

# ReMoGen: Real-time Human Interaction-to-Reaction Generation via Modular Learning from Diverse Data

## Supplementary Material

### A. Overview

This supplementary material provides a detailed description of the *ReMoGen* framework for real-time human interaction-to-reaction generation. Due to space constraints, the main paper only briefly summarizes a number of design choices and implementation details. Here we expand all core components so that the framework is fully specified and reproducible.

### B. Representations

#### B.1. Motion Representation

We adopt a SMPL-X-based motion representation aligned with our backbone design [49]. For each frame  $t$ , we construct a compact feature vector  $m_t$  using physically meaningful quantities derived from the root motion and the first 22 SMPL-X body joints. Specifically, we extract:

- Root translation  $\mathbf{t}_t$  in the canonical frame.
- Joint rotations  $\mathbf{R}_t$  in continuous 6D form for the root and body joints.
- Root motion deltas, including translation changes  $\Delta\mathbf{t}_t$  and orientation changes  $\Delta\mathbf{R}_t$  between consecutive frames.
- Joint positions  $\mathbf{J}_t$  obtained from the SMPL-X forward pass.
- Joint velocities  $\Delta\mathbf{J}_t$  computed as per-frame differences.

The final per-frame representation  $m_t$  is formed by concatenating all components above, and is used throughout the training and inference stages of our autoregressive latent-diffusion backbone.

**Canonicalization and Normalization.** Before computing features, all SMPL-X sequences  $\{\mathbf{t}_t, \mathbf{R}_t, \mathbf{J}_t\}_{t=1}^T$  are transformed into a consistent *ego-centric canonical frame*.

The pelvis joint of the first frame defines the canonical origin  $\mathbf{p}_0$ .

The body facing direction is estimated from the left and right hip joints (e.g., joints  $j_1$  and  $j_2$ ), which determines the canonical horizontal axes.

A rigid transform ( $\mathbf{R}_{\text{ego}}, \mathbf{t}_{\text{ego}}$ ) is then constructed from this reference frame and applied to the root translation  $\mathbf{t}_t$ , root orientation  $\mathbf{R}_t$ , and joint positions  $\mathbf{J}_t$  of the entire sequence, producing the canonicalized quantities  $\hat{\mathbf{t}}_t$ ,  $\hat{\mathbf{R}}_t$ , and  $\hat{\mathbf{J}}_t$ .

For human-human interaction scenarios, both the ego and the partner sequences are transformed using the same

ego-defined frame, ensuring that all agents' trajectories are expressed consistently relative to the canonicalized ego.

After canonicalization, we compute the global mean and standard deviation of all feature channels using only the training split and subsequently normalize every motion tensor.

**History and Future Segmentation.** We operate in a segment-based autoregressive regime. Each internal step  $i$  of the generator operates on:

- a history window  $M_h^i \in \mathbb{R}^{H \times D}$  of length  $H$ , and
- a future segment  $\hat{M}_f^i \in \mathbb{R}^{F \times D}$  of length  $F$ .

After generating the  $i$ -th future segment, we update the next history by sliding a window of size  $H$ :

$$M_h^{i+1} = [\text{concat}(M_h^i, \hat{M}_f^i)]_{-H}. \quad (1)$$

This corresponds to the update rule described in the main paper.

#### B.2. Scene Representation

Our framework supports human-scene and human-human-scene interactions by encoding the environment as a volumetric voxel occupancy grid [16, 17]. We adopt a unified 3D representation across all scene-aware datasets.

**Voxel Occupancy.** For datasets that natively provide voxelized scenes (e.g., LINGO), we directly use their occupancy volumes. Each scene is represented as a dense grid of size  $x \times y \times z$  with a spatial voxel resolution  $v_{\text{size}}$ . The voxel grids are loaded and cached in the dataset loader, and queried during training using point-based lookup.

**Ego Voxel Representation.** The Scene Encoder receives an *ego-centric voxel occupancy* volume that describes the local spatial layout around the moving subject. The ego voxel grid is defined over a fixed bounding box:

$$x \in [-0.6, 0.6], \quad y \in [-0.6, 0.6], \quad z \in [0.1, 1.2],$$

resulting in a compact  $32 \times 32 \times 32$  occupancy volume.

This bounding box corresponds to a localized workspace around the body. To construct ego voxels, we query the underlying scene occupancy grid at the projected 3D locations of the root joint (or goal positions).

### B.3. Ego Alignment.

All motion features are represented in our ego-centric canonical coordinate frame for consistent modeling across datasets. Before being fed into the network, any other motions or goal positions associated with the ego are also transformed into this canonical frame, ensuring that the model interprets all positional objectives relative to the ego-centered coordinate system.

However, scene occupancy queries must be performed in the global scene coordinate system. Thus, although the model operates in the canonical space, we temporarily convert the canonicalized ego motion back to world coordinates using the stored inverse canonical transform whenever querying the ego voxel grid.

The reconstructed global root positions are then used to sample occupancy values from the corresponding voxel volumes. This two-space design maintains consistent interaction between ego motion, partner motion, and scene geometry while enabling stable canonical-frame learning inside the model.

## C. Supplement Method Details

### C.1. Meta-Interaction Modules

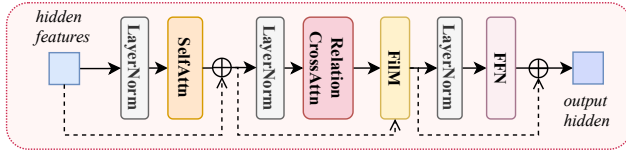


Figure A. Architecture of Meta-Interaction Block.

The universal prior does not model interactions by construction. To inject dynamic context from other agents and the scene while keeping the prior frozen, we introduce *Meta-Interaction Modules* (MIMs). Each MIM is an adapter that modulates intermediate activations in the latent denoiser via FiLM-style conditioning on external context.

**Meta-Interaction Block.** A Meta-Interaction Block (Fig. A) receives ego features  $h \in \mathbb{R}^{B \times T \times d}$  and context tokens  $c_{\text{context}} \in \mathbb{R}^{B \times T_c \times D}$  (from other agents or from the scene encoder). The block outputs a context-induced residual that is added to the ego features.

**Self-Attention.** We first apply standard self-attention to capture intra-sequence dependencies:

$$h' = h + \text{SelfAttn}(\text{LayerNorm}(h)). \quad (2)$$

**Relation-aware Cross-attention.** Given contextual embeddings  $c_{\text{context}}$ , we compute cross-attention using  $Q_h =$

$\text{LayerNorm}(h')W_Q, K_c = c_{\text{context}}W_K, V_c = c_{\text{context}}W_V$  where  $W_Q, W_K, W_V$  are learned projections.

The relation-aware cross-attention output is

$$r = \text{Softmax}\left(\frac{Q_h K_c^\top + B_{\text{rel}}}{\sqrt{d}}\right) V_c, \quad (3)$$

where  $B_{\text{rel}} \in \mathbb{R}^{T \times T_c}$  is a learnable relative positional bias encoding temporal or spatial offsets.

Each bias element is computed using a sinusoidal embedding of the relative index  $\Delta t = i - j$ :

$$b_{ij} = W_b[\sin(\omega \Delta t), \cos(\omega \Delta t)], \quad (4)$$

with a fixed frequency  $\omega = 0.25$  and a learned projection  $W_b$ .

