ARMFlow: AutoRegressive MeanFlow for Online 3D Human Reaction Generation

Zichen Geng[1], Zeeshan Hayder[2], Wei Liu[1], Hesheng Wang [3]*, Ajmal Saeed Mian[1]
[1]The University of Western Australia, Perth, WA, Australia
[2]Commonwealth Scientific and Industrial Research Organisation (CSIRO), Canberra, ACT, Australia
[3]Shanghai Jiao Tong University, Shanghai, China
zen.geng@research.uwa.edu.au
{wei.liu, ajmal.mian}@uwa.edu.au
zeeshan.hayder@data61.csiro.au
wanghesheng@sjtu.edu.cn

Abstract

3D human reaction generation faces three main challenges: (1) high motion fidelity, (2) real-time inference, and (3) autoregressive adaptability for online scenarios. Existing methods fail to meet all three simultaneously. We propose ARMFlow, a MeanFlow-based autoregressive framework that models temporal dependencies between actor and reactor motions. It consists of a causal context encoder and an MLP-based velocity predictor. We introduce Bootstrap Contextual Encoding (BSCE) in training, encoding generated history instead of the ground-truth ones, to alleviate error accumulation in autoregressive generation. We further introduce the offline variant ReMFlow, achieving state-of-the-art performance with the fastest inference among offline methods. Our ARMFlow addresses key limitations of online settings by: (1) enhancing semantic alignment via a global contextual encoder; (2) achieving high accuracy and low latency in a single-step inference; and (3) reducing accumulated errors through BSCE. Our single-step online generation surpasses existing online methods on InterHuman and InterX by about 30% in FID, while matching offline state-of-the-art performance despite using only partial sequence conditions. The official implementation is publicly available at: https://github.com/ZenGengChin/armflow_official

1. Introduction

Recent advances in generative modeling have led to remarkable progress in 3D human motion generation, covering a wide spectrum of tasks and methodologies. These include text-guided motion synthesis [4, 11, 13, 34, 38, 41, 49–51], human–object interaction [6, 9, 20, 32, 47], scene-conditioned motion generation, and multi-person interaction modeling. Among these, one particularly distinctive task is human reaction generation, which focuses on producing reactive human behaviors in response to other agents or stimuli. This task holds immediate practical potential in human–robot interaction, augmented, and virtual realities.

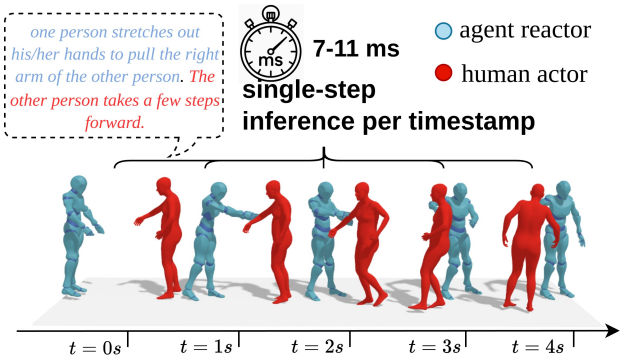

Figure 1. Our method only processes a single inference in each real-time step for online reaction generation, compared to the SOTA methods ReGenNet (35-78 ms), and CAMDM (45 ms). The text description is from the InterX [45] dataset.

Unlike *offline motion generation* tasks that rely on predefined conditions and tolerate second-level latency, *3D human reaction generation* demands real-time responsiveness, where the input condition evolves continuously and unpredictably. In such settings, even minimal computational delays can compromise the system's reactivity and realism. This imposes stricter requirements on the generation framework, introducing three key challenges: (1) achieving high inference efficiency to meet real-time constraints; (2) maintaining high-fidelity motion quality to ensure natural and expressive reactions; and (3) capturing long-term contextual dependencies to preserve semantic consistency and generate accurate, context-aware motions.

Existing approaches have made attempts to address these

*Corresponding author

challenges, yet fundamental limitations remain. To enable online generation, autoregressive architectures are essential, as they inherently model temporal dependencies between past and future motions. Methods such as CAMDM [3] and HumanX [18] adopt autoregressive diffusion models, where a fixed-length historical window serves as the conditioning context for current denoising steps. However, this design faces two key limitations: (1) the fixed context window hinders scalability and leads to inevitable information loss, causing semantic drift over long sequences; and (2) although accelerated samplers such as DDIM have been adopted, they still require multiple denoising steps (typically $\geq$8), which is computationally demanding, especially under fine-grained temporal resolutions.

To address the aforementioned three challenges and two key limitations of prior works, we propose ARMFlow, a scalable autoregressive architecture capable of generating high-fidelity human reactions in a *single-step inference*. Our method is built upon the recently proposed MeanFlow [8] paradigm, which enables one-step generation as opposed to multi-step denoising or iterative integration required by diffusion [9, 26] or traditional flow-based models. This is the first work that leverages MeanFlow for human motion generation, demonstrating the fastest inference speed while maintaining superior motion realism compared to existing approaches. To overcome the limitations of fixed-length contextual windows commonly used in prior autoregressive models, we introduce a causal context encoder that encodes the entire motion history with causal masking. This design prevents the loss of information beyond the context window and preserves global temporal semantics, ensuring coherent long-term motion generation.

Furthermore, we observe that training the model with only clean past motions makes it overly sensitive to noise during autoregressive generation, as it never learns to handle imperfect or accumulated prediction errors in its historical context. To mitigate this issue, we propose Bootstrap Context Encoding (BSCE), where the model uses predicted motion histories—rather than ground-truth ones during training to construct the contextual conditions for the flow velocity predictor. As training progresses, the predicted motion sequences gradually approximate real trajectories, leading to an adaptive curriculum that naturally reduces context noise over time. We further increase the number of bootstrap iterations throughout training, effectively introducing controlled noise and enhancing model robustness against accumulated prediction errors. This mechanism accelerates convergence and improves autoregressive stability. Benefiting from MeanFlow's single-step inference, BSCE can efficiently generate augmented history samples without costly iterative denoising, resulting in significantly improved training efficiency. Together, these components form a self-consistent and efficient autoregressive generation framework based on MeanFlow dynamics.

Beyond the online setting, we further design a general offline variant, Reaction MeanFlow (ReMFlow), which serves as a versatile baseline for offline motion generation. Compared with existing state-of-the-art (SOTA) offline models, ReMFlow achieves superior generation quality and the fastest inference speed on both InterHuman and InterX benchmarks. More importantly, our online model ARMFlow not only achieves SOTA performance in real-time settings, notably over 40% on FID, but also performs on par with, and in some cases surpasses, many offline models that have access to the entire conditioning sequence simultaneously. In summary, our contributions are threefold:

1. We propose ARMFlow, a flexible and scalable method consisting of a context encoder with an MLP velocity predictor that captures global semantic alignment in an autoregressive manner.
2. We extend the MeanFlow paradigm to the domain of 3D reaction generation, introducing a unified framework for both online (ARMFlow) and offline (ReMFlow) settings that achieves single-step, high-fidelity motion synthesis with real-time performance.
3. We propose a Bootstrap Context Encoding (BSCE) mechanism that effectively mitigates error accumulation in autoregressive generation, speeds up the model's convergence, and enhances inference robustness.

2. Related Works

Reaction Generation: 3D human motion generation is currently an active area of research, encompassing tasks such as Text-to-Motion [4, 11, 13, 33, 34, 38, 41, 49, 51], Human-Object Interaction [6, 9, 20, 32, 42, 47], to Human-Human Interaction [17, 23, 35, 43] generation. However, Action-Reaction Generation is still an understudied area. Tab. 1 lists current SOTA models for reaction generation.

Table 1. Current reaction generation models. AR: Autoregression.

Model	Online	Real-time	Long-context	AR
InterMask [17]	✗	✗	✓	✗
InterGen [23]	✗	✗	✓	✗
MARRS [44]	✗	✓	✓	✗
CAMDM [3]	✓	✓	✗	✓
ReGenNet [46]	✓	✓	✗	✗
HumanX [18]	✓	✓	✗	✓
Ours	✓	✓	✓	✓

Chopin et al. [5] first proposed Interformer, a Transformer-based [40] reaction synthesis model, but due to the use of traditional autoencoders to predict current actions, its prediction accuracy is less than ideal. Xu et al. [46] introduced the first online Transformer-decoder-based diffusion model, aiming to achieve both efficiency and accuracy in diffusion via DDIM [36]. Although this method

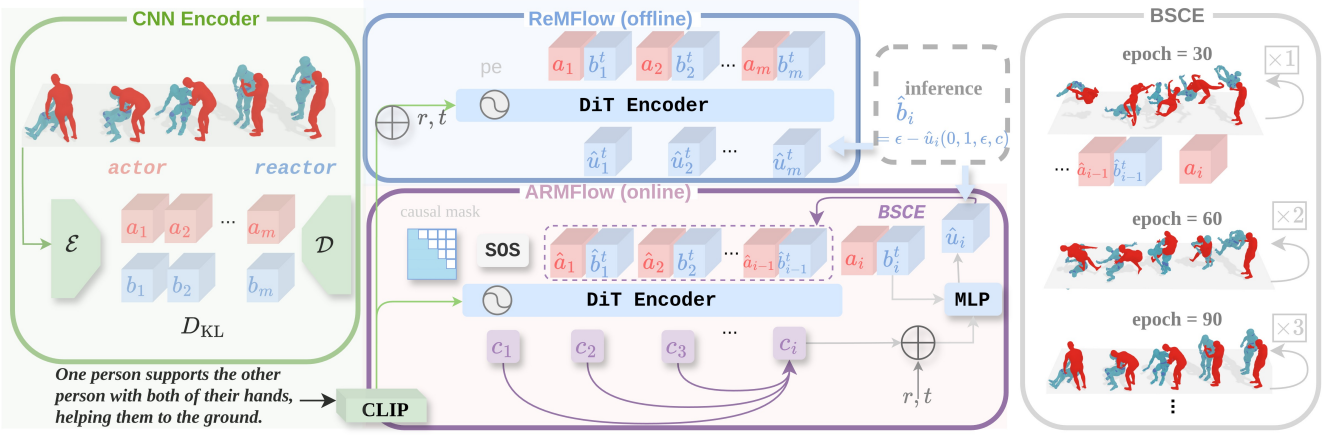

Figure 2. Overview of the proposed architecture for online and offline reaction generation. The framework consists of a CNN-based encoder to learn a compact latent space for the actor and the reactor. The ReMFlow is for offline generation based on the DiT architecture, and ARMFlow is the autoregressive online model consisting of a DiT context encoder and an MLP velocity predictor. A BSCE strategy is employed during online training progressively to reduce accumulated error in the autoregression.

