

Interact2Ar: Full-Body Human-Human Interaction Generation via Autoregressive Diffusion Models

Supplementary Material

This supplementary material aims to enhance the reproducibility and understanding of the work contributions. In Sec. A, we outline the implementation details of the state-of-the-art models used for comparison, provide detailed formulations of the loss functions employed in training, describe the newly proposed body-part-specific evaluators, detail the user study methodology, and explain the implementation of adaptive interaction capabilities. In Sec. B, we complement the quantitative evaluation with results using the original evaluators from the Inter-X dataset, present an extended ablation study examining different memory configurations across all evaluation settings, evaluate the impact of different text encoders, and introduce additional dyadic-specific metrics to assess interaction quality and physical plausibility, including foot sliding. In Sec. C, we describe the accompanying Supplementary Video, which includes additional visual examples and side-by-side comparisons with previous state-of-the-art methods, alongside new close-up visualizations of complex hand and body contacts, to better illustrate the Interact2Ar capabilities. Finally, in Sec. D, we detail the code and data availability to ensure full reproducibility of our work.

A. Implementation Details

A.1. State-of-the-art Implementations

We compared Interact2Ar with previous SOTA methods on the Inter-X dataset. We primarily compared against T2M [1], InterGen [5], and InterMask [2].

InterGen is the state-of-the-art baseline among the original baselines proposed by the Inter-X [13] dataset authors. However, given that the weights are not public, we re-trained InterGen using the original implementation details described in the paper, including a transformer encoder with 8 blocks and 8 heads (latent dimension 512, feed-forward dimension 1024). The model uses 1000 steps with DDIM-50 sampling and was trained for 5000 epochs using EMA and AdamW. Following Inter-X we used a learning rate 1×10^{-4} with weight decay 2×10^{-5} and a batch size of 128. Given that the motion representation that we use is not the same as in the InterHuman dataset [5], we used our loss adaptations using forward kinematics to train the model.

A.2. Losses

In this section, we provide detailed formulations and explanations for each component of our training loss function.

Representation Loss. The representation loss $\mathcal{L}_{\text{repr}}$ directly measures the ℓ_2 distance between the predicted and ground

truth SMPL-X parameters [8] in their raw representation space:

$$\mathcal{L}_{\text{repr}}(x, \hat{x}) = \|x - \hat{x}\|_2^2, \quad (1)$$

where $x \in \mathbb{R}^{T \times D}$ represents the ground truth SMPL-X parameters across T frames with dimensionality D , and \hat{x} denotes the predicted parameters. This loss operates directly on the body pose parameters, hand articulations, and global trajectory, providing a direct supervision signal in the learned representation space.

Root Orientation Loss. The root orientation loss $\mathcal{L}_{\text{orient}}$ specifically penalizes errors in the global root orientation of each individual:

$$\mathcal{L}_{\text{orient}}(r, \hat{r}) = \|r_a - \hat{r}_a\|_2^2 + \|r_b - \hat{r}_b\|_2^2, \quad (2)$$

where $r_a, r_b \in \mathbb{R}^{T \times 3}$ represent the ground truth root orientations for individuals a and b respectively, and \hat{r}_a, \hat{r}_b are the corresponding predictions. This loss ensures that the global facing direction and body orientation of each person are accurately captured, which is crucial for modeling proper spatial relationships in interactions.

A.2.1. Kinematic Losses

We compute the following geometric losses through forward kinematics (FK), which converts SMPL-X parameters to 3D joint positions: $p = \text{FK}(x)$.

Joint Position Loss. The global joint position loss \mathcal{L}_{pos} penalizes discrepancies in the predicted 3D locations of body joints:

$$\mathcal{L}_{\text{pos}}(p, \hat{p}) = \|p_a - \hat{p}_a\|_2^2 + \|p_b - \hat{p}_b\|_2^2, \quad (3)$$

where $p_a, p_b \in \mathbb{R}^{T \times N_j \times 3}$ are the ground truth global joint positions for both individuals with N_j joints per person, and \hat{p}_a, \hat{p}_b are the corresponding predicted positions. This loss enforces spatial accuracy in the generated motions.