**FiLM-style Modulation.** The cross-attention output  $r$  is mapped to FiLM parameters  $(\gamma, \beta)$  via a linear layer. We modulate  $h'$  using

$$h_{\text{mod}} = (1 + \tanh \gamma) \odot h' + \tanh \beta. \quad (5)$$

**Feed-Forward Network (FFN).** A transformer-style FFN refines the modulated features:

$$h_{\text{ffn}} = h_{\text{mod}} + \text{FFN}(\text{LayerNorm}(h_{\text{mod}})), \quad (6)$$

The final context-induced residual is

$$\Delta_{\text{context}} = h_{\text{ffn}} - h. \quad (7)$$

This residual is injected into the denoiser at predefined transformer layers, enabling interaction-aware modulation.

### C.2. Compose MIMs During Inference

**Composable Inference.** Multiple Meta-Interaction Modules (e.g., HHI and HSI branches) can be activated simultaneously. At each injection layer, the system first produces per-branch residuals  $\Delta_i \in \mathbb{R}^{B \times T \times d}$ , such as  $\Delta_{\text{others}}$  and  $\Delta_{\text{scene}}$ . Their contributions are combined through a weighted sum:

$$\Delta_{\text{total}} = \sum_i \alpha_i \Delta_i, \quad (8)$$

where each coefficient  $\alpha_i$  controls the strength of a module. These weights may be **set manually by the user** or **predicted automatically by an external model**, such as a task-specific rule system or a large language model that interprets textual instructions.

**L2-Norm Clamping for Stability.** A naïve weighted sum can produce overly large residuals, especially when multiple modules reinforce each other. To ensure stable interaction conditioning while preserving the frozen prior’s motion manifold, we normalize  $\Delta_{\text{total}}$  using a *per-sample L2 norm clamp*.

Let

$$\|\Delta_{\text{total}}\|_2 = \|\Delta_{\text{total}}\|_{2,\text{flatten}},$$

where the norm is computed over the  $(T \times d)$  dimensions for each sample. Similarly, let  $m$  be the largest norm among the individual modules:

$$m = \max_i \|\Delta_i\|_{2,\text{flatten}}.$$

We compute a scale factor

$$s = \min\left(1, \frac{m}{\|\Delta_{\text{total}}\|_2 + \varepsilon}\right), \quad (9)$$

with  $\varepsilon$  a small constant for numerical stability. The fused modulation is then:

$$\Delta_{\text{final}} = s \cdot \Delta_{\text{total}}. \quad (10)$$

**Effect of the Clamp.** This operation guarantees that the fused residual never exceeds the magnitude of the strongest individual branch, preventing runaway updates and ensuring that the decoded motion remains on the prior’s learned manifold. The model therefore supports **arbitrary combinations** of HHI and HSI cues—including user-driven, programmatic, or LLM-derived scaling—while maintaining temporal stability and preserving the geometry of the pretrained motion prior.

### C.3. Frame-wise Segment Refinement (FWSR)

Segment-based autoregressive rollout provides strong long-horizon stability, but the model updates its reaction only once per segment, introducing latency. To improve responsiveness under real-time constraints, we adopt a lightweight *Frame-wise Segment Refinement (FWSR)* module that applies per-frame latent corrections without re-running the full diffusion process. FWSR operates directly on the clean VAE latent  $z_0 \in \mathbb{R}^{d_z}$  predicted for the segment by the universal prior.

**FWSR Module Overview.** The FWSR module receives the clean latent  $z_0$ , the ego-dynamic feature sequence  $M_h$ , and dynamic context  $X_{\text{dyn}}$  such as other-agent motion. Dynamic context tokens are first processed by a lightweight refinement block consisting of a projection layer and a self-attention encoder to obtain  $c_{\text{dyn}}$ . The processed dynamic tokens, together with  $M_h$ , are then fed into a relation-aware cross-attention layer followed by a FiLM head. The architectural components mirror the Meta-Interaction Modules (MIM), but operate directly in the decoder latent space.

**Relation-Aware Attention.** At each refinement step, the module constructs relation features jointly from the ego-dynamic features  $M_h$  and each processed dynamic context token  $c_{\text{dyn}}$ . Let

$$Q = z_0 W_Q, \quad K = c_{\text{dyn}} W_K, \quad V = c_{\text{dyn}} W_V,$$

where  $W_Q, W_K, W_V \in \mathbb{R}^{d_z \times d_z}$  are learned projections. Relation features  $(M_h, X_{\text{dyn}})$  are used to compute relative positional bias  $B_{\text{rel}}$ . Cross-attention is then computed as:

$$\Delta_{\text{attn}} = \text{RelAttn}(Q, K, V, \text{RelBias}(M_h, X_{\text{dyn}})),$$

allowing the ego latent to incorporate fine-grained, interaction-dependent refinements.

**FiLM Latent Modulation.** The attention output is mapped to FiLM parameters  $(\gamma, \beta)$ , and applied to refine the latent:

$$\Delta_{\text{FWSR}} = (1 + \tanh \gamma) \odot z_0 + \tanh \beta - z_0.$$

**Safe Latent Scaling via Decoder Sensitivity.** Directly adding  $\Delta_{\text{FWSR}}$  may push the latent into directions where the decoder is overly sensitive. To stabilize the refinement, we use a *decoder sensitivity vector*  $s \in \mathbb{R}^{d_z}$ , estimated using finite-difference gradients of the decoder with respect to each latent dimension.

We apply per-dimension scaling:

$$\Delta_{\text{safe},d} = \frac{\Delta_{\text{FWSR},d}}{1 + \beta s_d},$$

where  $\beta$  is a hyper-parameter controlling suppression strength. We use  $\beta = 1.0$ . This protects highly sensitive latent dimensions, suppressing artifacts while leaving robust directions expressive.

**Final Latent Refinement.** The refined latent is obtained as:

$$\tilde{z} = z_0 + \Delta_{\text{safe}}.$$

FWSR performs this refinement once per frame at negligible cost, enabling low-latency adjustments and improved responsiveness while keeping all backbone and interaction modules frozen.

## D. Experimental Setup Details

### D.1. Datasets and Pre-processing

We summarize datasets used for different stages and the pre-processing steps that standardize them into a unified representation.

**HumanML3D.** The **Universal Single-Person Prior** is trained on HumanML3D [12]. We use the official train/val/test splits and its original text annotations.

**Inter-X.** The Meta-Interaction Module for Human-Human Interaction is trained on Inter-X [41], which contains paired SMPL-X sequences of two interacting subjects. We use its official train/val/test splits.

To increase coverage and balance role diversity, we perform data augmentation by treating **both p1 and p2 as the ego** in separate passes. For each choice of ego, the remaining person is treated as the conditioning partner, and we construct history-future motion segments for the ego while using the partner’s motion as dynamic context.

Inter-X provides only *global* textual descriptions for each two-person interaction, without distinguishing the roles of each person. To obtain ego-specific conditioning texts, we use the DeepSeek V3 large language model to automatically decompose each original dual-person description into two separate role-specific descriptions.

We supply the LLM with the raw Inter-X action label and multiple dataset-provided textual descriptions, and instruct the model to identify the actor and reactor and rewrite the interaction into two independent third-person declarative sentences follows the style of HumanML3D. The final assignment of actor/reactor descriptions to p1 or p2 follows the dataset’s official `interaction_order` annotation. This process allows Inter-X’s coarse two-person textual descriptions to be converted into well-aligned, role-specific conditioning signals for training our HHI Meta-Interaction Module.