performs online generation, it inherently lacks autoregression (AR) (i.e., generating the current action based on previously generated content), leading to temporal inconsistency. Ji et al. [18] addressed the autoregressive online generation issue by conditioning on a fixed window of previous context to guide the generation of the next timestep's action. However, a fixed-window design imposes significant scalability constraints: it cannot be adapted to finer temporal resolutions, and approaches such as [3] are unable to perform inference from scratch without any prior context. Moreover, when processing long sequences, the limited receptive field fails to capture distant dependencies, leading to cumulative error and a lack of global contextual understanding. R2R [1] aims to solve this by introducing full-history encoding, but its 50-step inference limits its efficiency.

Autoregressive generative models have been extensively explored, particularly for image generation. Some approaches tokenize continuous representations into discrete token spaces in an LLM-style manner [2, 39, 48, 49], while others adopt diffusion-based methods [14, 16, 21, 22] in an autoregressive fashion. Another line of research employs autoregressive flow-based architectures [10, 37]. Although autoregressive models are capable of capturing strong dependencies within data distributions, a core challenge lies in mitigating error accumulation that naturally occurs during sequential inference. T2M-GPT [49] addresses this issue by introducing random token corruption during training, forcing the model to correct corrupted past contexts. Subsequently, Tian et al. [39] proposed a hierarchical data representation to alleviate cumulative errors. However, these techniques are not directly applicable to conventional autoregressive models—especially diffusion-based architectures with fixed contextual windows [14, 16, 21]. To overcome this limitation, Li et al. [22] proposed a general framework combining an autoregressive context encoder with a lightweight MLP denoiser, leveraging MAE-style random masking [15] to reduce error propagation along the autoregressive direction. Nevertheless, such random masking schemes are less suitable for specially ordered or real-time generation tasks, where temporal consistency is essential. As a result, enhancing the robustness of real-time autoregressive models remains unsolved.

3. Preliminary

MeanFlow [8] is a recently proposed generative paradigm that aims to produce high-fidelity samples with only a single inference step. Unlike traditional Flow Matching (FM) [24, 27] approaches, which learn the instantaneous velocity field by estimating the continuous trajectory between data and noise distributions through numerous integration steps, MeanFlow instead models the mean velocity over the entire trajectory. Specifically, instead of computing the instantaneous flow $\mathbf{v}(x, t)$ at each time step, MeanFlow evaluates the average transport velocity between any two points r and t in the field. This simplification allows the model to generate samples by a single-step integration of the mean field from $r = 0$ (noise) to $t = 1$ (data), substantially reducing inference time while maintaining strong generative quality. Our method builds upon MeanFlow, enabling efficient reaction generation with single-step inference. Given an instantaneous velocity field $v(z_\tau, \tau)$, the average velocity between timesteps r and t is defined as:

$$u(z_t, r, t) = \frac{1}{t - r} \int_r^t v(z_\tau, \tau)\, d\tau, \tag{1}$$

which captures the cumulative dynamics over the interval $[r, t]$ and provides a smooth approximation of trajectory

evolution. Differentiating both sides using the Leibniz rule yields the training objective:

$$\mathcal{L}(\theta) = \mathbb{E}\big\|u_\theta(z_t, r, t) - \mathrm{sg}(u_{\text{tgt}})\big\|_2^2, \tag{2}$$

$$u_{\text{tgt}} = v(z_t, t) - (t-r)\big(v(z_t, t)\,\partial_z u_\theta + \partial_t u_\theta\big), \tag{3}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator to stabilize optimization. The Jacobian–vector products (JVPs) for $\partial_z u_\theta$ and $\partial_t u_\theta$ are computed via automatic differentiation, avoiding explicit Jacobian construction and reducing memory overhead while maintaining accurate gradient signals. Unlike conventional flow-based methods that uniformly sample timestep pairs, we adopt a *biased sampling strategy* where a proportion of pairs satisfy $r = t$. This allocation strengthens the learning of the instantaneous velocity $v(z_t, t)$, improves convergence, and enhances fidelity for high-frequency motion components, which are critical for reactive behaviors. A key advantage of MeanFlow is single-step inference, eliminating iterative refinement, which drastically reduces inference cost compared to multi-step diffusion or flow-based models, making it suitable for large-scale generation.

To improve controllability without additional inference overhead, Geng et al. [8] incorporate *classifier-free guidance* (CFG) during training rather than post-hoc blending by constructing a Ground-Truth Field. The modified objective becomes:

$$\mathcal{L}(\theta) = \mathbb{E}\big\|u_\theta^{\text{cfg}}(z_t, r, t \mid c) - \mathrm{sg}(u_{\text{tgt}})\big\|_2^2, \tag{4}$$

$$u_{\text{tgt}} = \tilde{v}_t - (t-r)\big(\tilde{v}_t\,\partial_z u_\theta^{\text{cfg}} + \partial_t u_\theta^{\text{cfg}}\big), \tag{5}$$

$$\tilde{v}_t = \omega v_t + (1-\omega) u_\theta^{\text{cfg}}(z_t, t, t), \tag{6}$$

where c denotes the conditioning signal, and $\omega > 1$ controls the blend between conditional and unconditional velocity fields. Training with CFG not only enhances the model's robustness and fidelity, but also reduces the extra calculation in inference for unconditional output.

4. Method

Given a text description $text$ and an actor motion sequence $x_a \in \mathbb{R}^{T,D}$, where T denotes the number of frames and D the pose dimension, our goal is to synthesize a realistic reactor motion $x_b \in \mathbb{R}^{T,D}$. The generation can occur in two modes: *sequential (online)*, where frames are predicted progressively, or *offline*, where the entire sequence is generated in a single pass. To achieve this, we first compress actor and reactor motions into a shared latent space using the CNN-VAE described in Section 4.1, which benefits both offline and online reaction generation. Building on this representation, we introduce **ReMFlow** (Section 4.2) for offline generation and **ARMFlow** (Section 4.3) for online generation, both leveraging the MeanFlow framework and a DiT-based backbone [31]. Finally, we propose the *Bootstrap Context Encoding* (BCSE), which is inherently suited for ARMFlow and mitigates error accumulation during autoregression.

4.1. CNN-VAE for Motion Compression

To improve generation efficiency and robustness, we adopt a 1D-CNN VAE following T2M-GPT [49] to compress actor–reactor motion sequences $\mathbf{x} = \{\mathbf{x}_a, \mathbf{x}_b\}$ into temporal latent representations $\{\mathbf{a}, \mathbf{b}\}$. This serves to (1) reduce motion dimensionality for efficient computation and (2) provide a structured latent space for autoregressive modeling.

Unlike canonicalized actor motion, reactor motion is highly dynamic and lacks standardization, making discrete quantization prone to accuracy loss. We therefore use continuous latent tokens with KL-divergence regularization instead of discrete codes. This embeds motion into a compact yet expressive latent space, accelerating convergence and improving reconstruction. The causal 1D-CNN preserves temporal dependencies, enabling efficient disentanglement of latent tokens in online generation and enhancing flexibility in sequential contexts. The VAE objective is:

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q(z|x)}[\log p(x|z)] - \mathrm{KL}(q(z|x)\|p(z)), \tag{7}$$

where $q(z|x)$ is the encoder, $p(x|z)$ the decoder, and $p(z)$ the Gaussian prior. To ensure high-fidelity reconstruction, we employ inverse kinematic (IK) loss for joint positions and velocity loss for motion smoothness.

4.2. ReMFlow for Offline Reaction Generation

We implement **ReMFlow** for offline generation using a **DiT**-based encoder with multimodal conditioning. Text prompts are encoded via a CLIP text encoder [28] and fused with timestep embeddings to form a global condition vector, which is injected into DiT through adaptive layer normalization (AdaLN) to modulate intermediate activations. Actor tokens a_i are concatenated with interpolated reactor tokens b_i^t, encoded from a CNN-based VAE to capture local appearance and dynamics of the reactor. These tokens are aligned along the last dimension, summed with positional encodings, and processed by DiT to predict average velocities $\mathbf{u} = \{u_i\}_{i=1}^m$ over tokenized spatiotemporal patches. To enable unconditional learning and support CFG, a proportion of text and actor tokens are replaced with null tokens $\varnothing$ during training, which regularizes the model and stabilizes optimization under varying conditioning strengths.

This design allows ReMFlow to jointly model actor dynamics and reactor responses under flexible conditioning while maintaining computational efficiency. Compared to iterative diffusion models, our approach achieves significant speedup and scalability due to single-step inference and average-velocity prediction, making it practical for large-scale offline reaction generation tasks without sacrificing fidelity or semantic alignment.