Joint Velocity Loss. To promote temporal smoothness and physical plausibility, we apply a velocity loss on the joint positions:

$$\mathcal{L}_{\text{vel}}(v, \hat{v}) = \|v_a - \hat{v}_a\|_2^2 + \|v_b - \hat{v}_b\|_2^2, \quad (4)$$

where $v_t = p_t - p_{t-1}$ represents the joint velocities computed as the difference between consecutive frames. This loss discourages unnatural jittering and encourages smooth, realistic motion trajectories.

Foot Contact Loss. The foot contact loss $\mathcal{L}_{\text{foot}}$ reduces artifacts such as foot skating and floating:

$$\mathcal{L}_{\text{foot}}(f, \hat{f}) = \sum_{i \in \{\text{feet}\}} \|v_i \odot f_i\|_2^2 + \|\hat{v}_i \odot \hat{f}_i\|_2^2, \quad (5)$$

where v_i denotes the velocity of foot joint i , $f_i \in \{0, 1\}^T$ is a binary contact indicator (1 when the foot is in contact with the ground, 0 otherwise), and \odot represents element-wise multiplication. This loss penalizes foot motion when contact is detected, enforcing physical constraints.

Pairwise Joint Distance Map Loss. To capture the fine-grained spatial relationships between the two individuals, we introduce the distance map loss $\mathcal{L}_{\text{dist}}$:

$$\mathcal{L}_{\text{dist}}(d, \hat{d}) = \|(D(p_a, p_b) - D(\hat{p}_a, \hat{p}_b)) \odot M\|_2^2, \quad (6)$$

where $D(p_a, p_b) \in \mathbb{R}^{T \times N_j \times N_j}$ computes the pairwise Euclidean distance between all joints of individual a and all joints of individual b :

$$D(p_a, p_b)_{i,j} = \|p_a^{(i)} - p_b^{(j)}\|_2, \quad (7)$$

with $p_a^{(i)}$ and $p_b^{(j)}$ denoting the positions of the i -th joint of person a and j -th joint of person b , respectively. The binary mask $M \in \{0, 1\}^{T \times N_j \times N_j}$ activates the loss only for joint pairs in close proximity in the ground truth, focusing supervision on spatial relationships that are most critical for realistic interactions. This ensures that the spatial proximity patterns between the two individuals match the ground truth, which is essential for generating realistic interactive behaviors such as handshakes, hugs, and other contact-based interactions.

Loss Weighting. The weighting coefficients $\{\lambda_{\text{repr}}, \lambda_{\text{orient}}, \lambda_{\text{pos}}, \lambda_{\text{vel}}, \lambda_{\text{foot}}, \lambda_{\text{dist}}\}$ are determined through grid search to balance the contribution of each loss term. These weights are calibrated to normalize the magnitude differences across loss components, ensuring that each term contributes meaningfully to the optimization process. The specific values used in our experiments are: $\lambda_{\text{repr}} = 1.0$, $\lambda_{\text{orient}} = 0.1$, $\lambda_{\text{pos}} = 1.0$, $\lambda_{\text{vel}} = 1.0$, $\lambda_{\text{foot}} = 1.0$, and $\lambda_{\text{dist}} = 0.5$.

A.3. Evaluators

Sec. 4 introduced an improved evaluation pipeline over the original evaluator provided on the Inter-X dataset. This new pipeline better assesses interaction quality and provides more granular information. To achieve this, we introduced a set of new body-part-specific evaluators retrained to have deeper knowledge of the global information of the interactants.

For all evaluators, we used the architecture proposed in [1], where a motion and a text feature extractor are trained via contrastive learning, and these encoded representations are used to calculate the remaining metrics. Using this architecture, we trained 3 evaluators for 300 epochs at a learning rate of 1×10^{-4} to generate feature vectors of size 512. The full evaluator was trained using information from all SMPL-X joints, the body evaluator using only the base SMPL [6] joints, and the hand evaluator using the additional 30 joints used for hands.

