Towards Immersive Human-X Interaction: A Real-Time Framework for Physically Plausible Motion Synthesis

Supplementary Material

Appendix

A. Overview

In this document, we provide additional technical details, extended experimental results, and further discussions that complement and elaborate upon the material in the main paper. Specifically, Section B offers a detailed description of our implementation, covering the training of both the motion diffusion model and the tracking policy, as well as clarifications on reproducing the baseline methods. In Section C, we present additional experiments—including ablation studies—and describe the evaluation metrics in greater depth. We also explore how our approach generalizes to other settings, illustrate further results with additional visualizations, and discuss findings from our user study. Finally, Section D lists extended failure cases of our method and offers insight into potential directions for future research. Through this supplementary material, we aim to provide a more comprehensive view of our approach, offer clarity on the nuances of the methodology, and furnish evidence of its robustness and versatility.

B. Supplementary Implementation Details

B.1. Details of Motion Diffusion Training

Scheduled training. To enhance stability and generalization of auto-regressive models, we use the scheduled training strategy [3, 48, 68, 73] to progressively supervise the training process with current sampling distribution. The interaction denoising network is trained on sequences of N interaction windows by integrating its own full-sampling predictions into the interaction history. Instead of always relying on the ground-truth data for past motion sequences, we substitute in the model's own earlier prediction. This approach let the model encounter an inference-time distribution, such as previously unseen interactions or unconventional combinations of actor states and text prompts.

To ease the transition to these more challenging scenarios, we adopt a three-phase training schedule:

- **Fully Supervised Phase**. The model is initially trained using only ground-truth motion history.
- **Mixed Training Phase** The ground-truth history is gradually replaced by rollout history with a probability that increases linearly from 0 to 1, allowing the model to slowly adjust to relying on its own outputs.
- **Rollout Training Phase** Finally, the model is trained exclusively using the sampled inteaction history.

In practice, we train each stage for 100K iteration steps and set the consecutive window number $N = 3$. The training algorithm is shown in Alg. 2.

B.2. Details of Reaction Policy Training

Actor-aware Reaction Policy Policy observations include generated and captured actor motion to adjust original imitation rewards and prevent interpenetration. The distances of every joint's positions are summed to get the adaptive weight $w(\hat{y}, y_{real})$, we use a linear interpolate between goal rewards depending on synthesis and realistic data. The rewards based on real action y_{real} are the imitation reward at default joint position and root distance reward up to 1, these rewards tend to keep the reactor away from the unpredictable actor,

$$r(s,a,s',\hat{x},\hat{y},y_{real}) = (1-w(\hat{y},y_{real})) * r^{\text{PHC}}(s,a,\hat{x}) + w(\hat{y},y_{real}) * (r^{\text{default}}(s,a) + r^{\text{root}}(s,y_{real})), \tag{10}$$

where $r^{\text{PHC}}(s^n, \hat{x}^n)$ is the reward components of PHC in full-motion imitation.

The imitation reward r^{PHC} proposed by PHC refers to $0.5r^g + 0.5r^{\text{amp}} + r^{\text{energy}}$. The weight $w(\hat{y}, y_{real})$ is derived from the cosine similarity S_C of 24 SMPL joints at frame i, then:

$$w(\hat{y},y_{real}) = \frac{1}{2}(1-\frac{1}{T}\sum\nolimits_{i=1}^{T} S_C(\hat{y}_i, y_i^{real})) \tag{11}$$

Once the prediction and capture conflict, w increases and the policy prioritizes higher rewards from the terms $r^{\text{defualt}} + r^{\text{root}}$. $r^{\text{defualt}} = 0.5e^{-100\|p_i^{\overline{x}} - p_i\|}$, representing the imitation reward with the average standing or walking motion $\overline{x}$