**LINGO.** The Meta-Interaction Module for Human-Scene Interaction is trained on LINGO [16], which provides SMPL-X motions paired with detailed 3D indoor scenes. We use the dataset’s original text annotations as the semantic descriptions for each motion sequence without additional rewriting.

For scene representation, we follow the Z-up-aligned official LINGO protocol and use the provided *voxelized occupancy grids*. Each indoor environment is represented as a dense occupancy volume defined over the following axis-aligned bounds:

$$x \in [-3.0, 3.0], \quad y \in [-4.0, 4.0], \quad z \in [0.0, 2.0],$$

which is discretized into a grid of size  $300 \times 400 \times 100$  with a spatial resolution of **0.02 m**. This large scene-centric voxel structure captures room-scale spatial layouts, obstacles, and affordances and is used as the conditioning input for our scene encoder.

Since LINGO does not release ground-truth motions for a held-out test set, we construct our own splits by randomly

partitioning the available training data into **70% training**, **10% validation**, and **20% testing**. All reported HSI results in our paper are based on this split.

**EgoBody.** We use EgoBody [47] to evaluate Human-Human-Scene generalization and compositionality. For textual conditioning, we adopt the **sequence-level motion descriptions** provided by Motion-X, which offers annotations aligned with each full interaction sequence.

To maintain consistency with our HHI preprocessing, we apply the same data augmentation strategy used in Inter-X by **swapping the ego and the other agent**. In each augmented version, the chosen ego is canonicalized and the partner is treated as the conditioning sequence.

EgoBody only provides scene geometry in **mesh format**. To convert the scenes into a unified voxel representation, we first normalize each scene by translating the mesh such that its **XY center lies at the (0, 0)** and its **lowest Z coordinate aligns with the ground plane ( $z = 0$ )**. We then voxelize the normalized mesh using the **same occupancy configuration as LINGO**.

**General Pre-processing.** For all datasets, we directly use the provided SMPL-X parameters and convert motions into a consistent **Z-up** coordinate convention. All SMPL-X models are loaded with the `neutral` gender configuration for consistency across datasets. We then obtain body joints by forwarding the SMPL-X parameters through the neutral-gender model.

For all ego sequences, we discard the dataset-provided body shape and set  $\beta = \mathbf{0}$  to eliminate cross-subject shape variation. For interaction scenarios involving other agents, we retain the original SMPL-X  $\beta$  parameters for the partner(s) to preserve their identity and body proportions as provided in each dataset.

All datasets are uniformly downsampled to **10 FPS** using equal-interval sampling to match the real-time generation setting of our framework.

For single-person datasets (HumanML3D) and human-human interaction datasets (Inter-X), we shift the mesh vertically such that the lowest vertex of the first-frame mesh lies exactly on the ground plane. For interaction datasets that include scenes (LINGO, EgoBody), we *do not modify* the absolute spatial placement; the original world-coordinate positions and contact relationships with the environment are strictly preserved.

## D.2. Training Details

Our model is trained in a progressive manner that first learns a universal motion prior, then incorporates interaction-specific conditioning through Meta-Interaction Modules, and optionally adds a Frame-wise Segment Refinement module for enhancing responsiveness to dynamic contexts.

**Universal Single-Person Motion Prior.** We begin by training a universal motion prior on HumanML3D under a purely single-person setting. The model predicts future motion segments using an autoregressive latent diffusion process with a history window of  $H = 2$  frames and a prediction length of  $F = 8$  frames. We adopt a DDPM-based denoising formulation with **10 diffusion steps**. All feature **mean and standard deviation statistics** used for normalization are computed exclusively during this stage and **reused unchanged** across all subsequent interaction datasets (Inter-X, LINGO, EgoBody), ensuring consistent feature scaling throughout training. This phase consists of **250k** steps for VAE training and **400k** steps for training the diffusion denoiser.

**Meta-Interaction Learning.** Once the universal prior is learned, we introduce Meta-Interaction Modules that incorporate external interaction cues such as other agent’s motion (Inter-X) or scene occupancy (LINGO). During this process, the universal backbone remains *frozen*. Each Meta-Interaction module is trained for **65k** steps using its respective dataset.

**Frame-wise Segment Refinement (FWSR).** This module refines autoregressive predictions through per-frame residual updates while keeping both the universal prior and Meta-Interaction parameters fixed. As FWSR is not required for training the core interaction model, it is applied only when there are dynamic contexts like others motion. The refinement module is trained for **65k** steps.

### D.3. Evaluation Protocol and Metrics

**Evaluation Protocol.** We evaluate all models in an *On-line Interaction-to-Reaction* setting, where the ego agent continuously observes interaction cues (other-agent motion or scene occupancy) and produces an immediate reaction. To ensure fair comparison, we strictly follow each baseline’s original inference procedure:

- **Baselines with autoregressive rollout.** If a baseline provides an online or autoregressive generation mechanism, we use it unchanged. The model receives the same history window as specified by its official implementation, without adding any auxiliary history embeddings.
- **Baselines without autoregressive rollout.** If a baseline does not support online generation, we follow the “segment prediction and stitching” protocol: the model predicts motion in fixed-length segments, and consecutive segments are stitched together by recursively feeding the end of the previous segment into the next prediction window. No modifications or additional temporal encodings are introduced beyond what the baseline originally supports.

- **Models without history conditioning.** For baselines that do not learn history embeddings or do not accept temporal context explicitly, we do *not* introduce any additional history mechanism. They receive only the inputs allowed by their native design.

All evaluations are conducted at 10FPS after canonicalization and standardization, consistent with our training setup.

**Metrics.** We adopt a comprehensive set of quantitative metrics widely used in human motion synthesis and interaction modeling. These metrics cover semantic alignment, distributional realism, motion diversity, physical smoothness, and real-time system performance.

**Embedding-based Metrics.** Following standard practice in HumanML3D, we extract normalized motion and text features using the pretrained t2m motion and language encoders [12]. The following metrics are computed in this embedding space:

- **FID.** compares the embedded feature distributions of generated and ground-truth motions, measuring distributional realism. Lower FID indicates that the generated motion better matches the statistics of natural human movement.
- **R-Precision.** evaluates semantic alignment by retrieving the correct textual description of a motion among a batch of 64 candidates. We report top-3 accuracy, reflecting how often the ground-truth description ranks among the three most similar texts based on cosine similarity.
- **Diversity.** quantifies variation among generated motions by computing the average embedding-space distance between multiple samples, indicating the model’s ability to avoid mode collapse.
- **MM-Dist.** measures motion-text semantic consistency by computing the embedding-space distance between a generated motion and its paired textual description. Lower MM-Dist. corresponds to stronger alignment between generated motion and intended semantics.

**Peak Jerk.** This metric captures high-frequency artifacts by measuring the maximum time derivative of joint acceleration. Lower jerk values correspond to smoother, more physically plausible motion trajectories.

**Latency.** We report the average per-frame computation time (in seconds) under our Online Interaction-to-Reaction protocol, measured over 1,000 generated frames. This reflects the model’s responsiveness and suitability for real-time deployment.