We additionally made the evaluators more robust, as demonstrated in Tab. 1. Based on the findings of [7], we decided to train an evaluator using only joint positions. Since the positioning between different individuals has great importance for interactions, we represented joint positions using global coordinates. These coordinates are calculated using a forward kinematic function on SMPL-X rotations predicted by our model.

A.4. User Study

The user study was performed with 35 different participants to rank 10 different interactions extracted from: ground truth, our model, InterMask, and InterGen. The 35 participants were in the range of 25 to 55 years old, from different nationalities, all having higher degrees of study (bachelor's or more). Among the participants, there was a similar distribution of individuals familiarized with the human motion generation task and not. Fig. A presents a real frame from one of the videos that the users had to rank. In the video, there is a textual description at the top, and there are 4 videos randomly shuffled for each of the possible options. From each video, the participant had to rank each interaction based on the alignment with the textual description and the quality of the hand generation. We also included the ground truth provided by the dataset for this textual description, so the user always had an aligned interaction with the text and could rank all videos based on the overall quality. To ensure even distribution of the interaction motions, all the textual descriptions were extracted from the test set, and every one pertained to a different action category.

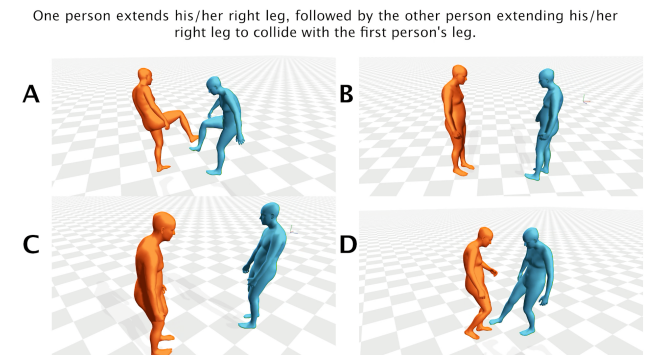


Figure A. **User study.** A sample of the user study where 35 participants evaluated and ranked the text alignment and the hand quality of Interact2Ar and baseline methods.

A.5. Adaptive Interactions

Temporal Motion Composition. We implement this with a large context window that includes all previously generated actions. Given the Mixed Memory approach we proposed, the memory buffer \mathcal{M} accesses this information and