Algorithm 2 Scheduled training for auto-regressive interaction diffusion

```
Input: denoiser $\mathcal{G}_\theta$ with parameters $\theta$, reactor dataloader $\mathcal{X}$, actor dataloader $\mathcal{Y}$, text dataloader $\mathcal{C}$, total diffusion steps $T$, consecutive window number $N$, optimizer $\mathcal{O}$, training loss $\mathcal{L}$, max iteration $I_{max}$, auto-regressive interaction sampler $\mathcal{S}$.
$iter \leftarrow 0$
while $iter < I_{max}$ do
    $[\mathbf{x}^0, \mathbf{x}^1, ..., \mathbf{x}^N] \sim \mathcal{X}, [\mathbf{y}^0, \mathbf{y}^1, ..., \mathbf{y}^N] \sim \mathcal{Y}$
        ▷ sample $N$ interaction windows from dataset
    $\mathcal{Z} \leftarrow \{\textsc{Canonicalize}(\mathbf{x}^0, \mathbf{y}^0)\}$
        ▷ initialize interaction history
    for $i \leftarrow 1$ to $N$ do        ▷ number of rollouts
        $\mathbf{z}_0^i \leftarrow \textsc{Canonicalize}(\mathbf{x}^i, \mathbf{y}^i)$
        $t \sim \mathcal{U}[0, T)$
        $\mathbf{z}_t^i \leftarrow \textsc{Forward\_diffusion}(\mathbf{z}_0^i, t)$
        $c^i \leftarrow \textsc{Collect}(\mathcal{C}), \mathbf{z}^{i-1} \leftarrow \textsc{Collect}(\mathcal{Z})$
        $\hat{\mathbf{z}}_0^i = \mathcal{G}_\theta(\mathbf{z}_t^i, \mathbf{z}^{i-1}, t, c^i)$
            ▷ next interaction denoising
        $\nabla \leftarrow \nabla_\theta \mathcal{L}(\mathbf{z}_0^i, \hat{\mathbf{z}}_0^i, \mathbf{z}^{i-1})$
        $\theta \leftarrow \mathcal{O}(\theta, \nabla)$        ▷ back propagation
        $p \leftarrow \textsc{Schedule\_probability}(iter, I_{max})$
            ▷ sample schedule training probability
        if rand() $> p$ then
            $\mathbf{z}_T^i \leftarrow \textsc{Forward\_diffusion}(\mathbf{z}_0^i, T)$
                ▷ maximum noising
            $\hat{\mathbf{z}}_0^i = \mathcal{S}(\mathcal{G}_\theta, \mathbf{z}_T^i, \mathbf{z}^{i-1}, T, c^i)$
                ▷ full sampling loop
            $\hat{\mathbf{x}}_0^i = \textsc{Recover}(\hat{\mathbf{z}}_0^i)$
            $\tilde{\mathbf{z}}_0^i = \textsc{Canonicalize}(\hat{\mathbf{x}}_0^i, \mathbf{y}^i)$
            $\mathcal{Z} \leftarrow \mathcal{Z} \cup \tilde{\mathbf{z}}_0^i$
                ▷ set predicted reaction into history
        else
            $\mathcal{Z} \leftarrow \mathcal{Z} \cup \mathbf{z}_0^i$
                ▷ use dataset interaction history
        end if
        $iter \leftarrow iter + 1$
    end for
end while
```

as its goal (p: joint position). While r^{root} penalizes root proximity to the actor, with no reward given beyond 0.4 meters: $\max(\|p^{real}_{root,i} - p_{root,i}\|, 0.4)$. All the above process based on the precise capture for real actors. For noisy captures, we add slight noise to the real actor motion during policy training and simulate huge observation noises via mismatching generated actor and captured actor sequences. While domain randomization cannot fully erase every possible noise without additional observation modalities (e.g., inertial measurement units).

C. Additional Experiments and Results

C.1. Evaluations on InterHuman dataset

On the InterHuman dataset as shown in Tab 4, our method shows a slight decline in Div$_{cd}$ compared to CAMDM. We attribute this to the actor's movements constraining the reactor's actions: the incorporation of Interaction Loss reduces the distance between the actor and reactor to encourage contact. In contrast, CAMDM treats the actor as a conditioning factor, imposing fewer restrictions on the reactor's movements. More details about experiments within online text-guided reaction setting can be found in C.3.

C.2. Details of Evaluation Metrics

The evaluation framework for synthesized reactor sequences encompasses three principal dimensions: (1) FID, Diversity and Multimodal Distance (MMDist) for Reaction Quality, which assesses the reactor's motion independently of the actor, following[14, 58]; (2) Penetration, Floating and Skating for Physical Plausibility, which evaluates the adherence of the reactor's motion to physical constraints, following[59, 69]; and (3) Interpenetration Volume (IV), FID$_{cd}$ and Div$_{cd}$ for Interaction Quality which examines the realism of interactions between the actor and the reactor, following[36, 52]. All the motion and text features are extracted with the pretrained checkpoints in [14]. The metrics are defined in detail as follows.