4.3. ARMFlow for Online Generation

ARMFlow overcomes the limitations of fixed-window architectures by enabling encoding from scratch and retaining all past information, thereby preserving global semantics. As illustrated in Fig. 2, the architecture follows the spirit of MAR [22], comprising a DiT-based context encoder and a lightweight MLP velocity predictor. The former encodes historical motion and text semantics and supplies them as partial conditions to the MLP predictor.

During training, we concatenate actor and reactor history tokens (a_i, b_i) along the last dimension and prepend a learnable start-of-sequence token $\langle \text{sos} \rangle$ to ensure that inference remains feasible at t=0. The concatenated sequence is fed into the DiT backbone, while text conditioning is injected through normalization layers (via infusion), and a causal mask is applied to enforce forward-only temporal dependencies for autoregressive learning. The encoded historical context c_i is then combined with the upsampled timesteps (r, t) and used as conditions for an AdaLN-modulated MLP. At the current step, the interpolated reactor sample $b_i^{r,t}$ from the velocity field is passed to the MLP velocity predictor, which outputs the average velocity $\hat{u}_{r,t}$. Unlike offline generation, the conditioning input here includes not only the current actor but also the accumulated history; consequently, when performing classifier-free guidance (CFG), we augment the null token with a *null history* to match the online conditioning interface.

At inference time, the start token $\langle \text{sos} \rangle$ is first encoded as the initial history. A Gaussian noise token ϵ_1 is then paired with the current actor token a_1 and fed into the MLP with (r=0, t=1) to predict the average velocity $\hat{u}_1$. The current reactor token is obtained by $b_1 = \epsilon_1 - \hat{u}_1$. The predicted reactor token is cached together with the corresponding actor token to update the history, and the process is iterated autoregressively until actor tokens end.

4.4. Bootstrap Contextual Encoding

Drift over long sequences is an inherent limitation of autoregressive models. HumanX [18] addresses this through a history-rollout training strategy, where the ground-truth (GT) history condition is gradually replaced by the model's generated history. However, this approach has three main drawbacks. First, to stabilize training, HumanX gradually subtracts the GT history from the generated ones, slowing the convergence. Second, as the model converges, the generated history becomes increasingly similar to the GT, leading to insufficient self-augmentation. Third, HumanX only replaces the reactor history while keeping the actor unchanged, which lead to overfitting to the actor's motion.

To address these limitations, we propose Bootstrap Contextual Encoding (BSCE). As detailed in Algorithm 1, BSCE replaces both actor and reactor histories with generated samples from the very beginning of training. As the model progressively aligns its outputs with the GT, the number of autoregressive iterations is gradually increased on schedule. This amplified accumulated error introduces additional noise, which in turn enhances the model's robustness and generalization. Moreover, since HumanX relies on a multi-step diffusion model, its rollout is computationally expensive during the training phase. In contrast, our MeanFlow-based generator performs single-step inference, providing substantial efficiency gains and further highlighting BSCE's natural compatibility with MeanFlow.

Algorithm 1 Bootstrap Contextual Encoding (BSCE)

procedure BSCE($G_\theta, \mathcal{Y}, \mathcal{X}, \mathcal{C}, \mathcal{S}_t, K_{\max}, I_{\max}$)
  $K \leftarrow 1$
  for iteration $= 1$ to $I_{\max}$ **do**
    Sample (x_a, x_b, c) from $\mathcal{Y}, \mathcal{X}, \mathcal{C}$
    Initialize context buffer $\mathcal{Z} \leftarrow \{\langle \text{sos} \rangle\}$
    $(\{a_i\}, \{b_i^{\text{gt}}\}) \leftarrow$ TOKENIZE(x_a, x_b)
    for $i = 1$ to K **do**
      $c_i \leftarrow$ ENCODECONTEXT$(\mathcal{Z}, c)$
      Sample $(r, t) \sim \mathcal{S}_t$, $\epsilon_i \sim \mathcal{N}(0, I)$
      $\hat{u}_{r,t} \leftarrow G_\theta(\epsilon_i, a_i \text{ or } b_i, r, t, c_i)$
      a_i or $b_i \leftarrow \epsilon_i - \hat{u}_{r,t}$
      Append (a_i, b_i) to $\mathcal{Z}$
      $\mathcal{L}_i \leftarrow$ LOSS$(\hat{u}_{r,t})$
    end for
    Update $\theta \leftarrow \theta - \eta \nabla_\theta \left(\frac{1}{K} \sum_i \mathcal{L}_i \right)$
    $K \leftarrow$ SCHEDULED$(iteration, K_{\max})$
  end for
end procedure

5. Experiments

Datasets. We evaluate our approach on two widely adopted benchmarks for text-conditioned human interaction synthesis: *InterHuman* [23] and *InterX* [45]. *InterHuman* comprises 7,779 interaction sequences, while *InterX* provides 11,388 sequences, each annotated with 3 textual descriptions. Compared with the other noisy and small datasets like NTU-120 and CHI3D[7, 25], these two datasets are of better motion qualities and have solid and public metrics for evaluations for fair comparisons.

The *InterHuman* dataset is built upon the AMASS [29] skeleton, which includes 22 joints with the root joint. Each joint is represented following the HumanML3D [11] convention as $\{\mathbf{p}_g, \mathbf{v}_g, \mathbf{r}_{6d}\}$, where $\mathbf{p}_g \in \mathbb{R}^3$ denotes global position, $\mathbf{v}_g \in \mathbb{R}^3$ global velocity, and $\mathbf{r}_{6d} \in \mathbb{R}^6$ the local 6D rotation. This results in a motion tensor $\mathbf{m}_p \in \mathbb{R}^{N \times 22 \times 12}$, accompanied by a binary foot-contact indicator $fc \in \mathbb{R}^2$.

In contrast, *InterX* adopts the SMPL-X [30] skeleton, which has 55 articulated joints covering the body, hands, and facial regions, along with the root orientation. Each joint rotation and root orientation is encoded as $\mathbf{r}_{6d}$, while

Table 2. Comparison of online methods on InterHuman and InterX datasets.

Dataset	Model	FID ↓	R-Prec@1 ↑	R-Prec@2 ↑	R-Prec@3 ↑	MM Dist ↓	Diversity →	MModality ↑
InterHuman	Ground Truth	$0.273^{\pm.007}$	$0.452^{\pm.008}$	$0.610^{\pm.009}$	$0.701^{\pm.008}$	$3.755^{\pm.008}$	$7.948^{\pm.064}$	-
	InterFormer [5]	$4.871^{\pm.049}$	$0.302^{\pm.004}$	$0.457^{\pm.004}$	$0.542^{\pm.005}$	$3.845^{\pm.001}$	$7.482^{\pm.045}$	$0.254^{\pm.029}$
	CAMDM [3]	$4.000^{\pm.046}$	$0.335^{\pm.005}$	$0.492^{\pm.005}$	$0.587^{\pm.005}$	$3.828^{\pm.001}$	$7.547^{\pm.025}$	**1.581**$^{\pm.026}$
	ReGenNet [46]	$4.176^{\pm.085}$	$0.355^{\pm.005}$	$0.508^{\pm.005}$	$0.600^{\pm.004}$	$3.817^{\pm.001}$	$7.480^{\pm.033}$	$0.442^{\pm.012}$
	R2R [1]	$2.795^{\pm.062}$	$0.431^{\pm.005}$	$0.591^{\pm.004}$	$0.674^{\pm.004}$	$3.793^{\pm.002}$	$7.693^{\pm.028}$	$0.517^{\pm.013}$
	ARMFlow (**Ours**)	**2.178**$^{\pm.054}$	**0.441**$^{\pm.005}$	**0.605**$^{\pm.005}$	**0.699**$^{\pm.005}$	**3.783**$^{\pm.002}$	**7.745**$^{\pm.024}$	$0.369^{\pm.008}$
InterX	Ground Truth	$0.002^{\pm.000}$	$0.435^{\pm.005}$	$0.628^{\pm.004}$	$0.736^{\pm.004}$	$3.574^{\pm.013}$	$8.947^{\pm.078}$	-
	InterFormer [5]	$0.304^{\pm.009}$	$0.301^{\pm.003}$	$0.469^{\pm.003}$	$0.571^{\pm.002}$	$4.604^{\pm.009}$	$8.579^{\pm.061}$	$0.289^{\pm.009}$
	CAMDM [3]	$0.429^{\pm.011}$	$0.312^{\pm.004}$	$0.480^{\pm.003}$	$0.587^{\pm.003}$	$4.468^{\pm.020}$	$8.467^{\pm.072}$	**1.460**$^{\pm.027}$
	ReGenNet [46]	$0.071^{\pm.003}$	$0.402^{\pm.005}$	$0.584^{\pm.004}$	$0.690^{\pm.004}$	$3.843^{\pm.011}$	$9.011^{\pm.053}$	$0.738^{\pm.021}$
	R2R [1]	$0.063^{\pm.003}$	$0.412^{\pm.005}$	$0.598^{\pm.005}$	$0.704^{\pm.004}$	$3.745^{\pm.011}$	$8.873^{\pm.055}$	$1.074^{\pm.019}$
	ARMFlow (**Ours**)	**0.042**$^{\pm.003}$	**0.420**$^{\pm.004}$	**0.606**$^{\pm.004}$	**0.711**$^{\pm.004}$	**3.728**$^{\pm.012}$	**8.939**$^{\pm.071}$	$1.203^{\pm.029}$

the root translation and rotation are represented by $\mathbf{t}_r$ and $\mathbf{r}_r$, respectively, forming $\mathbf{m}_p \in \mathbb{R}^{N\times 56\times 6}$.