	Methods	R-Precision \uparrow			FID \downarrow	MM Dist \downarrow	Diversity \rightarrow	MModality \uparrow	PJ \uparrow	AUJ \downarrow
		Top 1	Top 2	Top 3						
Full	<i>Ground Truth</i>	0.429 \pm .00	0.626 \pm .00	0.736 \pm .00	0.002 \pm .00	3.536 \pm .01	9.734 \pm .08	—	0.021 \pm .00	3.944 \pm .00
	T2M [1]	0.184 \pm .01	0.298 \pm .01	0.396 \pm .01	5.481 \pm .38	9.576 \pm .01	2.771 \pm .15	2.761 \pm .04	1.889 \pm .12	124.9 \pm 17.9
	InterGen [5]	0.327 \pm .00	0.506 \pm .00	0.619 \pm .00	2.601 \pm .10	4.580 \pm .02	9.200 \pm .11	3.999 \pm .11	2.132 \pm .12	84.16 \pm 3.1
	InterMask [2]	0.403 \pm .01	0.595 \pm .00	0.705 \pm .01	0.399 \pm .01	3.705 \pm .02	9.046 \pm .07	2.261 \pm .08	2.328 \pm .21	61.74 \pm .50
	Interact2Ar*	0.433 \pm .00	0.623 \pm .00	0.727 \pm .00	0.255 \pm .01	3.618 \pm .01	9.066 \pm .07	2.820 \pm .08	2.110 \pm .12	54.97 \pm .67
	Interact2Ar	0.441 \pm .00	0.631 \pm .00	0.737 \pm .00	0.148 \pm .01	3.581 \pm .01	9.147 \pm .06	1.529 \pm .05	0.136 \pm .00	8.837 \pm .17
Body	<i>Ground Truth</i>	0.431 \pm .00	0.621 \pm .00	0.725 \pm .00	0.002 \pm .00	3.371 \pm .01	8.931 \pm .06	—	0.014 \pm .00	3.832 \pm .00
	T2M [1]	0.310 \pm .00	0.468 \pm .00	0.570 \pm .00	3.346 \pm .05	4.465 \pm .02	8.113 \pm .09	0.604 \pm .03	1.857 \pm .10	115.6 \pm 17.7
	InterGen [5]	0.349 \pm .00	0.528 \pm .00	0.637 \pm .00	1.708 \pm .05	4.363 \pm .02	9.253 \pm .09	3.596 \pm .11	2.116 \pm .13	77.09 \pm 3.9
	InterMask [2]	0.401 \pm .00	0.594 \pm .00	0.707 \pm .00	0.741 \pm .03	3.483 \pm .01	8.915 \pm .08	1.589 \pm .05	2.549 \pm .16	59.41 \pm .69
	Interact2Ar*	0.447 \pm .00	0.643 \pm .00	0.750 \pm .00	0.273 \pm .02	3.231 \pm .01	9.074 \pm .08	2.347 \pm .08	2.040 \pm .17	55.30 \pm .51
	Interact2Ar	0.446 \pm .00	0.639 \pm .00	0.744 \pm .00	0.212 \pm .01	3.287 \pm .02	9.055 \pm .08	1.470 \pm .05	0.123 \pm .00	6.620 \pm .13
Hands	<i>Ground Truth</i>	0.372 \pm .00	0.556 \pm .00	0.663 \pm .00	0.002 \pm .00	3.893 \pm .01	8.611 \pm .07	—	0.017 \pm .00	2.480 \pm .00
	T2M [1]	0.325 \pm .00	0.486 \pm .00	0.590 \pm .00	2.114 \pm .05	4.296 \pm .02	8.129 \pm .06	0.595 \pm .04	1.767 \pm .09	113.1 \pm 17.9
	InterGen [5]	0.331 \pm .00	0.504 \pm .00	0.617 \pm .00	4.664 \pm .19	4.560 \pm .03	9.654 \pm .12	4.155 \pm .12	2.049 \pm .12	70.47 \pm 2.9
	InterMask [2]	0.380 \pm .00	0.568 \pm .00	0.681 \pm .00	0.383 \pm .02	3.729 \pm .02	8.689 \pm .08	1.883 \pm .09	2.201 \pm .16	59.77 \pm .76
	Interact2Ar*	0.402 \pm .00	0.592 \pm .00	0.704 \pm .00	0.242 \pm .01	3.639 \pm .02	8.972 \pm .08	2.877 \pm .08	1.923 \pm .09	51.03 \pm .49
	Interact2Ar	0.393 \pm .00	0.584 \pm .00	0.695 \pm .00	0.238 \pm .01	3.711 \pm .01	8.853 \pm .07	1.754 \pm .05	0.120 \pm .00	7.474 \pm .16

Table A. Comparison of our model (Interact2Ar) to the state of the art in human-human interaction motion generation on the Inter-X dataset. *Interact2Ar model is the version without autoregressive generation. All evaluations have been executed 20 times to elude the randomness of the generation. \pm indicates the 95% confidence interval. We highlight the **best** and the second best results.

enables Interact2Ar to generate seamless transitions while accounting for the complete action history.

Real-Time Disturbance Adaptation. To effectively assess the real-time adaptation of Interact2Ar, we randomly translated one individual in the XZ plane between different sub-motion generations. This simulates noisy contexts and disturbances produced by the environment or other individuals. Traditional diffusion models and Masked VQ-VAE Transformers, such as InterMask, cannot enable this capability because they produce the whole sequence at once, preventing adaptation until the entire motion is generated.