**Contact-based Metrics.** We report *Collision Rate* and other contact-based metrics in Tab. A. *Collision Rate* is defined as the ratio of frames in which body joints intersect with the scene mesh or other agents’ meshes. For other contact-based metrics, we follow SymBridge [2], which adopts contact-related metrics from the human-object interaction (HOI) setting. Specifically, we report *Contact Preci-*



| Method                        | FID ↓  | R-Precision<br>(Top-3) ↑ | MM Dist. ↓ | Diversity→ | Contact<br>Precision ↑ | Contact<br>Recall ↑ | Collision(%)↓ |
|-------------------------------|--------|--------------------------|------------|------------|------------------------|---------------------|---------------|
| GT                            | 0.000  | 0.472                    | 4.051      | 4.084      | —                      | —                   | —             |
| ReGenNet [42]                 | 11.622 | 0.269                    | 6.092      | 3.452      | 0.549                  | 0.505               | 0.665         |
| FreeMotion [8]                | 3.383  | 0.284                    | 5.438      | 3.394      | 0.497                  | 0.321               | 0.545         |
| FreeMotion <sup>off</sup> [8] | 0.492  | 0.417                    | 4.330      | 3.913      | 0.496                  | 0.484               | 1.220         |
| SymBridge [2]                 | 2.569  | 0.355                    | 4.955      | 3.598      | 0.614                  | 0.447               | 0.646         |
| Ours                          | 0.181  | 0.464                    | 4.076      | 3.911      | 0.527                  | 0.415               | 0.989         |
| Ours+FWSR                     | 0.166  | 0.462                    | 4.076      | 4.003      | 0.533                  | 0.423               | 0.961         |

Table A. Evaluation with contact-based metrics on HHL.

sion and *Contact Recall*, which measure the alignment between predicted and ground-truth contact patterns.

## E. Additional Experimental Results

### E.1. Contact-based Metrics Results

We provide additional contact-based evaluation to complement the main results (Tab. A). ReMoGen achieves substantially better motion quality (FID) while maintaining competitive contact alignment compared to prior methods.

We note that contact outcomes in interaction scenarios are inherently *multi-modal*. Multiple distinct contact patterns can correspond to equally plausible interactions depending on timing, spatial configuration, and agent behavior. As a result, metrics that directly compare predicted contacts to a single ground-truth pattern may underestimate valid alternative interactions.

### E.2. Latency Breakdown

We provide a detailed breakdown of runtime (Tab. B) to complement the end-to-end latency reported in the main paper. The total latency corresponds to end-to-end inference time measured on a single GPU. Time not explicitly assigned to the listed components is attributed to data I/O and other implementation overhead.

## F. Robustness Analysis

We analyze the robustness of ReMoGen along two dimensions: (i) sensitivity to semantic inputs, and (ii) sensitivity to different context encoder choices.

### F.1. Robustness to Semantic Inputs

We evaluate the robustness of ReMoGen to high-level semantic inputs by testing two settings: (i) *no-text*, where no semantic signal is provided, and (ii) *shuffled-text*, where the input text conflicts with the physical interaction context.

As shown in Fig. B, ReMoGen remains stable in both settings. When semantic inputs are absent or inconsistent,

| Component                              | Time (s) |
|--|----------|
| Denoising (total, 10 cond + 10 uncond) | 0.29     |
| Denoiser                               | 0.067    |
| Meta-Interaction Modules               | 0.069    |
| per step                               | 0.0136   |
| Decoding                               | 0.005    |
| Frame-wise Segment Refinement          | 0.047    |
| decoding                               | 0.005    |
| FWSR module                            | 0.0017   |
| per iteration                          | 0.0067   |
| Pre/Post-processing (canonicalization) | ~0.02    |
| Total (8 frame or less)                | ~0.36    |
| per frame                              | ~0.047   |

Table B. Latency breakdown of ReMoGen.

interaction-driven cues dominate motion generation, leading to physically plausible reactions. This behavior is consistent with our design, where text is consumed only by the universal prior, while Meta-Interaction Modules are text-agnostic and react directly to dynamic interaction cues.

| Encoder Type | FID ↓ | R-Precision ↑ | MM Dist. ↓ | Div. → |
|--------------|-------|---------------|------------|--------|
| GT           | 0.000 | 0.472         | 4.051      | 4.084  |
| MLP          | 0.236 | 0.464         | 4.109      | 3.890  |
| Transformer  | 0.223 | 0.457         | 4.135      | 3.816  |
| TCN (Ours)   | 0.181 | 0.464         | 4.076      | 3.911  |

Table C. Encoder sensitivity evaluation on HHL.

### F.2. Robustness to Encoder Design

We evaluate the sensitivity of ReMoGen to different context encoders by replacing the TCN encoder with alternative architectures, including MLP and Transformer-based encoders.

As shown in Tab. C, almost all variants consistently outperform baselines (in Tab. 1), and the performance differences between encoder choices are relatively small. This indicates that the effectiveness of ReMoGen does not de-

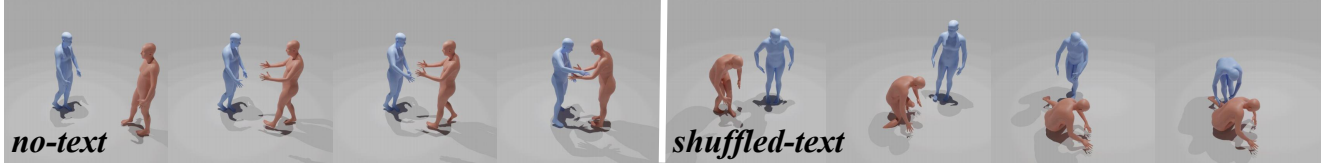


Figure B. Results under *no-text* and *shuffled-text* settings. In *no-text*, the original intent is “Shake hands”. In *shuffled-text*, the original intent is “Help others up” while the input is “Standing still”.

pend on a specific encoder design, but rather on the proposed modular framework and training strategy.

## G. Qualitative Experimental Results

This section presents extended qualitative evaluations of ReMoGen across Human-Human, Human-Scene, and Human-Human-Scene interaction scenarios. These visualizations complement the quantitative results summarized in Tab. 1, Tab. 2, Tab. 3, Tab. 4, and Tab. 5, and further illustrate how our modular learning design achieves stable, responsive, and semantically aligned motion generation.

### G.1. Comparison Results

**Human-Human Interaction.** We present two representative examples: “*One person approaches and opens their arms to **embrace** the back and waist of another person*” and “*One person **supports** the lower part of the other person’s left **arm** with both hands*”. These cases require fine-grained interpersonal coordination, accurate contact modeling, and temporally coherent reaction timing.

As shown in Fig. C and quantitatively validated in Tab. 1, ReMoGen consistently produces more coordinated, semantically accurate, and physically plausible reaction motions than competing methods. Baselines often exhibit unnatural timing, misaligned contact, or unstable body dynamics, whereas ReMoGen preserves interpersonal coordination and better matches ground-truth behavior.

**Human-Scene Interaction.** Fig. D, together with the quantitative scores in Tab. 2, demonstrates ReMoGen’s strong scene awareness.

We present two representative examples: “***stand up** from seat*” and “***walk back left** while holding small plant in right hand*”. Our model produces spatially consistent behaviors—such as standing up, walking with objects, or navigating around obstacles—while preserving motion coherence. Baselines struggle with scene grounding and often generate misaligned or unstable movements.

### G.2. Human-Human-Scene Results

**Zero-shot Results.** Fig. E demonstrates the zero-shot results. We present an example of the textual input “*The person explains on the blackboard*”. Due to the significant do-

main gap, all three zero-shot results show limited semantic alignment. Using only the HHI branch or only the HSI branch leads to incomplete behavior. However, the composition of interaction modules generates coherent in-scene interactions.

**Few-step Adaptation Results.** We present an example: “*Two people are engaged in a casual chat.*” This example includes motion implicitly involving the action of sitting down. This scenario requires the model to recognize mixed human-human-scene cues, maintain conversational posture alignment, and correctly represent the subtle transition into a seated position.