FID. Frechet Inception Distance (FID) is a widely adopted metric for quantifying the quality of generated motion by measuring the statistical divergence between feature distributions of real and synthesized samples. It evaluates the fidelity of synthesized motion sequences by comparing their latent-space representations to those of ground-truth motion data.

Diversity. Diversity measures the variance of action categories across all generated motion sequences. Specifically, we sample two subsets of motion sequences, each containing the same number of samples S_d from the set of generated motion sequences across all action categories, denoted as $\{\mathbf{v}_1, ..., \mathbf{v}_{S_d}\}$ and $\{\mathbf{v}'_1, ..., \mathbf{v}'_{S_d}\}$. The diversity of the generated motions is then defined as Diversity $= \frac{1}{S_d}\sum_{i=1}^{S_d} \|\mathbf{v}_i - \mathbf{v}'_i\|_2$. In our experiments, we set $S_d = 200$. As the diversity of the generated motions approaches that of the real dataset, the generated motions exhibit greater diversity, leading to improved alignment with real-world motion distributions.

MMDist. Multimodal Distance (MMDist) evaluates the alignment fidelity between generated motion features and

Methods	Reaction			Physics			Interaction		
	FID↓	Div.→	MMDist. ↓	Pene. ↓	Skat.↓	Float. ↓	IV ↓	FID$_{cd}$ ↓	Div$_{cd}$ →
Ground Truth	0.008	3.865	4.306	0.000	0.032	21.273	0.045	0.532	9.109
InterFormer [8]	9.475	2.645	10.893	1.113	1.021	37.927	0.354	5.734	5.302
InterGen [32]	6.379	3.033	7.881	0.266	0.279	23.398	0.210	3.101	6.414
ReGenNet [66]	2.257	3.459	5.754	0.427	0.396	24.804	0.169	1.733	7.485
CAMDM [7]	2.166	4.161	6.212	0.296	0.177	24.261	0.341	3.701	**8.269**
Human-X	**1.995**	**3.767**	**5.638**	0.216	0.048	22.165	**0.106**	**1.582**	7.754
Human-X*	2.350	3.245	6.071	**0.064**	**0.008**	**12.026**	0.134	1.634	7.015

Table 4. **Action-to-Reaction** with online unconstrained reaction setting on **InterHuman**[32] dataset. A higher or lower value is better for ↑ or ↓, and → means the value closer to ground truth is better. * denotes the method with the physics tracker

their corresponding text features by measuring the average Euclidean distance. Formally, given N paired motion-text samples, we extract motion and text features using pretrained extractors, denoted as f_i^{motion} and f_i^{text} respectively. The MMDist is computed as: MMDist $= \frac{1}{N}\sum_{i=1}^{N}\left\|f_i^{motion} - f_i^{text}\right\|_2$, where lower MMDist indicates that the generated motion aligns more closely with the textual description.

Penetration, Floating and Skating. **Penetration** measures the distance between the lowest body mesh vertex below the ground and the ground surface, assessing whether the character exhibits ground penetration. **Floating** computes the distance between the lowest body mesh vertex above the ground and the ground surface, evaluating whether the character is unnaturally floating. **Skating** identifies foot joints that maintain ground contact across consecutive frames and calculates their average horizontal displacement, assessing the presence of foot sliding artifacts.

Interpenetration Volume. Interpenetration Volume (IV) measures the collision volume between the meshes of the actor and reactor, serving as a penalty term to discourage mesh interpenetration and unintended collisions between the two entities.

FID$_{cd}$ and Div$_{cd}$. In every frame, we select 10 key joints for both the actor and the reactor, chosen from the complete set of available joints. The selected joints encompass the pelvis, knees, feet, shoulders, head, and two wrists. A matrix $M \in \mathbb{R}^{10\times10}$ is then constructed based on the pairwise distances between the selected joints of the two agents, serving as the interactive feature. Based on this representation, we compute the FID$_{cd}$ and Div$_{cd}$, which are used to supervise and assess the interaction quality between the two characters.