Sequential Multi-Person Interactions. We implemented this using two couples performing 2 different actions with their respective memories. Once they finish their actions, we take one individual from each couple and generate a new interaction with the newly formed couple. For the memories, we retain the original memories from the initial couples and create a new memory using information from the new couple. This enables seamless interaction while maintaining access to previously performed actions. While this implementation only generates sequential multi-human interactions, the idea can be expanded to generate parallel multi-human interactions as proposed in [14].

B. Quantitative Evaluation

B.1. Original Evaluators

In Tab. 1, we present a quantitative evaluation of the robustness of our newly proposed evaluators with respect to the

original ones provided in the Inter-X dataset. Additionally, we provide the main metrics of our model in Tab. 2 using those newly trained evaluators. In Tab. A, we performed the same evaluation using the original full-body evaluator trained directly with the SMPL-X representation, alongside body- and hand-specific evaluators using the same representation. We can observe in this case that Interact2Ar still outperforms previous methods. However, these differences are not as large as when using our evaluators. It can even be observed that in the body- and hand-specific evaluators, the non-autoregressive version of Interact2Ar obtains slightly better metrics than the autoregressive one. These smaller differences occur because rotation-based evaluators penalize diffusion models compared to VQ-VAE approaches. Furthermore, the original evaluators do not account for degradations in the global positioning of the individuals, which makes them incapable of detecting small differences, such as those between the autoregressive and non-autoregressive versions.

B.2. Extended Memory Ablation

In Tab. 3, we present an extended ablation study where different memory configurations have been tested to determine which provides the best trade-off between quality in terms of metrics and memory size. Tab. B is an extended version of this ablation where all the different evaluators have been used. As can be observed, the overall tendency that we observed for the full evaluator remains consistent in the body and hands evaluators. While adding more memory can re-