Evaluation Metrics. For reaction generation, we mainly focus on fidelity and semantic correspondence. To comprehensively evaluate the performance of our model, we adopt a suite of feature-space metrics by Liang et al. [23]. To assess the realism and fidelity of generated interactions, we compute the Fréchet Inception Distance (FID) and between the feature distributions of generated and ground-truth motions and their Diversity. To evaluate the semantic alignment between text prompts and generated motions, we use R-Precision and Multimodal Distance (MMDist), which measure the consistency between text input and generated motions. To assess generative quality beyond accuracy, we report Multimodality, quantifying the model's ability to produce multiple plausible motions for the same text prompts.

Baselines. We divide our evaluation into two parts. For offline generation, we compare our approach with several state-of-the-art methods, including In2IN, which first synthesizes actor–reactor motions with an MDM prior and subsequently refines them; ReGenNet, a diffusion model with a transformer-decoder backbone; InterMask, a discrete autoregressive model based on masking; and an extended variant of MLD that performs diffusion in a latent space. For online generation, we benchmark against InterFormer, a conventional transformer-based autoregressive architecture; the online configuration of ReGenNet; CAMDM, which adopts a transformer encoder and conditions on a fixed-length context window; and R2R, an autoregressive diffusion diagram. Notably, these online baselines are not explicitly designed for real-time synthesis. Although ReGenNet can apply causal masking, its decoder architecture only conditions on the actor's previous frame and cannot exploit the reactor's past states, which often leads to temporal discontinuities. CAMDM, on the other hand, relies on a short fixed window of 10 frames, limiting long-range contextual modeling and causing pronounced drift in extended sequences, particularly on InterHuman. Furthermore, these designs do not support generation from scratch, rendering inference at t=0 infeasible. In contrast, our method operates at a finer 4-frame granularity for real-time synthesis, scales more favorably than CAMDM, and does not depend on any initial motion. To ensure a fair comparison, we provide CAMDM with additional support by supplying an initial motion sequence, which compensates for its inability to generate from scratch, while keeping its 10-frame context window as specified by the authors.

Implementation Details. For the CNN encoder, we use the same architecture for both datasets except for the input dimension. Following T2M-GPT, the encoder has a hidden size of 256, two residual convolutional downsampling blocks, and three hidden layers per block. The main model adopts a standard DiT backbone for both ARMFlow (online) and ReMFlow (offline), with 512 hidden dimensions, 7 transformer layers, and 8 attention heads with skip connections. The MLP-based velocity predictor in ARMFlow is a lightweight 5-layer network, with identical settings across datasets. For MeanFlow, timesteps are sampled from a logit-normal distribution, with the probability of instantaneous velocity sampling ($r = t$) set to 0.25. In ReMFlow, classifier-free guidance (CFG) is set to $\omega = 1.8$ for InterHuman and $\omega = 2.0$ for InterX. For ARMFlow, we use $\omega = 1.8$ on InterHuman and $\omega = 1.2$ on InterX. All models are trained with a batch size of 64. The CNN-VAE is trained for 2000 epochs on InterHuman and 800 on InterX. ReMFlow is trained for 800 and 500 epochs, while ARMFlow converges faster with 500 and 300 epochs, respectively. All experiments use the AdamW optimizer with a learning rate of 1×10^{-4} and are conducted on NVIDIA H100 GPUs in an HPC environment.

5.1. Quantative Results

Online Generation. We first consider the more challenging online generation setting. Tab. 2 compares our ARMFlow with existing online models, ReGenNet and

Table 3. Comparison of offline methods on InterHuman and InterX datasets.

Dataset	Model	FID ↓	R-Prec@1 ↑	R-Prec@2 ↑	R-Prec@3 ↑	MM Dist ↓	Diversity →	MModality ↑
InterHuman	Ground Truth	$0.273^{\pm.007}$	$0.452^{\pm.008}$	$0.610^{\pm.009}$	$0.701^{\pm.008}$	$3.755^{\pm.008}$	$7.948^{\pm.064}$	-
	InterGen [23]	$9.183^{\pm.174}$	$0.325^{\pm.004}$	$0.467^{\pm.004}$	$0.546^{\pm.005}$	$3.859^{\pm.001}$	$7.305^{\pm.047}$	$1.270^{\pm.023}$
	in2IN [35]	$7.913^{\pm.251}$	$0.362^{\pm.004}$	$0.504^{\pm.008}$	$0.589^{\pm.007}$	$3.832^{\pm.002}$	$7.709^{\pm.040}$	$1.165^{\pm.034}$
	MLD* [4]	$3.588^{\pm.076}$	$0.405^{\pm.005}$	$0.561^{\pm.007}$	$0.649^{\pm.006}$	$3.798^{\pm.002}$	$7.663^{\pm.040}$	$1.124^{\pm.048}$
	ReGenNet* [46]	$2.930^{\pm.052}$	$0.362^{\pm.005}$	$0.513^{\pm.005}$	$0.605^{\pm.004}$	$3.815^{\pm.001}$	$7.582^{\pm.064}$	$\mathbf{1.737}^{\pm.020}$
	InterMask [17]	$3.453^{\pm.061}$	$0.451^{\pm.007}$	$0.610^{\pm.006}$	$0.701^{\pm.005}$	$3.782^{\pm.002}$	$7.710^{\pm.046}$	$1.361^{\pm.032}$
	ReMFlow (**Ours**)	$\mathbf{2.433}^{\pm.042}$	$\mathbf{0.452}^{\pm.006}$	$\mathbf{0.618}^{\pm.005}$	$\mathbf{0.708}^{\pm.005}$	$\mathbf{3.778}^{\pm.001}$	$\mathbf{7.714}^{\pm.029}$	$0.685^{\pm.018}$
InterX	Ground Truth	$0.002^{\pm.000}$	$0.435^{\pm.005}$	$0.628^{\pm.004}$	$0.736^{\pm.004}$	$3.574^{\pm.013}$	$8.947^{\pm.078}$	-
	InterGen [23]	$0.238^{\pm.038}$	$0.352^{\pm.004}$	$0.542^{\pm.005}$	$0.643^{\pm.004}$	$4.212^{\pm.029}$	$8.773^{\pm.067}$	$1.552^{\pm.029}$
	MLD* [4]	$0.148^{\pm.020}$	$0.414^{\pm.004}$	$0.607^{\pm.006}$	$0.712^{\pm.004}$	$3.655^{\pm.020}$	$8.893^{\pm.053}$	$1.875^{\pm.088}$
	ReGenNet [46]	$0.093^{\pm.022}$	$0.407^{\pm.004}$	$0.589^{\pm.004}$	$0.705^{\pm.003}$	$3.762^{\pm.015}$	$8.841^{\pm.072}$	$\mathbf{1.937}^{\pm.069}$
	InterMask [17]	$0.399^{\pm.013}$	$0.429^{\pm.005}$	$0.622^{\pm.005}$	$0.731^{\pm.005}$	$3.584^{\pm.017}$	$8.911^{\pm.057}$	$0.859^{\pm.033}$
	ReMFlow (**Ours**)	$\mathbf{0.058}^{\pm.005}$	$\mathbf{0.440}^{\pm.004}$	$\mathbf{0.636}^{\pm.004}$	$\mathbf{0.743}^{\pm.003}$	$\mathbf{3.570}^{\pm.013}$	$\mathbf{8.948}^{\pm.068}$	$1.607^{\pm.039}$

Figure 3. Qualitative comparison with ReGenNet on InterHuman dataset. The problematic interactions are marked with red dashed lines, including penetrations and semantic misalignment, and the correct ones are marked with green.

CAMDM. On the long-sequence InterHuman dataset, both baselines fail to achieve satisfactory results. CAMDM is mainly limited by its short context window, which prevents encoding the full history and leads to severe temporal drift. Although ReGenNet leverages the full actor history, it initializes the reactor input from Gaussian noise at each inference step, preventing adaptive refinement of the reactor representation based on its own history and resulting in temporally inconsistent actions. In contrast, ARMFlow significantly outperforms existing methods in both FID and semantic consistency. Compared with R2R (the second-best in FID), ARMFlow achieves a 28% relative improvement. Moreover, on the shorter-sequence InterX dataset, our method also achieves state-of-the-art performance. ReGenNet performs relatively well on this dataset, likely because its short-term attention mechanism is more effective for shorter sequences. Notably, online generation does not necessarily degrade performance—in our experiments, ARMFlow even achieves a lower (better) FID in the online setting than in the offline one, while the offline version, benefiting from full actor-conditioned information, yields slightly higher semantic consistency.

Offline Generation. For the offline generation task, we evaluate the generality of ReMFlow and compare it with current SOTA methods in Tab. 3. InterMask, based on discrete random masking, performs well on semantic metrics such as R-precision and MM-Distance but yields a significantly worse FID. This mainly results from its shared

VQ-VAE encoder, which over-compresses motion features and degrades reconstruction quality—one reason we avoid discrete representations. ReGenNet operates directly on raw data without compression and thus avoids this loss; however, on the long-sequence InterHuman dataset its self-attention struggles to capture fine-grained semantics, leading to weaker semantic consistency despite a strong FID. MLD lies between these extremes, reflecting a common trend in motion generation: stronger compression improves semantic alignment but often reduces reconstruction fidelity. InterMask achieves high semantic consistency at the cost of efficiency, requiring at least 20 inference steps (0.77 s). In contrast, ReMFlow attains state-of-the-art performance with a single forward pass, greatly improving inference speed while maintaining generation quality. Similar trends appear on the InterX dataset, where shorter sequences allow ReGenNet to achieve relatively strong semantic alignment.

5.2. Qualitative Comparison

We present online generation results qualitatively on the InterHuman dataset, which has longer sequences, and the competition is more challenging. We compare our method with ReGenNet, and visualize the results in Fig. 3, which indicate that our approach not only achieves better interperson contact quality but also demonstrates stronger semantic alignment with the text prompts.