As shown in Fig. F and quantitatively verified in Tab. 3, ReMoGen rapidly adapts to the EgoBody dataset when initialized from the universal prior. With only 2k-10k finetuning steps, our model surpasses scratch-trained baselines that require hundreds of thousands of updates. The qualitative results illustrate that the adapted model produces stable, well-coordinated mixed-domain interactions.

**Results in Diverse Scene.** We present four representative examples illustrating the diversity and robustness of ReMoGen across complex multi-agent and scene-rich environments: “*Two individuals are engaged in a casual chat*”, “*The woman is learning TaiChi*”, “*A woman stretches her arms above her head*”, and “*A teacher is explaining on the blackboard*”. These scenarios span a wide range of interaction types—from individual scene-aware motions to multi-person collaborative activities.

Fig. G showcases ReMoGen’s ability to generate diverse, semantically rich interactions in complex scenes, including TaiChi movements, stretching, teaching, and conversation. These results qualitatively reflect the strong mixed-modality generalization trends reported in Tab. 3.

### G.3. Ablation Study Results

**Universal Prior Ablation.** Fig. H visualizes the effects observed quantitatively in Tab. 4. Models trained from scratch suffer from limited data and exhibit low diversity and unstable dynamics. Joint finetuning erodes pretrained kinematic knowledge. Only our modular prior-guided ap-

proach preserves both natural motion structure and interaction semantics.

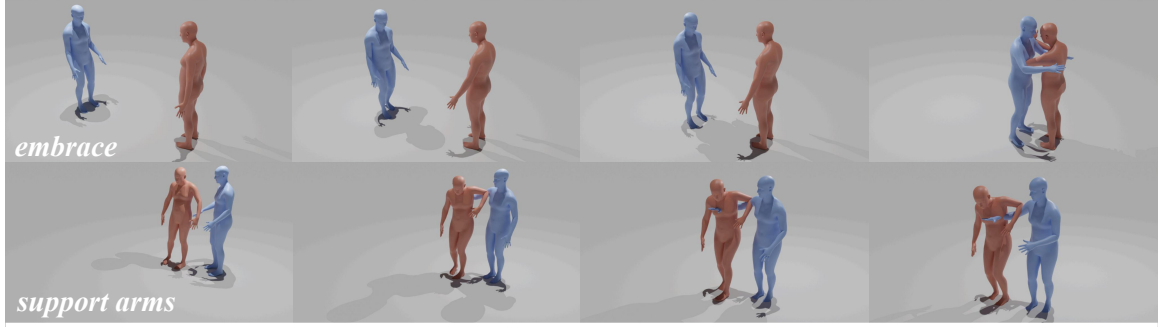
**Frame-wise Segment Refinement Ablation.** The qualitative comparisons in Fig. 1 correspond directly to the latency and accuracy trends in Tab. 5. Slide-style inference reacts quickly but introduces artifacts, while segment-only inference lacks responsiveness. Our Frame-wise Segment Refinement achieves a favorable balance—real-time responsiveness with stable, coherent motion.

## H. Limitations

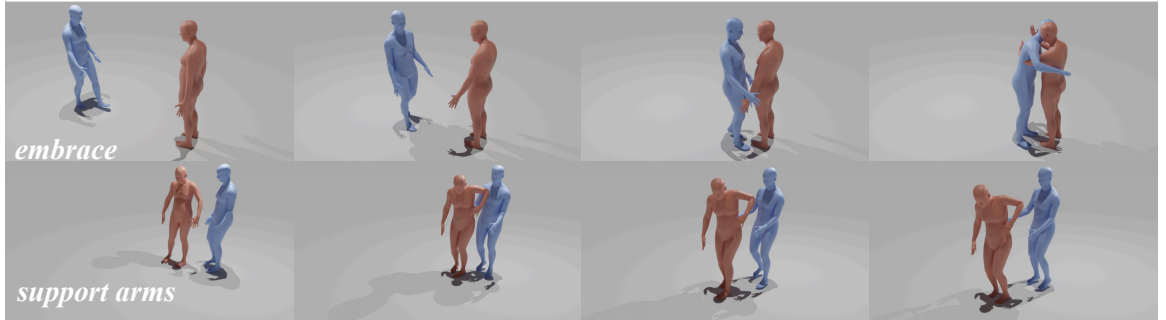
While ReMoGen achieves effective performance across heterogeneous interaction domains and supports real-time reaction generation, several limitations warrant further exploration. Our framework is built upon a latent diffusion prior that abstracts motion into a compressed latent space. While effective for generalization, this design may compromise fine-grained spatial accuracy, particularly in close-contact or high-precision interactions between humans and scene objects. While our current module composition strategy is simple and intuitive, more effective training-free fusion methods may further improve performance in mixed-interaction settings.



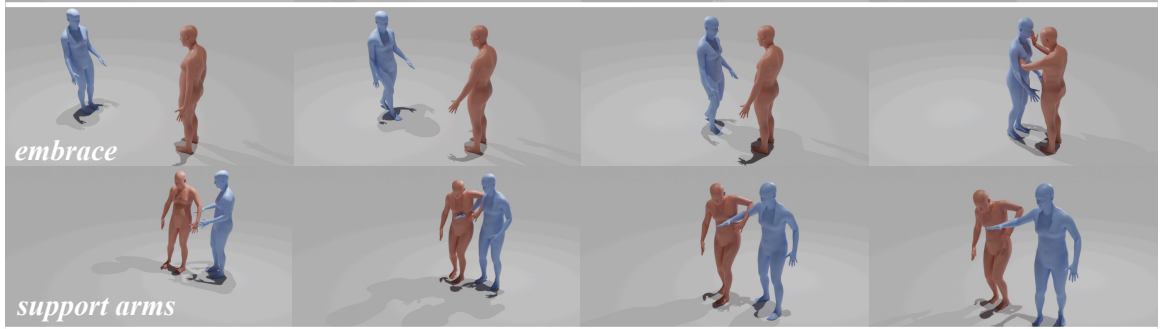
***ReGenNet***



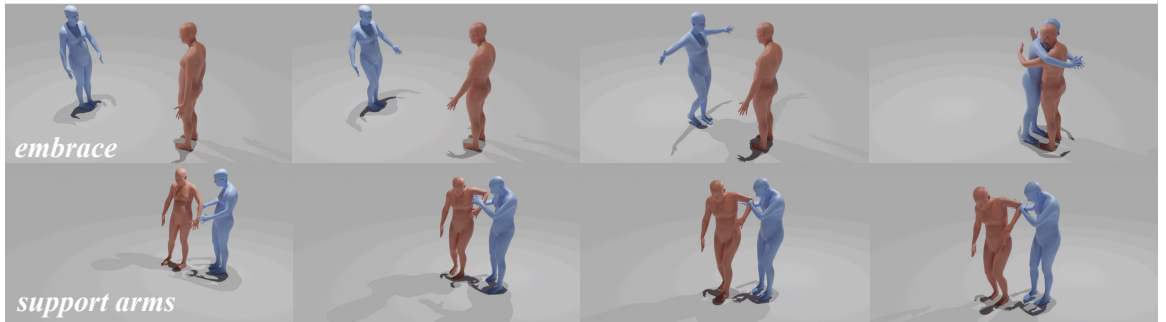
***FreeMotion  
(offline)***



***SymBridge***



***Ours***



***GT***



Figure C. Qualitative comparisons on Human–Human Interaction tasks. For typographical reasons, we have presented the optimal offline version of FreeMotion. Our method produces smoother and more coordinated reactions aligned with the intent, whereas baselines exhibit unnatural timing, misaligned contact, or unstable body dynamics.