	Method	m_s	m_l	\mathcal{M}	R-Precision \uparrow			FID \downarrow	MM Dist \downarrow	Diversity \rightarrow	MModality \uparrow
					Top 1	Top 2	Top 3				
Full	<i>Ground Truth</i>	-	-	-	0.431 \pm .00	0.631 \pm .00	0.740 \pm .00	0.002 \pm .00	3.318 \pm .01	8.973 \pm .10	-
	Interact2Ar Regular Memory	15	-	15	0.449 \pm .00	0.653 \pm .00	0.765 \pm .00	0.283 \pm .01	3.141 \pm .01	9.275 \pm .08	1.491 \pm .05
		30	-	30	0.445 \pm .00	0.656 \pm .00	0.769 \pm .00	0.346 \pm .01	3.122 \pm .01	9.240 \pm .06	1.471 \pm .04
		60	-	60	0.458 \pm .00	0.665 \pm .00	0.776 \pm .00	0.316 \pm .01	3.071 \pm .02	9.285 \pm .07	1.394 \pm .06
		90	-	90	0.452 \pm .00	0.660 \pm .01	0.771 \pm .00	0.412 \pm .01	3.103 \pm .01	9.264 \pm .07	1.344 \pm .04
		120	-	120	0.456 \pm .00	0.665 \pm .00	0.774 \pm .00	0.413 \pm .01	3.084 \pm .01	9.299 \pm .07	1.336 \pm .04
	Interact2Ar Mixed Memory	15	15	18	0.448 \pm .00	0.655 \pm .00	0.769 \pm .00	0.289 \pm .01	3.131 \pm .01	9.255 \pm .06	1.461 \pm .05
		15	45	24	0.453 \pm .00	0.661 \pm .00	0.773 \pm .00	0.277 \pm .01	3.095 \pm .01	9.305 \pm .07	1.427 \pm .04
		15	75	30	0.453 \pm .00	0.661 \pm .00	0.771 \pm .00	0.279 \pm .01	3.110 \pm .01	9.318 \pm .06	1.346 \pm .05
		15	105	36	0.458 \pm .00	0.664 \pm .00	0.773 \pm .00	0.325 \pm .01	3.089 \pm .01	9.346 \pm .06	1.363 \pm .04
Body	<i>Ground Truth</i>	-	-	-	0.462 \pm .01	0.655 \pm .00	0.758 \pm .00	0.001 \pm .00	3.272 \pm .01	9.002 \pm .07	-
	Interact2Ar Regular Memory	15	-	15	0.466 \pm .00	0.667 \pm .00	0.774 \pm .00	0.399 \pm .01	3.194 \pm .02	9.319 \pm .09	1.483 \pm .05
		30	-	30	0.471 \pm .00	0.672 \pm .00	0.780 \pm .00	0.428 \pm .01	3.168 \pm .01	9.195 \pm .06	1.439 \pm .05
		60	-	60	0.474 \pm .01	0.677 \pm .00	0.781 \pm .00	0.397 \pm .01	3.158 \pm .01	9.244 \pm .08	1.388 \pm .05
		90	-	90	0.471 \pm .00	0.678 \pm .00	0.781 \pm .00	0.475 \pm .01	3.188 \pm .01	9.260 \pm .07	1.337 \pm .04
		120	-	120	0.470 \pm .00	0.676 \pm .00	0.781 \pm .00	0.463 \pm .01	3.168 \pm .01	9.311 \pm .07	1.304 \pm .04
	Interact2Ar Mixed Memory	15	15	18	0.471 \pm .00	0.676 \pm .00	0.780 \pm .00	0.354 \pm .01	3.169 \pm .01	9.268 \pm .07	1.442 \pm .04
		15	45	24	0.469 \pm .00	0.672 \pm .00	0.779 \pm .00	0.352 \pm .01	3.173 \pm .01	9.271 \pm .08	1.421 \pm .04
		15	75	30	0.474 \pm .00	0.673 \pm .00	0.778 \pm .00	0.331 \pm .01	3.175 \pm .01	9.276 \pm .06	1.362 \pm .05
		15	105	36	0.477 \pm .00	0.678 \pm .00	0.782 \pm .00	0.411 \pm .01	3.141 \pm .01	9.321 \pm .06	1.343 \pm .04
Hands	<i>Ground Truth</i>	-	-	-	0.399 \pm .00	0.597 \pm .00	0.713 \pm .00	0.002 \pm .00	3.312 \pm .01	8.370 \pm .05	-
	Interact2Ar Regular Memory	15	-	15	0.414 \pm .00	0.620 \pm .00	0.734 \pm .00	0.206 \pm .01	3.165 \pm .01	8.580 \pm .07	1.540 \pm .06
		30	-	30	0.420 \pm .01	0.627 \pm .00	0.744 \pm .00	0.302 \pm .01	3.129 \pm .01	8.515 \pm .07	1.495 \pm .06
		60	-	60	0.425 \pm .00	0.630 \pm .00	0.748 \pm .00	0.273 \pm .01	3.112 \pm .01	8.595 \pm .07	1.405 \pm .06
		90	-	90	0.423 \pm .01	0.630 \pm .01	0.747 \pm .00	0.350 \pm .01	3.126 \pm .01	8.607 \pm .07	1.354 \pm .05
		120	-	120	0.424 \pm .00	0.634 \pm .01	0.748 \pm .00	0.378 \pm .01	3.111 \pm .01	8.589 \pm .08	1.370 \pm .05
	Interact2Ar Mixed Memory	15	15	18	0.421 \pm .00	0.628 \pm .00	0.743 \pm .00	0.258 \pm .01	3.135 \pm .01	8.551 \pm .08	1.504 \pm .06
		15	45	24	0.422 \pm .00	0.629 \pm .00	0.745 \pm .00	0.257 \pm .01	3.111 \pm .01	8.614 \pm .08	1.439 \pm .05
		15	75	30	0.424 \pm .00	0.632 \pm .00	0.745 \pm .00	0.245 \pm .01	3.124 \pm .01	8.608 \pm .06	1.372 \pm .05
		15	105	36	0.430 \pm .00	0.635 \pm .00	0.748 \pm .00	0.269 \pm .01	3.114 \pm .01	8.576 \pm .07	1.384 \pm .05

