SINGLE STREAM, SINGLE STEP: SIMPLIFYING 3D HUMAN–HUMAN INTERACTION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating realistic 3D human–human interactions requires plausible individual motion and coherent coordination between two individuals. Existing approaches often use separate per-person streams or specialized interaction modules, while some additionally rely on learned motion autoencoders. These designs complicate the architecture or training pipeline, and iterative sampling adds inference cost. We introduce InterFlow, an end-to-end autoregressive flow-matching model that simplifies interaction generation using a single-stream diffusion transformer. InterFlow operates directly in motion space without a variational autoencoder, representing the motions of both individuals as a unified token sequence and modeling temporal and interpersonal dependencies through shared self-attention. Its fixed-window formulation enables blockwise generation with a computational cost per block independent of the total generated sequence length. We further introduce FastInterFlow, which combines rollout-based score distillation with forward-kinematics-based adversarial distillation to obtain a one-step generator from the multi-step teacher. FastInterFlow retains the autoregressive formulation and generates each motion block in a single network evaluation. Empirical results demonstrate the effectiveness of InterFlow's streamlined architecture for modeling interactions. Together, the two models provide a simple framework combining unified motion modeling with efficient blockwise autoregressive generation.

1 INTRODUCTION

Generating realistic 3D human–human interactions requires modeling deeply coupled spatiotemporal dynamics: motions plausible in isolation can easily fail to form a coherent interaction without precise physical contact, relative orientation, and temporal synchrony. The central challenge is therefore to capture the joint evolution of two individuals while preserving individual motion fidelity.

Existing approaches typically tackle these dependencies through specialized architectures or learned motion representations. One paradigm employs multi-stream networks with explicit coupling mechanisms: InterGen (Liang et al., 2024) connects dual denoisers via mutual cross-attention, in2IN (Ruiz-Ponce et al., 2024) cross-conditions separate streams, and Interact2Ar (Ruiz-Ponce et al., 2026) introduces parallel denoiser branches dedicated separately to trajectories, bodies, and articulate hands. Another paradigm compresses motion into discrete tokens via learned autoencoders before modeling interactions with masked transformers, such as InterMask (Javed et al., 2024). While effective, these designs introduce notable drawbacks: branch-segregated pipelines require handcrafted architectural routing that fragments the motion manifold, while autoencoder latent spaces impose lossy bottlenecks that smooth out delicate contacts and finger gestures. These limitations motivate directly modeling joint dynamics in continuous raw motion space without dedicated interaction modules or autoencoders.

Inference efficiency for extended interactions poses an additional hurdle. Full-sequence generation cannot accommodate dynamic context updates, while unconstrained autoregression suffers from quadratic computational growth and severe exposure bias. Furthermore, iterative numerical sampling creates a latency bottleneck during sequential block rollouts. An effective framework must therefore bound the autoregressive computational footprint while condensing sampling steps.

To address these challenges, we introduce **InterFlow**, an end-to-end framework operating directly in raw motion space without an autoencoder. InterFlow represents both interactants as a unified

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

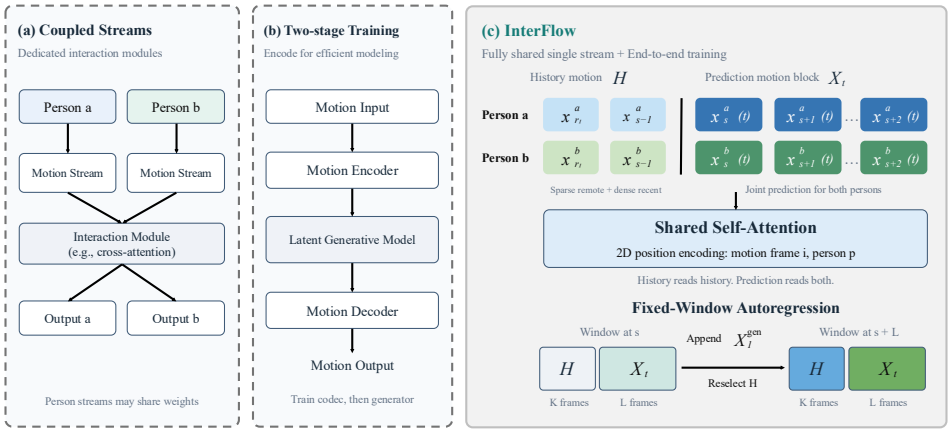


Figure 1: **Interaction generation through shared attention.** (a) Separate person streams coupled by dedicated attention and (b) generation in a learned motion space illustrate two design patterns prevalent in prior work . (c) InterFlow directly represents each person at each frame as an independent motion token. Two-dimensional coordinates (i, p) identify the motion frame and person identity within shared self-attention. A history H with a fixed budget K conditions a prediction block X_t of L frames. After flow integration, the generated block X_1^{gen} is appended, and the history is re-sampled for the next boundary $s + L$. Token counts are schematic.

sequence of individually addressable *person–frame tokens*. Equipped with factorized 2D Rotary Position Embeddings (2D RoPE), a standard single-stream Diffusion Transformer (DiT) inherently captures both individual kinematics and interpersonal coordination through shared self-attention. InterFlow generates motion in fixed-window blocks, bounding the per-block computational cost independently of total sequence duration, while clean endpoint (x_1) parameterization and context diffusion forcing ensure stable high-dimensional optimization and rollout robustness.