5.3. Ablation Study

We validate the effectiveness of our proposed designs and individual modules. First, for the generative approaches, we kept the same network architecture while replacing the objective function with Diffusion Models and Rectified Flow, respectively, and performed task-specific parameter tuning for them including CFG strength. Experimental results in Tab. 4 and Tab. 5 show that MeanFlow consistently achieves the best FID scores and semantic alignment across both online and offline generation tasks, as well as on both datasets. Moreover, its single-step generation greatly accelerates the training process when integrated with BSCE, demonstrating the broad adaptability of MeanFlow to various generation tasks. Furthermore, we compare BSCE with two alternative strategies: using ground-truth contextual encoding (GTE) and progressive rollout. BSCE significantly outperforms both approaches, confirming the effectiveness and superiority of our proposed method.

6. Conclusion

We presented ARMFlow, the first MeanFlow-based framework for real-time 3D human reaction generation, achieving single-step inference while maintaining high-fidelity and context-awareness. Our causal context encoder and Bootstrap Context Encoding (BSCE) effectively address long-term dependency modeling and autoregressive error accumulation, enabling robust and efficient online generation. We also introduce ReMFlow as a versatile offline baseline, demonstrating state-of-the-art performance with unprecedented inference speed. Overall, our work establishes a scalable and practical framework for both online and offline human motion generation, paving the way for real-world applications in Human-Robot Interaction, AR, and VR.

Despite these strengths, our method has certain limitations. First, from an engineering perspective, the current implementation does not provide elastic delay handling for the autoregressive small window, which may lead to minor asynchronous behaviors when multiple agents interact. Second, due to the nature of MeanFlow, it does not support post-hoc classifier guidance like diffusion-based models, preventing the use of optimization-based corrections to further refine generated motions. Addressing these limitations represents promising directions for future work.

Table 4. Ablation study for online generation on methods.

	Model	R-Prec@3 ↑	FID ↓	MM Dist ↓	Diversity →
InterHuman	DDIM 10	$0.689^{\pm.004}$	$3.528^{\pm.049}$	$3.803^{\pm.002}$	$7.691^{\pm.037}$
	DDIM 50	$\underline{0.697}^{\pm.005}$	$3.449^{\pm.062}$	$3.794^{\pm.002}$	$7.702^{\pm.028}$
	DDPM	$0.686^{\pm.004}$	$3.757^{\pm.068}$	$3.806^{\pm.002}$	$\mathbf{7.803}^{\pm.035}$
	Rectified Flow 10	$0.692^{\pm.004}$	$\underline{2.449}^{\pm.062}$	$3.796^{\pm.001}$	$7.702^{\pm.028}$
	ARMFlow **(Ours)**	$\mathbf{0.699}^{\pm.005}$	$\mathbf{2.178}^{\pm.054}$	$\mathbf{3.783}^{\pm.002}$	$\underline{7.745}^{\pm.024}$
InterX	DDIM 10	$0.692^{\pm.004}$	$0.093^{\pm.005}$	$3.802^{\pm.014}$	$8.870^{\pm.066}$
	DDIM 50	$\underline{0.705}^{\pm.004}$	$0.064^{\pm.004}$	$\underline{3.733}^{\pm.013}$	$8.895^{\pm.062}$
	DDPM	$0.695^{\pm.004}$	$0.106^{\pm.006}$	$3.792^{\pm.015}$	$\underline{8.920}^{\pm.088}$
	Rectified Flow 10	$0.698^{\pm.005}$	$\underline{0.059}^{\pm.004}$	$3.784^{\pm.011}$	$8.913^{\pm.072}$
	ARMFlow **(Ours)**	$\mathbf{0.711}^{\pm.004}$	$\mathbf{0.042}^{\pm.003}$	$\mathbf{3.728}^{\pm.012}$	$\mathbf{8.939}^{\pm.071}$

Table 5. Ablation study for offline generation on methods.

	Model	R-Prec@3 ↑	FID ↓	MM Dist ↓	Diversity →
InterHuman	DDIM 10	$0.677^{\pm.006}$	$3.931^{\pm.050}$	$3.800^{\pm.002}$	$7.622^{\pm.025}$
	DDIM 50	$\underline{0.694}^{\pm.004}$	$2.918^{\pm.068}$	$\underline{3.789}^{\pm.003}$	$7.656^{\pm.040}$
	DDPM	$0.681^{\pm.005}$	$3.595^{\pm.051}$	$3.797^{\pm.002}$	$7.663^{\pm.035}$
	Rectified Flow 10	$0.678^{\pm.005}$	$\underline{2.906}^{\pm.044}$	$3.801^{\pm.002}$	$\mathbf{7.775}^{\pm.045}$
	ReMFlow**(Ours)**	$\mathbf{0.708}^{\pm.005}$	$\mathbf{2.433}^{\pm.042}$	$\mathbf{3.778}^{\pm.001}$	$\underline{7.714}^{\pm.029}$
InterX	DDIM 10	$0.699^{\pm.005}$	$0.127^{\pm.036}$	$3.805^{\pm.012}$	$8.804^{\pm.059}$
	DDIM 50	$0.700^{\pm.004}$	$\underline{0.095}^{\pm.076}$	$3.749^{\pm.013}$	$8.842^{\pm.077}$
	DDPM	$0.671^{\pm.005}$	$3.595^{\pm.051}$	$3.893^{\pm.012}$	$8.857^{\pm.063}$
	Rectified Flow 10	$\underline{0.737}^{\pm.007}$	$0.103^{\pm.015}$	$\underline{3.645}^{\pm.012}$	$\underline{8.941}^{\pm.087}$
	ReMFlow**(Ours)**	$\mathbf{0.743}^{\pm.003}$	$\mathbf{0.058}^{\pm.005}$	$\mathbf{3.570}^{\pm.013}$	$\mathbf{8.948}^{\pm.068}$

Table 6. Ablation on the training strategies for online autoregressive diffusion. GTE stands for ground-truth encoding, Rollout is the strategy used in HumanX[18], and BSCE is our strategy.

	Model	R-Prec@3 ↑	FID ↓	MM Dist ↓	Diversity →
InterHuman	Ground Truth	$0.701^{\pm.008}$	$0.273^{\pm.007}$	$3.755^{\pm.008}$	$7.948^{\pm.064}$
	ARMFlow-GTE	$0.602^{\pm.004}$	$5.136^{\pm.040}$	$3.813^{\pm.002}$	$7.728^{\pm.033}$
	ARMFlow-Rollout	$0.675^{\pm.005}$	$4.161^{\pm.055}$	$3.798^{\pm.002}$	$\mathbf{7.802}^{\pm.038}$
	ARMFlow**(Ours)**	$\mathbf{0.699}^{\pm.005}$	$\mathbf{2.178}^{\pm.054}$	$\mathbf{3.783}^{\pm.002}$	$7.745^{\pm.024}$
InterX	Ground Truth	$0.736^{\pm.004}$	$0.002^{\pm.000}$	$3.574^{\pm.013}$	$8.947^{\pm.078}$
	ARMFlow-GTE	$0.630^{\pm.003}$	$0.548^{\pm.007}$	$4.667^{\pm.020}$	$8.435^{\pm.068}$
	ARMFlow-Rollout	$0.670^{\pm.004}$	$0.192^{\pm.006}$	$4.321^{\pm.013}$	$8.743^{\pm.082}$
	ARMFlow**(Ours)**	$\mathbf{0.711}^{\pm.004}$	$\mathbf{0.042}^{\pm.003}$	$\mathbf{3.728}^{\pm.012}$	$\mathbf{8.939}^{\pm.071}$

Acknowledgements

This research was supported by the Australian Research Council (ARC) Discovery Project DP240101926. We gratefully acknowledge this support.

References

[1] Zhi Cen, Huaijin Pi, Sida Peng, Qing Shuai, Yujun Shen, Hujun Bao, Xiaowei Zhou, and Ruizhen Hu. Ready-to-react: Online reaction policy for two-character interaction generation. In *ICLR*, 2025. 3, 6

[2] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[3] Rui Chen, Mingyi Shi, Shaoli Huang, Ping Tan, Taku Komura, and Xuelin Chen. Taming diffusion probabilistic models for character control. In *ACM SIGGRAPH 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 2, 3, 6

[4] Xin Chen, Wen Jiang, Biao Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 7

[5] Baptiste Chopin, Hao Tang, Naima Otberdout, Mohamed Daoudi, and Nicu Sebe. Interaction transformer for human reaction generation. *IEEE Transactions on Multimedia (TMM)*, pages 1–13, 2023. 2, 6

[6] Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19888–19901, 2024. 1, 2

[7] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[8] Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025. 2, 3, 4

[9] Zichen Geng, Zeeshan Hayder, Wei Liu, and Ajmal Saeed Mian. Auto-regressive diffusion for generating 3d human-object interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 3131–3139, 2025. 1, 2

[10] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. 3

[11] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, 2020. 1, 2, 5, 3

[12] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4, 5

[13] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1900–1910, 2024. 1, 2

[14] Bo Han, Hao Peng, Minjing Dong, Yi Ren, Yixuan Shen, and Chang Xu. Amd: Autoregressive motion diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 2022–2030, 2024. 3

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[16] Emiel Hoogeboom, Alexey A Gritsenko, Jasmijn Bastings, Ben Poole, Rianne van den Berg, and Tim Salimans. Autoregressive diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022. 3

[17] Muhammad Gohar Javed, Chuan Guo, Li Cheng, and Xingyu Li. Intermask: 3d human interaction generation via collaborative masked modeling. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. 2, 7

[18] Kaiyang Ji, Ye Shi, Zichen Jin, Kangyi Chen, Lan Xu, Yuexin Ma, Jingyi Yu, and Jingya Wang. Towards immersive human-x interaction: A real-time framework for physically plausible motion synthesis. *arXiv preprint arXiv:2508.02106*, 2025. 2, 3, 5, 8