Table B. **Ablation study on memory configurations** for Interact2Ar across different evaluation settings. m_s and m_l represent the context window used for each memory. $m_l = -$ indicates models not using Mixed Memory. For models using Mixed Memory, $\delta = 5$. The total number of frames used in the full memory is $\mathcal{M} = m_s + m_l/\delta$.

sult in more informative generations, it also increases the complexity of the task that the denoiser has to learn, resulting in a non-linear improvement of metrics as the memory size increases. However, what is more noticeable is the significant increase in FID, which is generally used to determine motion quality, when using Mixed Memory.

B.3. Text Encoder Ablation

CLIP [9] is used as the text encoder for the textual descriptions of the interactions, which are injected into the model as conditions. This decision was made to ensure consistency with previous works [2, 5, 10]. To evaluate the impact of a newer and more advanced encoder, Tab. C presents a comparison against Qwen3-VL-Embedding-2B [3]. As can be observed, the results show minimal differences, which can likely be attributed to the limited diversity of textual descriptions present in the dataset.

B.4. Additional Interaction Metrics

In addition to the standard metrics present in the Inter-X benchmark, Tab. D presents supplementary metrics from related tasks to provide further insights into interaction quality. Specifically, Contact Frequency [11] measures the ratio of frames where interactants are in contact, FID_{CD} [11] computes the FID using a feature vector derived from the pairwise distances of joints, and Interaction Volume Penetration [4] calculates the average penetration volume per sequence between the individuals involved in the interaction. As can be observed, our proposed method consistently achieves the best performance across these additional metrics.

Foot Sliding. Qualitative examples indicate that all evaluated methods suffer from foot sliding, a limitation that could be addressed with additional data or post-processing. Nevertheless, to quantitatively measure the physical plausibility of foot contacts, we include the physical foot contact (PFC)

Methods	R-Precision (Top 3) \uparrow	FID \downarrow	MM Dist \downarrow	Diversity \rightarrow	MModality \uparrow
<i>Ground Truth</i>	0.740 \pm .00	0.002 \pm .00	3.318 \pm .01	8.973 \pm .10	—
CLIP [9]	0.773 \pm .00	0.277 \pm .01	3.095 \pm .01	9.305 \pm .07	1.427 \pm .04
Qwen3-VL-Embedding-2B [3]	0.728 \pm .00	0.389 \pm .02	3.361 \pm .03	9.230 \pm .14	1.637 \pm .04

Table C. **Ablation study on the text encoder** used to encode textual conditions in Interact2Ar. We compare CLIP and Qwen3-VL-Embedding-2B using the original Inter-X evaluation metrics. The **best** and second best results are highlighted.

Methods	Contact Frequency \rightarrow	FID _{CD} \downarrow	Interaction Volume Penetration \downarrow	PFC \downarrow
<i>Ground Truth</i>	33.827 \pm 2.67	—	0.086 \pm .00	0.120 \pm .00
T2M [1]	47.975 \pm 2.52 (+14.1%)	4.037 \pm .46	0.768 \pm .06	2.454 \pm .19
InterGen [5]	14.867 \pm 1.47 (-19.0%)	4.999 \pm .41	0.441 \pm .02	4.834 \pm .25
InterMask [2]	<u>21.881</u> \pm 2.01 (-11.9%)	5.593 \pm .33	<u>0.437</u> \pm .03	<u>0.339</u> \pm .03
Interact2Ar	38.703 \pm 2.27 (+4.9%)	2.406 \pm .55	0.360 \pm .08	0.268 \pm .02

Table D. **Evaluation with additional dyadic-specific metrics.** Contact Frequency calculates the ratio of frames where interactants are in contact, FID_{CD} computes the FID using features derived from pairwise joint distances, Interaction Volume Penetration measures the average penetration volume per sequence, and PFC evaluates the physical plausibility of foot contacts. The **best** and second best results are highlighted.

score [12] in Tab. D. The results demonstrate that our approach yields the most realistic foot movements.