C.3. Additional Experiments on Diffusion Planner

In the online text-guided reaction setting, we also conduct experiments on the Inter-X and InterHuman datasets, evaluating performance across three key dimensions: Reaction Quality, Physical Plausibility, and Interaction Quality, as shown in Tab. 5 and 6. Unlike the experiments in the unconstrained reaction setting, we exclude InterFormer from the baselines since it does not support text-conditioned inputs. The results demonstrate that our approach outperforms previous baselines across all metrics, achieving state-of-the-art performance. Incorporating text supervision enables the generation of diverse and fine-grained human motion sequences, allowing for detailed customization based on the provided descriptions, while also leading to slight improvements in various evaluation metrics. However, it is noteworthy that even without textual input, our model already achieves strong performance. Although the inclusion of text information enhances the results, the overall improvement is relatively modest. This indicates that the model remains robust and effective even in the absence of additional textual guidance.

In additional ablation studies, we conduct experiments on four aspects: motion representation, the size of binary interaction field $\mathcal{I}^n$, numbers of the decoder layers l_{layers} and text encoder. All experimental results are presented in Tab. 7.

Representation. Similar to InterGen[32], we attempt to use a non-canonical representation for multi-person interaction motion. However, experimental results show a performance degradation across all metrics, with a particularly significant drop in interaction-related metrics. This performance degradation is primarily due to the fact that the non-canonical representation encodes features in the global coordinate system. In contrast, our approach defines the coordinate origin at the root joint of the reactor and represents motion in a relative coordinate system. This relative formu-

Methods	Reaction			Physics			Interaction		
	FID↓	Div.→	MMDist. ↓	Pene. ↓	Skate. ↓	Float. ↓	IV ↓	FID$_{cd}$ ↓	Div$_{cd}$ →
Ground Truth	0.002	6.028	3.524	0.000	0.023	7.956	0.024	0.235	11.471
InterGen [32]	5.211	4.498	6.070	0.257	0.136	11.319	0.256	3.265	8.867
ReGenNet [66]	2.101	5.096	4.973	0.134	0.112	9.320	0.207	1.853	9.359
CAMDM [7]	1.359	5.676	5.052	0.145	0.121	9.371	0.138	1.982	9.140
Human-X	**0.926**	**5.951**	**3.909**	**0.112**	**0.087**	**8.218**	**0.072**	**1.607**	**9.822**

Table 5. **Action-to-Reaction** with online text-guided reaction setting on **Inter-X** [65] dataset, where a higher or lower value is better for ↑ or ↓, and → means the value closer to ground truth is better.

Methods	Reaction			Physics			Interaction		
	FID↓	Div.→	MMDist. ↓	Pene. ↓	Skat. ↓	Float.↓	IV ↓	FID$_{cd}$ ↓	Div$_{cd}$ →
Ground Truth	0.008	3.865	4.306	0.000	0.032	21.270	0.045	0.532	9.109
InterGen [32]	6.060	3.366	7.486	0.253	0.265	22.221	0.200	2.946	6.549
ReGenNet [66]	2.144	3.621	5.466	0.406	0.376	23.560	0.161	1.646	7.566
CAMDM [7]	2.058	4.043	5.901	0.281	0.168	23.047	0.324	3.516	8.311
Human-X	**1.889**	**3.807**	**5.356**	**0.205**	**0.046**	**21.052**	**0.101**	**1.513**	**8.384**

Table 6. **Action-to-Reaction** of online text-guided reaction setting on **InterHuman** [32] dataset, where a higher or lower value is better for ↑ or ↓, and → means the value closer to ground truth is better.

lation facilitates learning the spatial relationships between the two characters, whereas the global representation makes it substantially more challenging for the model to capture these interactions effectively.

When removing binary interaction field $\mathcal{I}$ and angular velocity of the roots $\dot{\mathbf{r}}^y$, temporal difference of the local joint positions $\dot{\mathbf{p}}^y$ separately, the model's ability to learn the authenticity of motion interaction and the temporal coherence between consecutive frames decreases, which naturally leads to a degradation in generation quality. However, introducing 6D representation of the joint rotations θ^y also results in performance degradation. This is because θ^y does not contribute to the model's understanding of the interaction between the two characters; instead, the inclusion of such redundant information hinders the model's comprehension ability.

