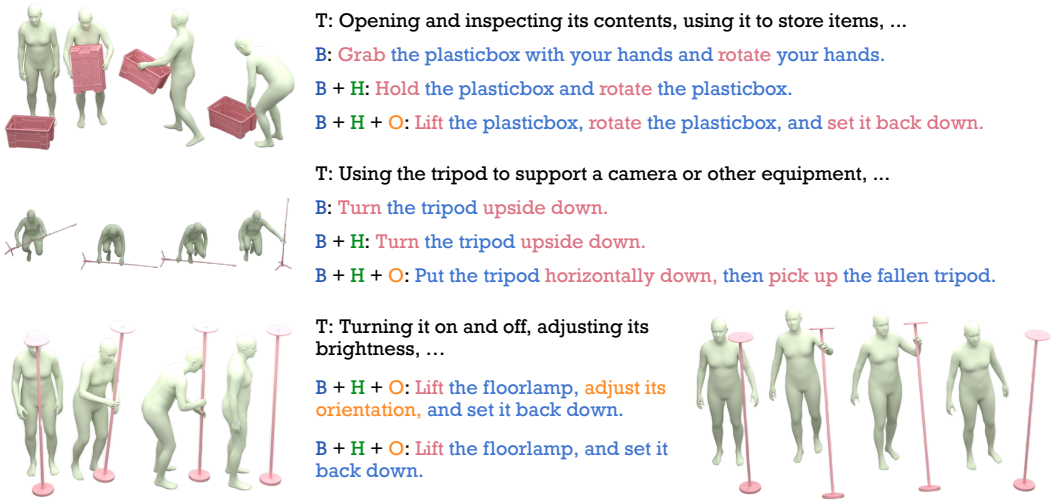

Figure A: **Qualitative HOI understanding.** Adding hand and object evidence produces more complete interaction descriptions. T denotes the reference description.

Stream Completion on Three Datasets. Table E compares three conditioning settings for object, hand, and body completion. Text/BPS-only uses no observed motion, Streams-only removes the text and BPS embeddings, and Full uses all conditions. For Text/BPS-only, the generated target is evaluated with the synchronized ground-truth non-target streams. In the other settings, observed streams are copied unchanged. Each dataset uses one TRACE checkpoint, with results averaged over two evaluation seeds.

B-MPJPE measures root-relative body-joint error, **Root** measures root-translation error, and **H-JPE** measures hand-joint error, all in mm. **Center** and **V2V** measure object-center and surface errors in cm. C_{prec}, C_{rec}, and C_{F1} are macro-averaged either-hand contact metrics under a 5cm threshold; contact is omitted when both hand and object are observed.

Across all datasets, observed streams provide the main completion signal. Full conditioning consistently yields the lowest object reconstruction errors and the highest contact scores for object completion. For hand and body completion, Streams-only remains comparable and sometimes produces lower paired reconstruction errors, showing that the observed trajectories already provide most of the required motion information.

Qualitative Evaluation of HOI Understanding. Figure A analyzes the effects of motion evidence and representation on HOI understanding. In the plastic box and tripod examples, partial body and hand observations yield coarse descriptions, while including all three streams recovers more complete object movement and temporal progression. The floor lamp example further compares two representations under the same complete input. Coupled flow features recognize the orientation adjustment, whereas direct raw motion encoding omits this detail. These results show that both complete HOI evidence and the interaction relations learned during generation benefit HOI understanding.

Table F: **Ablation study of stream modeling on OMOMO.** We report R-Precision with a batch size of 256.

Strategy	R-Precision ↑			FID ↓	MM Dist. ↓	Diversity →	FSR ↓	Pene. ↓	Contact →	Interaction ↑		
	Top 1	Top 2	Top 3							C_{prec}	C_{rec}	C_{F1}
GT	$0.318^{\pm 0.003}$	$0.560^{\pm 0.001}$	$0.731^{\pm 0.001}$	$0.000^{\pm 0.000}$	$1.499^{\pm 0.000}$	$7.400^{\pm 0.047}$	$0.012^{\pm 0.039}$	$0.067^{\pm 0.000}$	$0.262^{\pm 0.000}$	$1.000^{\pm 0.000}$	$1.000^{\pm 0.000}$	$1.000^{\pm 0.000}$
Independent	$0.325^{\pm 0.004}$	$0.570^{\pm 0.007}$	$0.731^{\pm 0.004}$	$0.120^{\pm 0.011}$	$1.701^{\pm 0.020}$	$7.467^{\pm 0.124}$	$\mathbf{0.015}^{\pm 0.078}$	$\mathbf{0.077}^{\pm 0.006}$	$0.174^{\pm 0.002}$	$0.849^{\pm 0.011}$	$0.651^{\pm 0.012}$	$0.696^{\pm 0.012}$
Unified	$\underline{0.339}^{\pm 0.003}$	$\underline{0.597}^{\pm 0.006}$	$\underline{0.759}^{\pm 0.003}$	$\underline{0.093}^{\pm 0.010}$	$\underline{1.590}^{\pm 0.018}$	$\mathbf{7.411}^{\pm 0.118}$	$0.022^{\pm 0.032}$	$0.097^{\pm 0.006}$	$\underline{0.218}^{\pm 0.002}$	$\underline{0.906}^{\pm 0.010}$	$\underline{0.842}^{\pm 0.011}$	$\underline{0.855}^{\pm 0.011}$
Coupled	$\mathbf{0.342}^{\pm 0.003}$	$\mathbf{0.601}^{\pm 0.006}$	$\mathbf{0.764}^{\pm 0.002}$	$\mathbf{0.087}^{\pm 0.009}$	$\mathbf{1.557}^{\pm 0.017}$	$\underline{7.422}^{\pm 0.122}$	$\underline{0.018}^{\pm 0.054}$	$\underline{0.089}^{\pm 0.005}$	$\mathbf{0.227}^{\pm 0.002}$	$\mathbf{0.917}^{\pm 0.009}$	$\mathbf{0.889}^{\pm 0.009}$	$\mathbf{0.894}^{\pm 0.010}$