C. Qualitative Evaluation

In addition to the qualitative examples shown in Fig. 5 and Fig. 6, we introduce a new set of examples and comparisons in the Supplementary Video. Due to the 4D nature of the representation that we generate, static images present a significant information loss. In the video, the quality of the motion and the side-by-side comparisons will facilitate understanding and highlight the qualitative differences between Interact2Ar and the previous SOTA, InterMask. However, to complement these dynamic results, Fig. B provides additional close-up static visualizations that further demonstrate the capability of Interact2Ar to generate realistic hand interactions involving complex body and hand contacts.

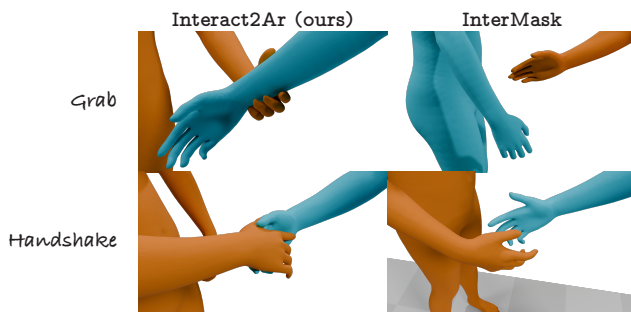


Figure B. **Close-up visualizations.** Zoomed-in views of hands during challenging interactions involving body and hand contacts.

D. Code and Data Availability

All the code and checkpoints related to this paper will be publicly released upon the acceptance of this paper. The

code will contain all the codebase used to declare, train, and evaluate Interact2Ar and the new set of evaluators. The checkpoints will include the Interact2Ar checkpoints alongside the checkpoints of the evaluators for providing a more robust and reliable evaluation of future works using Inter-X.

References

- [1] 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*, pages 5152–5161, 2022. 1, 2, 3, 5
- [2] Muhammad Gohar Javed, Chuan Guo, Li Cheng, and Xingyu Li. InterMask: 3d human interaction generation via collaborative masked modeling. *arXiv preprint arXiv:2410.10010*, 2024. 1, 3, 4, 5
- [3] Mingxin Li, Yanzhao Zhang, Dingkun Long, Chen Keqin, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026. 4, 5
- [4] Ronghui Li, Youliang Zhang, Yachao Zhang, Yuxiang Zhang, Mingyang Su, Jie Guo, Ziwei Liu, Yebin Liu, and Xiu Li. Interdance: Reactive 3d dance generation with realistic duet interactions. *arXiv preprint arXiv:2412.16982*, 2024. 4
- [5] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. InterGen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, 132(9):3463–3483, 2024. 1, 3, 4, 5
- [6] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2

- [7] Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. Rethinking diffusion for text-driven human motion generation: Redundant representations, evaluation, and masked autoregression. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27859–27871, 2025. [2](#)
- [8] 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)*, pages 10975–10985, 2019. [1](#)
- [9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. [4](#), [5](#)
- [10] Pablo Ruiz-Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and José García-Rodríguez. in2in: Leveraging individual information to generate human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1941–1951, 2024. [4](#)
- [11] Li Siyao, Tianpei Gu, Zhitao Yang, Zhengyu Lin, Ziwei Liu, Henghui Ding, Lei Yang, and Chen Change Loy. Duolando: Follower gpt with off-policy reinforcement learning for dance accompaniment. *arXiv preprint arXiv:2403.18811*, 2024. [4](#)
- [12] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 448–458, 2023. [5](#)
- [13] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-x: Towards versatile human-human interaction analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22260–22271, 2024. [1](#)
- [14] Wenning Xu, Shiyu Fan, Paul Henderson, and Edmond SL Ho. Multi-person interaction generation from two-person motion priors. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, 2025. [3](#)