[19] Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng. A large-scale rgb-d database for arbitrary-view human action recognition. New York, NY, USA, 2018. Association for Computing Machinery. 3

[20] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. *arXiv preprint arXiv:2312.03913*, 2023. 1, 2

[21] Tianyu Li, Calvin Qiao, Guanqiao Ren, KangKang Yin, and Sehoon Ha. Aamdm: Accelerated auto-regressive motion diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1813–1823, 2024. 3

[22] Tao Li, Yu Tian, Hang Li, Mingyuan Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024. To appear. 3, 5

[23] 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 (IJCV)*, pages 1–21, 2024. 2, 5, 6, 7, 3

[24] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative mod-

eling. In *11th International Conference on Learning Representations (ICLR)*, 2023. 3

[25] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 42(10):2684–2701, 2020. 5

[26] Jian Liu, Wei Sun, Hui Yang, Pengchao Deng, Chongpei Liu, Nicu Sebe, Hossein Rahmani, and Ajmal Mian. Diff9d: Diffusion-based domain-generalized category-level 9-dof object pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5520–5537, 2025. 2

[27] Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 3

[28] Zhenzhong Luo, Yuwen Xiong, Xuancheng Sun, Yang Long, Ting Yao, and Tao Mei. Learning transferable visual models from natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[29] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, pages 5442–5451, 2019. 5

[30] 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. 5

[31] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182. IEEE Computer Society, 2023. 4

[32] Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. *arXiv preprint arXiv:2312.06553*, 2023. 1, 2

[33] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[34] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2

[35] 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 (CVPR) Workshops*, pages 1941–1951, 2024. 2, 7

[36] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[37] NextStep Team, Chunrui Han, Guopeng Li, Jingwei Wu, Quan Sun, Yan Cai, Yuang Peng, Zheng Ge, Deyu Zhou, Haomiao Tang, Hongyu Zhou, Kenkun Liu, Ailin Huang, Bin Wang, Changxin Miao, Deshan Sun, En Yu, Fukun Yin, Gang Yu, Hao Nie, Haoran Lv, Hanpeng Hu, Jia Wang, Jian Zhou, Jianjian Sun, Kaijun Tan, Kang An, Kangheng Lin, Liang Zhao, Mei Chen, Peng Xing, Rui Wang, Shiyu Liu, Shutao Xia, Tianhao You, Wei Ji, Xianfang Zeng, Xin Han, Xuelin Zhang, Yana Wei, Yanming Xu, Yimin Jiang, Yingming Wang, Yu Zhou, Yucheng Han, Ziyang Meng, Binxing Jiao, Daxin Jiang, Xiangyu Zhang, and Yibo Zhu. Nextstep-1: Toward autoregressive image generation with continuous tokens at scale. *arXiv preprint arXiv:2508.10711*, 2025. 3

[38] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 1, 2

[39] Keyu Tian, Yicheng Jiang, Ziyang Yuan, Bo Peng, and Lijuan Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *Proceedings of the 37th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 84839–84865, 2024. NeurIPS 2024 Best Paper. 3

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems (NeurIPS)*, 30, 2017. 2

[41] Yin Wang, Zhiying Leng, Frederick WB Li, Shun-Cheng Wu, and Xiaohui Liang. Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22035–22044, 2023. 1, 2

[42] Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. Physhoi: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393*, 2023. 2

[43] Yabiao Wang, Shuo Wang, Jiangning Zhang, Ke Fan, Jiafu Wu, Zhucun Xue, and Yong Liu. Timotion: Temporal and interactive framework for efficient human-human motion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 7169–7178, 2025. 2

[44] YB Wang, Shuo Wang, JN Zhang, JF Wu, QD He, CC Fu, CJ Wang, and Yong Liu. Marrs: Masked autoregressive unit-based reaction synthesis. *arXiv preprint arXiv:2505.11334*, 2025. 2

[45] 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 *CVPR*, pages 22260–22271, 2024. 1, 5, 3

[46] Liang Xu, Yizhou Zhou, Yichao Yan, Xin Jin, Wenhan Zhu, Fengyun Rao, Xiaokang Yang, and Wenjun Zeng. Regennet: Towards human action-reaction synthesis. In *CVPR*, pages 1759–1769, 2024. 2, 6, 7, 3

[47] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. InterDiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023. 1, 2

[48] Lijun Yu, José Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alex Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David Ross, and Lu Jiang. Language model beats diffusion – tokenizer is key to visual generation. In *ICLR*, 2024. 3

[49] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 3, 4

[50] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2024.

[51] Chongyang Zhong, Lei Hu, Zihao Zhang, and Shihong Xia. Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 509–519, 2023. 1, 2

ARMFlow: AutoRegressive MeanFlow for Online 3D Human Reaction Generation

Supplementary Material

Table of Contents

7. Supplementary Introduction

This supplementary document provides additional experiments, visualizations, and detailed descriptions of the accompanying code and log files. Beyond the analyses included in this PDF, we also provide supplementary MP4 videos and the complete source code contained in the `armflow/` directory, including training and inference scripts as well as a comprehensive README.

- **supplementary.pdf** — Contains this pdf file:
- **supplementary.mp4** — Contains all supplementary figures:
- **armflow/** — Contains scripts for reproducing figures and tables:
  - `README.md`: Readme file for how to run our code.
  - `others/` : All other necessary files and folders.
- `logs/` — training logs for the trained methods.

Due to storage limitations on the submission platform, we are unable to include the full pre-trained model checkpoints. However, all training configurations, scripts, and instructions are provided to allow readers to reproduce the results from scratch. Please refer to the README file for environment setup, dataset preparation instructions, and training commands.

The supplementary materials include:

- Additional quantitative and qualitative experiments extending those presented in the main paper.
- Visualization results for BSCE, model ablations, comparative evaluations, and velocity fields.
- Complete source code for ARMFlow, including model definitions, training pipelines, dataset loaders, and evaluation tools.
- Log files for key experiments to facilitate reproducibility and verification.
- Demonstration videos showcasing generated motions on the InterHuman and InterX datasets.

Readers can follow the provided instructions to install dependencies, prepare the datasets, and train models using the released implementation.

8. Discussion on Efficiency

With recent advances in VR devices and human pose estimation, real-time interaction with virtual environments has become increasingly feasible, making low latency a critical requirement. As summarized in Table 7, our method achieves the lowest latency, ranging from only 7 to 11 ms per time step, with just a single inference step. The slight variation in latency is due to the increasing context size. In practical grounding applications, we can truncate the context using a maximum window length while still retaining essential information from the historical context.

It is worth noting that in our experiments, we set the response window length to 4 frames for 30 fps scenarios, corresponding to approximately 13 ms—an appropriate window size for VR processing or human pose estimation tasks. Our method's latency falls well within this range, ensuring real-time performance. In contrast, other approaches, such as CAMDM and ReGenNet, exhibit significantly higher latency, far exceeding the practical requirements for such applications.

Table 7. Comparison of model latency and number of inference steps. The $2\times$ means classifier free guidance

Model	Latency (ms)	Number of Inference Steps
CAMDM	45	2×8
ReGenNet	35-78	2×5
Ours	7-11	1

Figure 4. FID over BSCE on InterHuman Dataset

9. Discussion on BSCE

In this section, we further elaborate on the behavior and advantages of BSCE. As described previously, we introduce a fixed repeat epoch number e_r during training; for InterHuman, it is set to 50, and for the InterX datase,t it is set to 30. Figure 4 presents the FID curves throughout training, where BSCE is directly compared against the Rollout strategy [18] as well as the Ground Truth Encoding (GTE). For completeness, the full evaluation logs are included in the supplementary file under `logs`.

In the following, we provide several observations that summarize its empirical effects.

Faster convergence. The results in Fig. 4 indicate that BSCE leads to noticeably faster convergence. Although the FID is relatively high during the very early iteration, which is likely due to its more aggressive supervision pattern, it causes quicker convergence. Once the first stage completes, BSCE reaches a similar FID level to that of GTE. This suggests that the learning dynamics under BSCE allow the model to adjust more rapidly, even if the initial fluctuations appear larger.

Better fidelity. After the initial warm-up period, the trend becomes clearer. The model trained with BSCE consistently attains a lower FID compared to the other strategies, implying an improvement in generation fidelity. This behavior remains stable across epochs beyond the first, demonstrating that BSCE not only accelerates convergence but also ultimately guides the model toward a solution that better aligns with the target data distribution. It is worth noting that this improvement persists even though no additional architectural modifications are introduced, highlighting that the benefit primarily comes from the training paradigm itself rather than auxiliary model design.

Generalization for other methods. Here, we also apply our BSCE to the other methods and the results in the Table 8 suggest that the

Table 8. Ablation study for BSCE for online generation.