$\mathcal{I}$ Field Size. The size of the binary interaction field $\mathcal{I}$ refers to the number of joints selected in the contact map. When only the pelvis joint is selected, $\mathcal{I}$ fails to be effective, as the pelvis itself rarely engages in contact. On the other hand, selecting all 22 joints achieves the best performance on most metrics but introduces significant computational latency. Therefore, we ultimately select the six most critical joints, including the pelvis, head, both ankles, and both wrists, to ensure interaction quality while minimizing computational overhead.

Num of l_{layers}. We conduct experiments with 2, 4, 8, and 16 decoder layers and ultimately select 8 layers as the final choice, as it demonstrated the best overall performance.

Text Encoder. In ablation study, we attempt to use CLIP as the text encoder. As stated in CLOSD[59], CLIP[47] tends to focus on understanding image descriptions. Although it outperforms DistilBERT[51] in MMDist, it performs slightly worse on other metrics. Therefore, we ultimately adopt DistilBERT as the text encoder.

C.4. Additional Experiments on Reaction Policy

In this section, we present additional experimental results on the Inter-X dataset, comparing our actor-aware reaction policy with the baseline method PHC[38]. The results are shown in Tab. 8. IV stands for interpenetration volume and other metrics following PHC. During inference, we disable the inter-actor collision avoidance in simulation, otherwise, the interpenetration volume (IV) will constantly be zero. Our method achieves a significant reduction in interpenetration volume and a notable improvement in success rate, while maintaining comparable performance in other metrics. This demonstrates that our actor-aware reaction policy

Class	Settings	Reaction			Physics			Interaction			Latency
		FID↓	Div.→	MMDist. ↓	Pene. ↓	Skat. ↓	Float. ↓	IV ↓	FID_{cd} ↓	Div_{cd} →	(ms)
	Ground Truth	0.002	6.028	3.524	0.000	0.023	7.956	0.024	0.235	11.471	-
Representation	non-canonical	2.032	6.285	4.278	0.129	0.131	8.814	0.083	5.737	7.853	10.2
	w.o. $\mathcal{I}$	1.108	5.842	4.235	0.121	0.095	8.689	0.078	3.542	8.423	13.1
	w.o. $\dot{\mathbf{r}}^y$ and $\dot{\mathbf{p}}^y$	1.044	5.617	4.259	0.125	0.094	8.712	0.080	2.389	9.168	12.8
	add θ^y	1.363	5.956	4.116	0.117	0.089	8.634	0.075	1.864	9.705	15.0
$\mathcal{I}$ Field Size	1 × 1	1.095	5.945	4.305	0.133	0.102	8.916	0.087	1.802	9.223	11.9
	10 × 10	0.979	6.092	4.153	0.119	0.094	8.667	0.077	1.701	**9.781**	13.9
	22 × 22	0.982	6.067	4.118	0.121	**0.089**	8.650	**0.069**	1.697	9.399	15.1
Num of l_{layers}	2	2.729	5.611	4.222	0.124	0.096	8.701	0.079	1.827	7.952	3.8
	4	1.379	5.868	4.117	0.118	0.093	8.654	0.076	1.734	9.538	7.2
	16	1.007	5.935	4.183	0.121	0.094	8.677	0.078	1.705	9.654	25.5
Text Encoder	CLIP[47]	1.023	6.100	**4.073**	0.120	0.095	8.680	0.079	1.712	9.652	13.6
	Human-X	**0.975**	**6.063**	4.115	**0.118**	0.092	**8.650**	0.076	**1.694**	9.735	13.6

Table 7. Additional **ablation studies** with online reaction settings on the **Inter-X**[65] dataset, where a higher or lower value is better for ↑ or ↓, and → means the value closer to ground truth is better.

Methods	IV↓	Succ↑	E_{mpjpe}↓	E_{acc}↓	E_{vel}↓
PHC[33]	4.7	84.1%	47.6	11.7	9.1
Ours	1.4	**95.6%**	**37.3**	**10.5**	**4.2**
Ours-safety	**0.05**	84.0%	51.7	11.4	10.1

Table 8. Test performance of reaction policy on Inter-X dataset. (Ours-safety indicates actor-aware policy).

effectively enhances the reactor's motion quality and interaction realism.