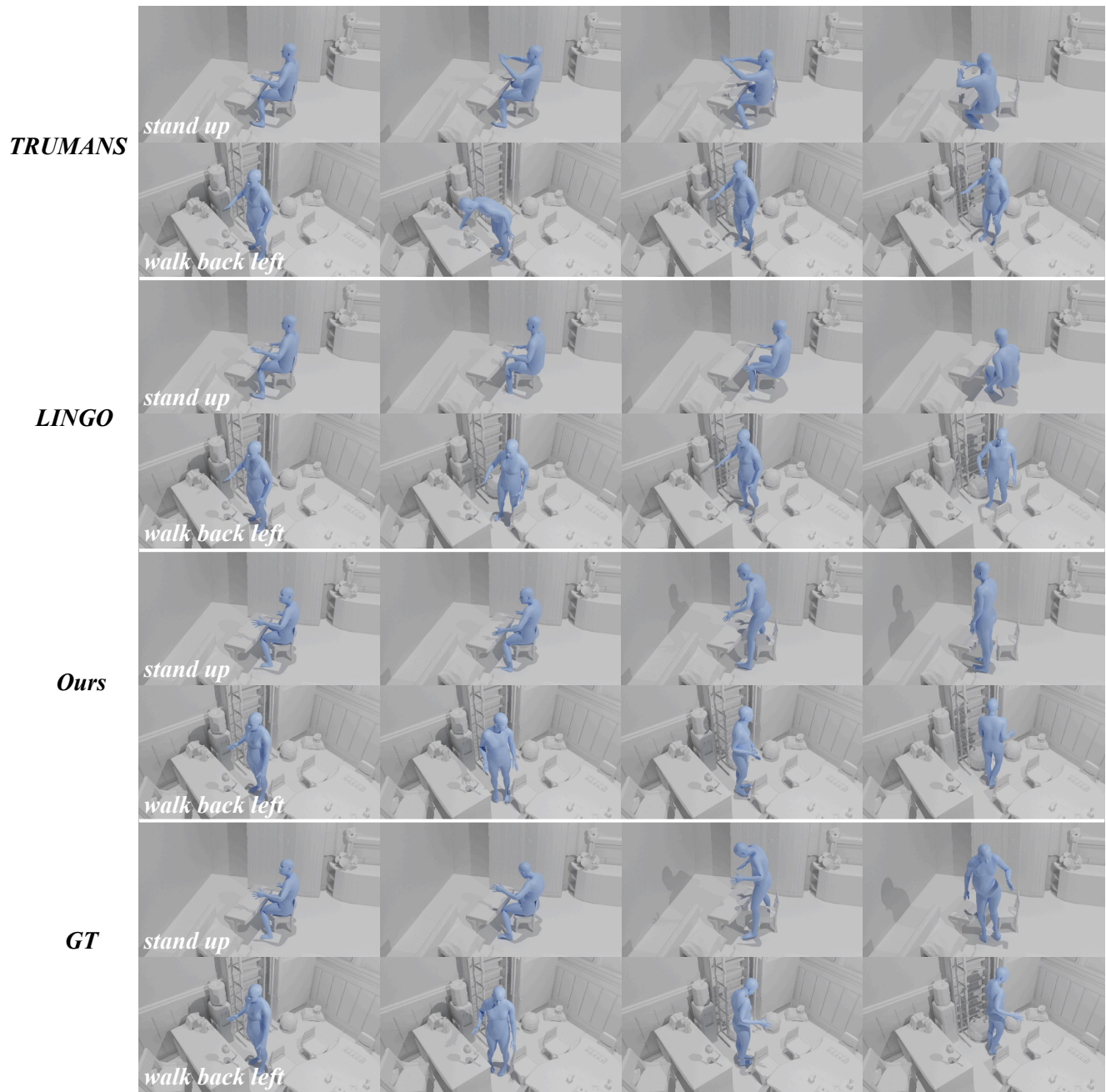


Figure D. Qualitative comparisons on Human–Scene Interaction tasks. All methods are evaluated with goal location provided.

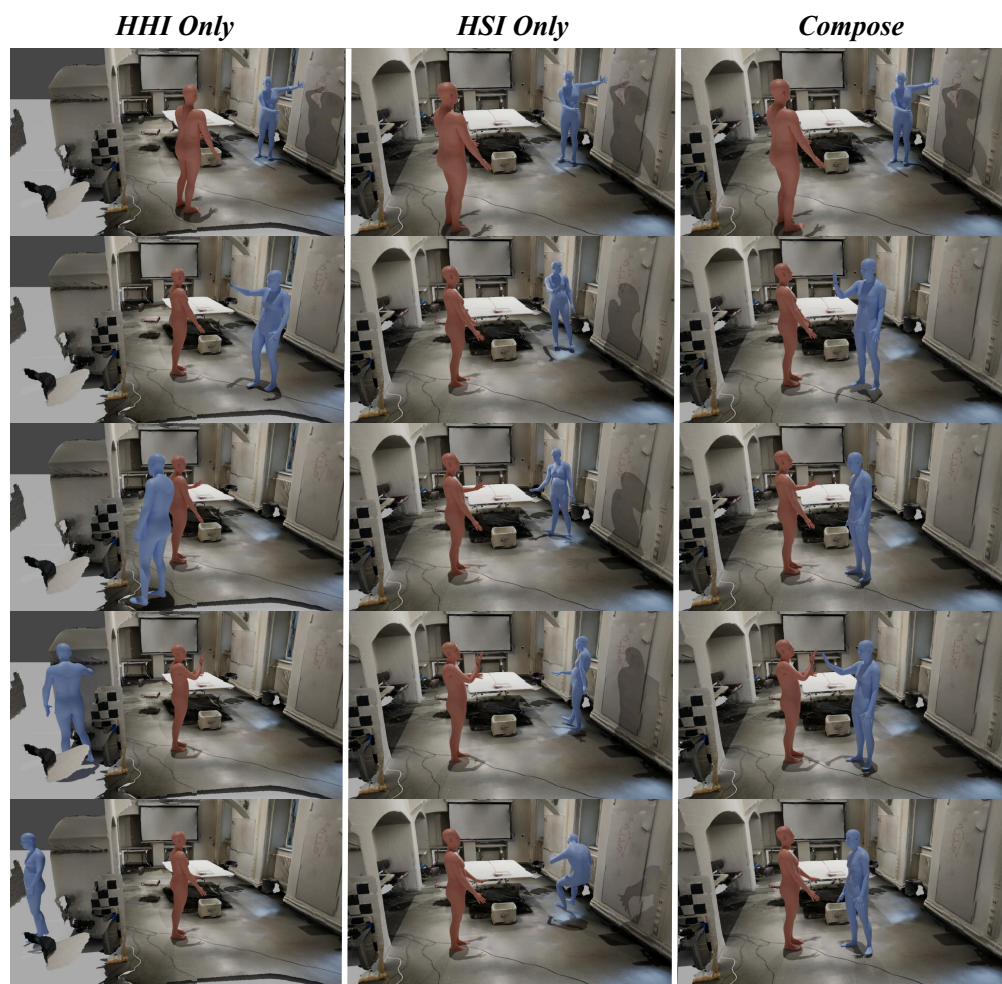


Figure E. Zero-shot Human–Human–Scene Interaction results.