	Model	R-Prec@3 ↑	FID ↓	MM Dist ↓	Diversity →
InterHuman	CAMDM	$0.587^{\pm.005}$	$4.000^{\pm.046}$	$3.828^{\pm.001}$	$7.547^{\pm.025}$
	CAMDM w. BSCE	$0.624^{\pm.005}$	$3.271^{\pm.059}$	$3.803^{\pm.001}$	$7.685^{\pm.024}$
	ReGenNet	$0.600^{\pm.004}$	$4.176^{\pm.085}$	$3.817^{\pm.001}$	$7.480^{\pm.033}$
	ReGenNet w. BSCE	$0.611^{\pm.004}$	$3.914^{\pm.055}$	$3.815^{\pm.001}$	$7.563^{\pm.033}$
	ARMFlow w/o. BSCE	$0.695^{\pm.005}$	$\underline{2.287}^{\pm.050}$	$3.796^{\pm.001}$	$7.702^{\pm.028}$
	ARMFlow **(Ours)**	$\mathbf{0.699}^{\pm.005}$	$\mathbf{2.178}^{\pm.054}$	$\mathbf{3.783}^{\pm.002}$	$\underline{7.745}^{\pm.024}$
InterX	CAMDM	$0.587^{\pm.003}$	$0.429^{\pm.011}$	$4.468^{\pm.020}$	$8.467^{\pm.072}$
	CAMDM w. VAE	$0.608^{\pm.003}$	$0.291^{\pm.007}$	$4.133^{\pm.018}$	$8.674^{\pm.075}$
	ReGenNet	$0.690^{\pm.004}$	$0.071^{\pm.003}$	$3.843^{\pm.011}$	$9.011^{\pm.053}$
	ReGenNet w. VAE	$0.689^{\pm.004}$	$0.066^{\pm.006}$	$3.792^{\pm.014}$	$\underline{8.927}^{\pm.061}$
	ARMFlow w/o. VAE	$0.710^{\pm.004}$	$\underline{0.045}^{\pm.003}$	$\mathbf{3.725}^{\pm.010}$	$8.917^{\pm.066}$
	ARMFlow **(Ours)**	$\mathbf{0.711}^{\pm.004}$	$\mathbf{0.042}^{\pm.003}$	$\underline{3.728}^{\pm.012}$	$\mathbf{8.939}^{\pm.071}$

10. Ablations on the CNN VAE

In this section, we conduct an ablation study on the CNN-VAE used in our model as shown in Table 9. First, we compare the performance of our model in the CNN-VAE latent space against that in the raw data space, and additionally compare it with the CAMDM and ReGenNet models. It is important to note that, since ReGenNet does not encode the historical actions of the reactor, we only augment the actor's actions in its experiments.

The results show that on the more challenging InterHuman dataset, which has longer average sequence lengths, CAMDM achieves a larger improvement. In contrast, ReGenNet shows less improvement due to its inability to leverage the historical context of the reactor. On the InterX dataset, which has shorter sequences and smaller action amplitudes, our method still provides improvement, although it is less pronounced than on InterHuman. These findings indicate that our approach is particularly effective for long-sequence scenarios.

Table 9. Ablation study for CNN VAE in online generation.

	Model	R-Prec@3 ↑	FID ↓	MM Dist ↓	Diversity →
InterHuman	CAMDM	$0.587^{\pm.005}$	$4.000^{\pm.046}$	$3.828^{\pm.001}$	$7.547^{\pm.025}$
	CAMDM w. VAE	$0.592^{\pm.004}$	$3.924^{\pm.050}$	$3.814^{\pm.001}$	$7.572^{\pm.018}$
	ReGenNet	$0.600^{\pm.004}$	$4.176^{\pm.085}$	$3.517^{\pm.001}$	$7.480^{\pm.033}$
	ReGenNet w. VAE	$0.638^{\pm.004}$	$3.801^{\pm.063}$	$3.805^{\pm.002}$	$7.623^{\pm.033}$
	ARMFlow w/o. VAE	$0.675^{\pm.005}$	$\underline{2.287}^{\pm.050}$	$3.796^{\pm.001}$	$7.702^{\pm.028}$
	ARMFlow **(Ours)**	$\mathbf{0.699}^{\pm.005}$	$\mathbf{2.178}^{\pm.054}$	$\mathbf{3.783}^{\pm.002}$	$\underline{7.745}^{\pm.024}$
InterX	CAMDM	$0.587^{\pm.003}$	$0.429^{\pm.011}$	$4.468^{\pm.020}$	$8.467^{\pm.072}$
	CAMDM w. VAE	$0.608^{\pm.003}$	$0.291^{\pm.007}$	$4.133^{\pm.018}$	$8.674^{\pm.075}$
	ReGenNet	$0.690^{\pm.004}$	$0.071^{\pm.003}$	$3.843^{\pm.011}$	$9.011^{\pm.053}$
	ReGenNet w. VAE	$0.689^{\pm.004}$	$0.066^{\pm.006}$	$3.792^{\pm.014}$	$\underline{8.927}^{\pm.061}$
	ARMFlow w/o. VAE	$0.710^{\pm.004}$	$\underline{0.045}^{\pm.003}$	$\mathbf{3.725}^{\pm.010}$	$8.917^{\pm.066}$
	ARMFlow **(Ours)**	$\mathbf{0.711}^{\pm.004}$	$\mathbf{0.042}^{\pm.003}$	$\underline{3.728}^{\pm.012}$	$\mathbf{8.939}^{\pm.071}$

11. Classifier-Free Guidance

In this section, we show the performance with different classifier-free guidance strength ω in Table 10. And we finally choose $w = 1.8$ for InterHuman dataset and $\omega = 1.2$ for InterX dataset.

Table 10. CFG for online generation on methods.

	Model	R-Prec@3 ↑	FID ↓	MM Dist ↓	Diversity →
InterHuman	$\omega = 1.0$	$0.673^{\pm.005}$	$2.436^{\pm.037}$	$3.790^{\pm.001}$	$7.716^{\pm.021}$
	$\omega = 1.2$	$0.692^{\pm.004}$	$\underline{2.273}^{\pm.058}$	$3.814^{\pm.001}$	$\mathbf{7.789}^{\pm.034}$
	$\omega = 1.8$	$\mathbf{0.699}^{\pm.005}$	$\mathbf{2.178}^{\pm.054}$	$\mathbf{3.783}^{\pm.002}$	$7.745^{\pm.024}$
	$\omega = 2.4$	$\underline{0.695}^{\pm.004}$	$2.518^{\pm.063}$	$\underline{3.784}^{\pm.002}$	$\underline{7.758}^{\pm.033}$
	$\omega = 3.0$	$0.669^{\pm.004}$	$2.873^{\pm.062}$	$3.792^{\pm.001}$	$7.738^{\pm.029}$
InterX	$\omega = 1.0$	$0.702^{\pm.003}$	$0.109^{\pm.005}$	$3.752^{\pm.020}$	$9.073^{\pm.085}$
	$\omega = 1.2$	$\mathbf{0.711}^{\pm.004}$	$\mathbf{0.042}^{\pm.003}$	$\mathbf{3.728}^{\pm.012}$	$8.939^{\pm.071}$
	$\omega = 1.8$	$\underline{0.709}^{\pm.003}$	$\underline{0.050}^{\pm.003}$	$\mathbf{3.728}^{\pm.012}$	$\underline{8.942}^{\pm.082}$
	$\omega = 2.4$	$0.704^{\pm.004}$	$0.059^{\pm.003}$	$\underline{3.749}^{\pm.011}$	$8.927^{\pm.075}$
	$\omega = 3.0$	$0.698^{\pm.003}$	$0.074^{\pm.004}$	$3.757^{\pm.012}$	$\mathbf{8.949}^{\pm.078}$

12. Comparison with CAMDM

In this section, we compare our result with CAMDM on the InterHuman [23] dataset, which uses a fixed-length window to encode the previous history. However, this results in information loss from a long history and a lack of consistency between windows. In contrast, we encode the whole history and use the history c_i at timestep i to guide the current generations. We use the same text prompts and action sequences and compare their results on the most challenging dataset InterHuman [23], suggesting that the sliding window is not stable in generating long-term motions. Especially in the case of boxing, which requires quick and vivid response. The corresponding results are shown in Fig. 5.

13. Compare on InterX

In this section, we compare the visualization resutl with ReGenNet [46] as shown in Fig. 6. Although the numerical results are similar on this dataset, we still find that our method outperforms ReGenNet in terms of visualization. Compared to ReGenNet, our ARMFlow is more capable of generating realistic motions, which reduces penetrations and performs better with semantic alignment and fidelity.

14. Evaluation Metrics

Frechet Inception Distance (FID): To quantify the discrepancy between the distributions of generated and real motion features, we employ the Frechet Inception Distance:

$$d_F(\mathcal{N}(\mu, \mathbf{\Sigma}), \mathcal{N}(\mu', \mathbf{\Sigma}'))^2 = \\ ||\mu - \mu'||^2 + \mathrm{Tr}(\mathbf{\Sigma} + \mathbf{\Sigma}' - 2\sqrt{\mathbf{\Sigma}\mathbf{\Sigma}'}), \tag{8}$$

where μ and μ' denote the mean vectors of features extracted from generated motions and ground truth, respectively, and $\mathbf{\Sigma}$, $\mathbf{\Sigma}'$ are the corresponding covariance matrices. Intuitively, FID captures the distance between two multivariate Gaussian approximations of the feature distributions. Lower FID values indicate a higher fidelity of generated samples. For motion sequences with extensive temporal length, we compute FID on high-level feature representations extracted via pre-trained action recognition models [11, 19], rather than on raw motion data. For the InterHuman dataset [45], a Transformer encoder is used to extract the global feature of the HHI sequences, while for the InterX dataset.

Diversity: Diversity reflects the variability among generated motions across different action categories. Specifically, two subsets of equal size S_d are randomly sampled from the generated motions, yielding feature sets $\{\mathbf{f_1}, \ldots, \mathbf{f_{S_d}}\}$ and $\{\mathbf{f'_1}, \ldots, \mathbf{f'_{S_d}}\}$. The diversity metric is defined as

$$\mathrm{Div} = \frac{1}{S_d} \sum_{i=1}^{S_d} ||\mathbf{f_i} - \mathbf{f'_i}||^2. \tag{9}$$

An ideal generative model should produce motions exhibiting diversity comparable to that observed in the ground truth.