Table G: **Ablation study of stream modeling on BEHAVE.** We report R-Precision with a batch size of 256.

Strategy	R-Precision ↑			FID ↓	MM Dist. ↓	Diversity →	FSR ↓	Pene. ↓	Contact →	Interaction ↑		
	Top 1	Top 2	Top 3							C_{prec}	C_{rec}	C_{F1}
GT	$0.559^{\pm 0.005}$	$0.910^{\pm 0.000}$	$0.926^{\pm 0.000}$	$0.000^{\pm 0.000}$	$1.483^{\pm 0.018}$	$6.940^{\pm 0.153}$	$0.181^{\pm 0.224}$	$0.094^{\pm 0.001}$	$0.204^{\pm 0.001}$	$1.000^{\pm 0.000}$	$1.000^{\pm 0.000}$	$1.000^{\pm 0.000}$
Independent	$0.414^{\pm 0.011}$	$0.550^{\pm 0.007}$	$0.643^{\pm 0.011}$	$0.842^{\pm 0.014}$	$3.392^{\pm 0.048}$	$7.056^{\pm 0.051}$	$0.194^{\pm 0.122}$	$\mathbf{0.142}^{\pm 0.007}$	$0.104^{\pm 0.003}$	$0.676^{\pm 0.004}$	$0.451^{\pm 0.005}$	$0.486^{\pm 0.005}$
Unified	$\underline{0.435}^{\pm 0.010}$	$\underline{0.580}^{\pm 0.006}$	$\underline{0.678}^{\pm 0.010}$	$\underline{0.716}^{\pm 0.012}$	$\underline{3.188}^{\pm 0.046}$	$\mathbf{6.956}^{\pm 0.030}$	$\underline{0.177}^{\pm 0.194}$	$0.174^{\pm 0.007}$	$\underline{0.136}^{\pm 0.003}$	$\underline{0.730}^{\pm 0.003}$	$\underline{0.622}^{\pm 0.004}$	$\underline{0.631}^{\pm 0.004}$
Coupled	$\mathbf{0.438}^{\pm 0.009}$	$\mathbf{0.584}^{\pm 0.004}$	$\mathbf{0.681}^{\pm 0.009}$	$\mathbf{0.688}^{\pm 0.017}$	$\mathbf{3.145}^{\pm 0.040}$	$\underline{6.978}^{\pm 0.047}$	$\mathbf{0.169}^{\pm 0.139}$	$\underline{0.163}^{\pm 0.006}$	$\mathbf{0.141}^{\pm 0.003}$	$\mathbf{0.737}^{\pm 0.002}$	$\mathbf{0.650}^{\pm 0.009}$	$\mathbf{0.653}^{\pm 0.003}$

Impact of Stream Modeling on Additional Datasets. Tables F and G extend the comparison of unified, independent, and coupled stream modeling to OMOMO and BEHAVE. The coupled formulation achieves the strongest overall generation and interaction results on both datasets, consistent

with the InterAct results in the main paper. This confirms that preserving separate stream states while coupling their evolution remains effective across different dataset scales and interaction distributions.

Table H: **Inference efficiency on InterAct.** All methods generate a batch of 64 interactions, each containing 300 frames, on one NVIDIA A800 GPU using BF16.

Method	Time / batch (s) ↓	Time / sequence (s) ↓	Relative latency (×) ↓
HOI-Diff	$35.20^{\pm 0.12}$	0.550	10.2
InterDiff	$42.88^{\pm 0.35}$	0.670	12.5
LIGHT	$192.76^{\pm 0.57}$	3.012	56.0
TRACE	$3.44^{\pm 0.01}$	0.054	1.0

Inference efficiency. We evaluate all methods on one NVIDIA A800 GPU using BF16, a batch size of 64, and sequences of 300 frames. Each method uses its official inference configuration. After five warm-up runs, we measure 10 runs with CUDA synchronization before and after timing. Runtime includes sampling and motion decoding but excludes data loading, rendering, and file I/O. As shown in Table H, TRACE requires 3.44 seconds per batch and is $10.2\times$, $12.5\times$, and $56.0\times$ faster than HOI-Diff, InterDiff, and LIGHT, respectively.