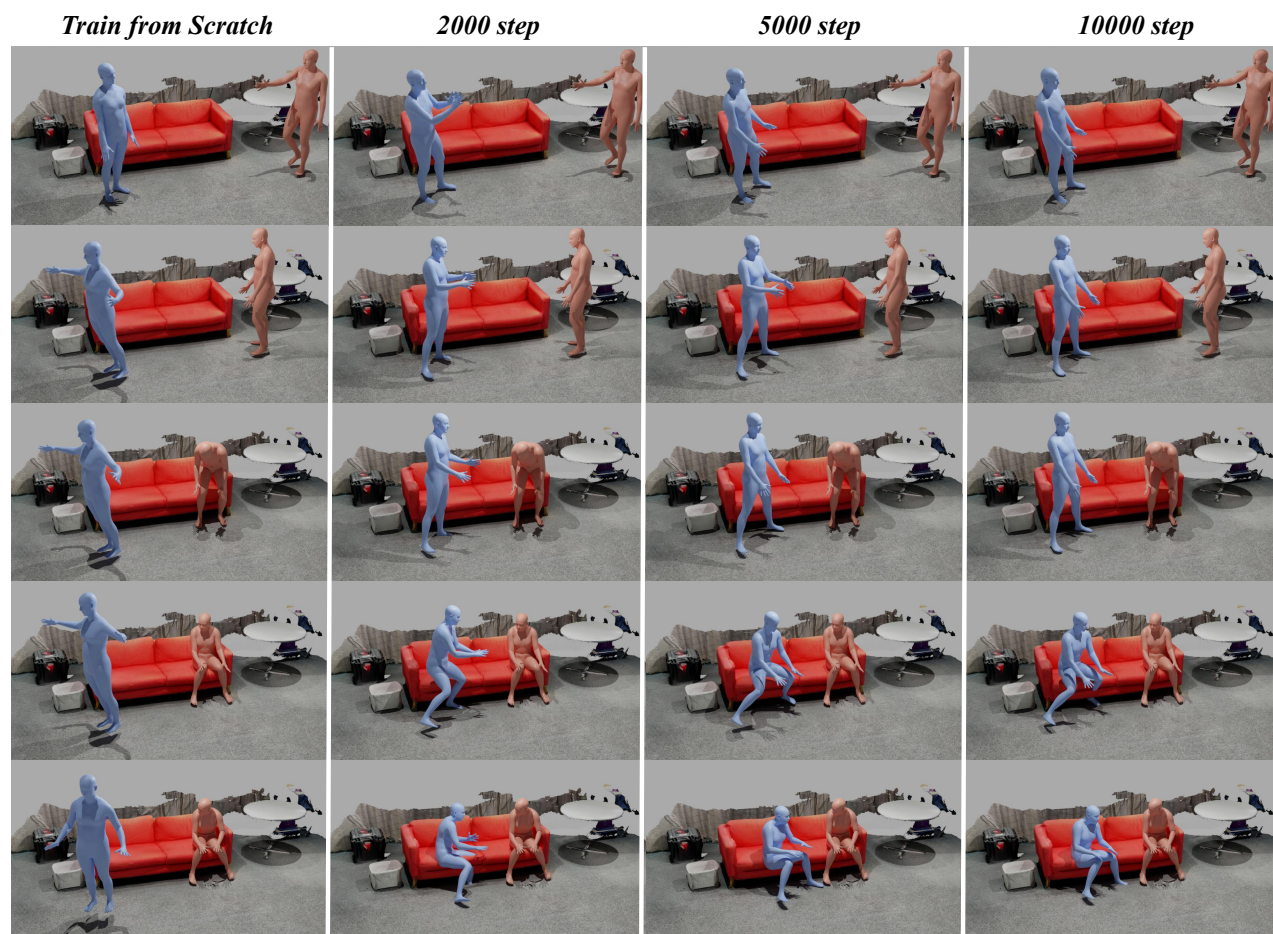


Figure F. Few-step finetuning on EgoBody. Initializing from our universal prior enables rapid adaptation, producing natural reactions after only a small number of updates.

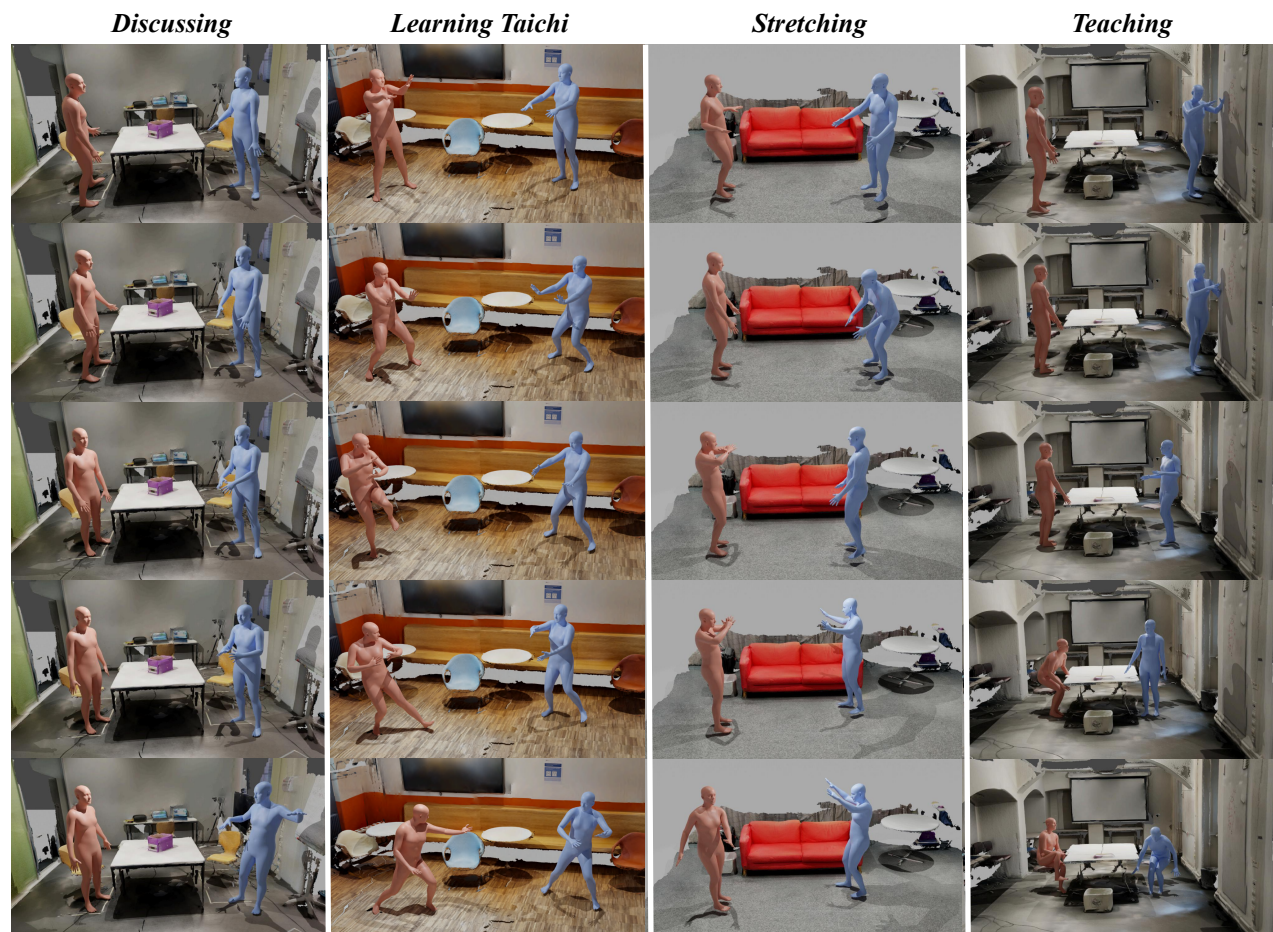


Figure G. Results in diverse Human–Human–Scene settings. ReMoGen generates semantically rich and varied interaction behaviors across different scenes and activity types. We present the results of finetuning 65k steps.