Multimodality: While diversity captures inter-class variance, multimodality measures intra-class variation. For motions corresponding to C distinct action types, we randomly

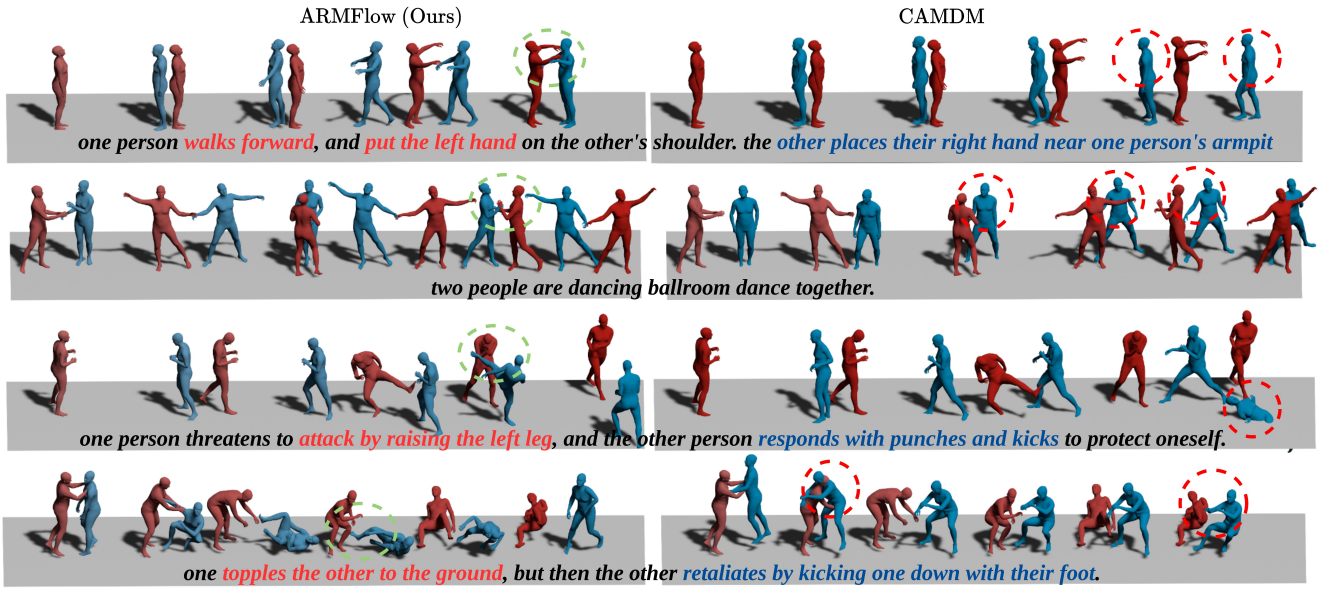

Figure 5. Comparison between ARMFlow and CAMDM on InterHuman Dataset

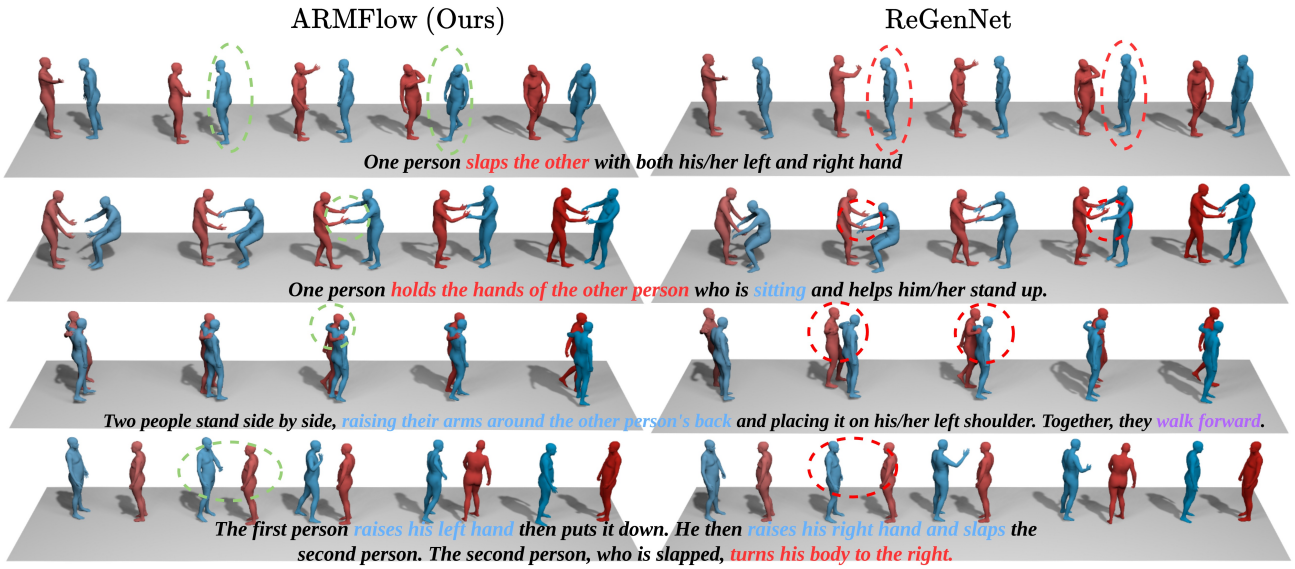

Figure 6. Comparison between ARMFlow and ReGenNet on InterX

select two subsets of size S_l for each action c, with feature vectors $\{\mathbf{f}_{c,1}, \ldots, \mathbf{f}_{c,S_l}\}$ and $\{\mathbf{f}'_{c,1}, \ldots, \mathbf{f}'_{c,S_l}\}$. The multimodality score is computed as

$$\text{Multimodality} = \frac{1}{C \cdot S_l} \sum_{c=1}^{C} \sum_{i=1}^{S_l} ||\mathbf{f}_{c,i} - \mathbf{f}'_{c,i}||^2. \quad (10)$$

Multimodal Distance (MM-Dist): This metric quantifies the alignment between generated motion features and their corresponding text embeddings. Given text $\mathbf{T}$ with feature $\mathbf{f_T}$ and generated motion $\mathbf{M}$ with feature $\mathbf{f_M}$, the multimodal distance is defined as

$$\text{MMD} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} ||\mathbf{f_{T,i}} - \mathbf{f_{M,i}}||^2}, \quad (11)$$

where n is the number of samples. Text features are extracted using a pre-trained text encoder [12]. This metric is also applicable to motions generated conditioned on other modalities, such as audio or music.

R-Precision: R-Precision, often referred to as motion-retrieval accuracy, assesses the top-K matching between generated motions and textual descriptions. Following the procedure of Guo et al. [12], for each generated motion, the ground-truth description is combined with 31 randomly sampled mismatched captions. Distances between motion features and all candidate text features are computed and ranked, and accuracy is reported for top 1, 2, and 3 retrieval positions.

15. Noisy Motion Condition

We evaluate ReGenNet, R2R, and ours with noisy input by adding the same noise level to the actor conditions (Tab 11). Due to the double augmentation of BSCE for the actor and the reactor, our method is more robust to noisy input.

Table 11. Results with noisy input

	Model	R-Prec@3 ↑	FID ↓	MM Dist ↓	Diversity →
InterH	Ground Truth	0.701 → 0.699	0.273 → 0.495	3.755 → 3.787	7.948 → 7.989
	ReGenNet	0.600 → 0.407	4.176 → 15.66	3.817 → 3.898	7.480 → 7.141
	R2R	0.674 → 0.641	2.795 → 6.138	3.793 → 3.812	7.693 → 7.702
	ARMFlow **(Ours)**	**0.699** → **0.664**	**2.178** → **4.101**	**3.783** → **3.794**	**7.745** → **7.788**
InterX	Ground Truth	0.736 → 0.732	0.002 → 0.028	3.574 → 3.605	8.947 → 8.931
	ReGenNet	0.690 → 0.635	0.071 → 0.740	3.843 → 4.165	9.011 → 8.573
	R2R	0.704 → 0.662	0.063 → 0.612	3.745 → 4.017	8.873 → 8.629
	ARMFlow **(Ours)**	**0.711** → **0.676**	**0.042** → **0.467**	**3.728** → **3.934**	**8.939** → **8.781**

16. User Study

We conducted a user study comparing R2R, ReGenNet, and our approach on both datasets, dividing them into four groups: boxing, dancing, cooperation, and daily actions. We created an anonymous Google Form and invited 25 participants to choose the most realistic motion sequence. Fig. 7 summarizes the 21 valid responses received. We can see that motions generated by our method are preferred in all four action groups. Notably, since daily actions lack of close contact, therefore, the preference for our methods are not obvious compared to motions with close contact like boxing, dancing, and cooperations.

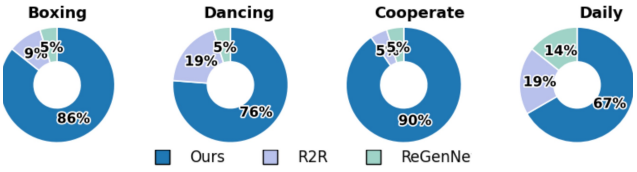

Figure 7. User study results.

17. Physical Results

We calculated (1) the contact success ratio for sequences containing contact, and (2) the penetrated volume per sequence (in Liters). Tab. 12 shows that our approach has the lowest penetration and highest contact ratio.

Table 12. Physical metrics on InterHuman and InterX dataset

	Model	Contact Success Ratio ↑	Penetration Score ↓
InterH	Ground Truth	1.000	0.004
	ReGenNet	0.763	0.075
	R2R	0.848	0.049
	ARMFlow **(Ours)**	**0.870**	**0.036**
InterX	Ground Truth	1.000	0.008
	ReGenNet	0.833	0.043
	R2R	0.795	0.024
	ARMFlow **(Ours)**	**0.854**	**0.020**