C.5. User Study

Metrics	Question
Diversity	Which interaction leads to a greater variety of reactions?
Consistency	Which interaction produces more realistic reactions?
Authenticity	Which interaction exhibits more realistic contact?

Table 9. Question settings for user questionnaire.

In order to assess the effectiveness of the proposed method, we conduct a user study involving 15 participants. They are asked to complete a questionnaire, with the specific questions provided in Tab 9. In the first phase of the experiment, participants are instructed to compare the video results generated by our method against those produced by **InterGen**[32], **ReGenNet**[66] and the ground truth, based on three key criteria: Diversity, Consistency, and Authenticity. In the second phase, participants will engage in an immersive interaction with the avatar using a VR headset. Following this experience, they evaluate our method in comparison to **CAMDM**[7] and select the motion sequence they perceive as superior. Each participant is presented with 20 video samples and allocated an average of 30 seconds for the VR experience, ensuring sufficient exposure to the generated results before providing their final assessment.

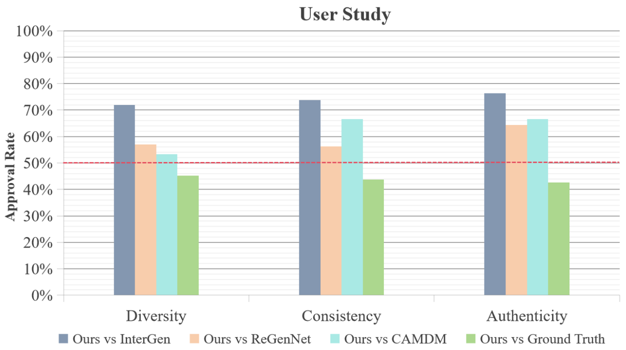

Figure 5. Extensive experimental results indicate that participants perceive our method to perform better in all three metrics: Diversity, Consistency, and Authenticity.

The final results, as presented in Figure 5, demonstrate that our method achieved the highest approval rate across all metrics among the participants, reaching state-of-the-art performance. This result validates that our approach not only ensures latency-free performance but also maintains high-quality motion generation.

D. Extended Limitation and Discussions

D.1. Discussion and Future Work

In this work, we introduce a novel auto-regressive framework that seamlessly integrates predictive interaction synthesis with actor-aware physical refinement. Meanwhile, we integrate our framework into a real-time VR system, demonstrating its effectiveness in immersive, unconstrained environments. Through comparative experiments, we identify considerable areas where our model can be further improved, as outlined below:

Further Memory and Planning. Currently, our model relies on the past 20 frames to predict the next 40 frames. However, this short temporal window may result in the loss of critical historical information. A potential direction could be leveraging the planning and decision-making capabilities of Large Language Models (LLMs) to guide the model[4, 33, 63], allowing it to incorporate a longer temporal context for more informed decision-making and improved generation quality.

Multi-Modal Action and Response. For now, our model is limited to generating the reactor's motion based solely on the actor's motion and textual input. However, in real-world scenarios, additional modalities such as audio[26, 61] and visual cues[11, 21] play a crucial role in motion decision-making. Developing an interactive system capable of processing multimodal inputs and outputs would enhance its generalizability and expand its potential applications.

Reactive Character Generalizability. The reactor can be made more diverse. In the future, we can extend its representation from the SMPL[37] to SMPL-X[43] and SMPL+H[50], enabling finer-grained control over facial expressions, hand gestures, and body shape. Additionally, the reactor can be replaced with humanoid robots[5, 6, 24], laying the groundwork for real-world deployment.

Various Interaction Context. Although we have achieved real-time interaction between two individuals, real-time interaction among multiple participants presents a significantly greater challenge. While systems like [4] can generate interactions between two individuals in specific scenarios through text-based action control, they lack the capability for real-time reaction generation. Implementing real-time, two-person interactions within specific contexts, or integrating dual participants with object interaction, remains an area for future exploration.

Customized Reaction Design. In real-life scenarios, individuals possess distinct personalities, leading to varied responses to the same action. Inspired by [4], we can assign personality traits to the reactor in the future, enabling it to generate personalized responses that cater to users' individual needs.