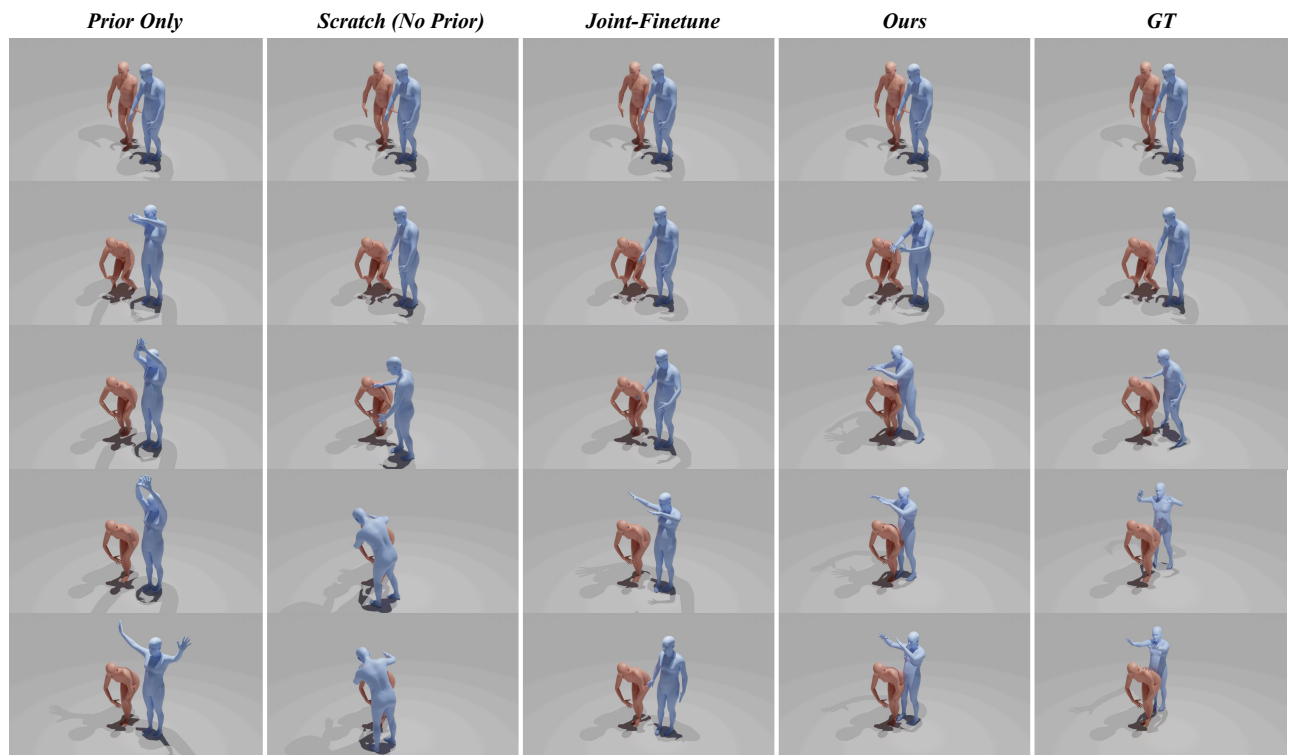


Figure H. Ablation on the Universal Motion Prior. We present “One person walks behind the other person and strikes a pose by extending both hands above the head.” Training from scratch or joint finetuning leads to unnatural or unstable reactions, while our prior-guided modular learning produces natural, coherent motion.



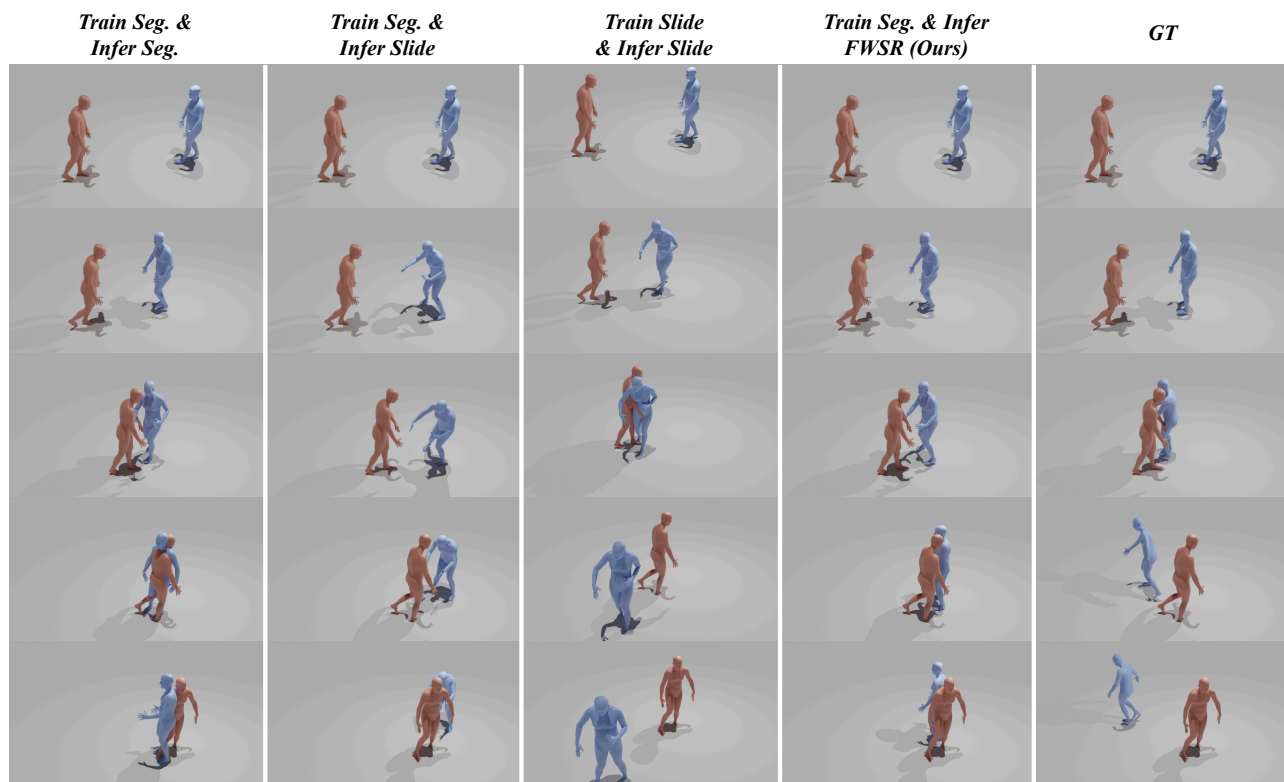


Figure I. Ablation on Frame-wise Segment Refinement. We present "A person runs towards someone and passes by them." FWSR provides fine-grained updates that improve responsiveness without sacrificing stability, outperforming both segment-only and naive slide-style inference.