To eliminate the multi-step sampling bottleneck, we further propose **FastInterFlow**, distilling InterFlow into a one-step blockwise generator. FastInterFlow introduces a dual-representation distillation strategy tailored to interactions: rollout-based score distillation matches trajectory probability fields under student-generated contexts, while forward-kinematics (FK) adversarial distillation enforces physical plausibility and contact coordination directly on explicit 3D joints.

Our main contributions are three-fold: (1) We propose InterFlow, a single-stream framework that models two-person interaction directly in raw motion space via unified person–frame tokens, dispensing with specialized interaction modules or autoencoders while bounding per-block compute via fixed-window autoregression; (2) We develop FastInterFlow, the first one-step generator for autoregressive interaction synthesis, combining continuous rollout score distillation with explicit 3D kinematic adversarial distillation for rapid, high-fidelity generation; and (3) We conduct extensive experiments and ablations on Inter-X (Xu et al., 2024), demonstrating superior whole-body and hand coordination while providing practical design guidance for multi-person motion modeling.

2 Related Work

2.1 Human Motion and Interaction Generation

Text-driven human motion generation has advanced rapidly with diffusion backbones (Tevet et al., 2022; Zhang et al., 2024; Chen et al., 2023) and paired corpora (Guo et al., 2022). While some methods discretize motion into tokens for autoregressive or masked modeling (Zhang et al., 2023; Jiang et al., 2023; Guo et al., 2024), learned codebooks (Kingma & Welling, 2013) impose representation bottlenecks that can compromise fine-grained geometric details and bias evaluation metrics (Meng et al., 2025b;a). Synthesizing two-person interactions introduces further complexity due to coupled interpersonal coordination (Zhu et al., 2023; Sui et al., 2026). Recent benchmarks (Liang et al., 2024; Xu et al., 2024) have driven various architectural paradigms, including mutual attention (Liang et al., 2024), dual-stream conditioning (Ruiz-Ponce et al., 2024), collaborative masked tokens (Javed et al., 2024), human–object inspired modules (Xu et al., 2023; 2025a), trajectory or reactive con-

trol (Wang et al., 2024; Fan et al., 2024; Ghosh et al., 2024; Li et al., 2024a;b; Xu et al., 2025b), long-horizon architectures (Tanke et al., 2023; 2025; Yu et al., 2026a), and segregated body-hand branches (Ruiz-Ponce et al., 2026). In contrast to treating coupling as an architectural primitive, InterFlow structures interactants as unified person–frame tokens within a single sequence, learning coordination directly in raw motion space via standard self-attention.

Full-sequence generation is inherently ill-suited for open-ended interactions where actors dynamically react to each other. Autoregressive methods address this by predicting local blocks from past context across discrete tokens (Zhang et al., 2023; Jiang et al., 2023), continuous control models (Chen et al., 2024b; Zhao et al., 2025; Xiao et al., 2025; Yu et al., 2026b), physics simulators (Tevet et al., 2025; Zhang et al., 2025), online reaction frameworks (Cen et al., 2025; Luo et al., 2026), and streaming processes (Liu et al., 2026; Zhao et al., 2026). To avoid unbounded computational growth, memory-bounded schemes retain sparse distant states alongside dense recent frames (Dai et al., 2019; Xiao et al., 2024). To alleviate exposure bias between training history and inference rollouts, diffusion forcing corrupts past contexts with noise (Chen et al., 2024a; Huang et al., 2026). We adapt this fixed-budget autoregressive paradigm to two-person interaction, combining sparse remote and dense recent history with heterogeneous noise schedules to ensure constant per-block compute and rollout stability.

2.2 Few-Step Generation

Continuous-time flow matching (Lipman et al., 2022) and rectified flow (Liu et al., 2022) learn deterministic probability trajectories, unified by stochastic interpolants (Albergo et al., 2025) and scaled via transformers (Ma et al., 2024; Barquero et al., 2024). To overcome the latency of iterative ODE sampling and classifier-free guidance (Ho & Salimans, 2022), few-step models such as consistency models (Song et al., 2023; Luo et al., 2023; Zhou et al., 2024; Dai et al., 2024) learn direct mappings to the data manifold. Alternatively, distillation compresses teacher models via progressive distillation (Salimans & Ho, 2022), adversarial distillation (Sauer et al., 2024), score distillation sampling (Poole et al., 2022), distribution matching distillation (Wang et al., 2023; Yin et al., 2024b;a), and mean flows (Geng et al., 2026), with recent extensions to autoregressive video (Huang et al., 2026; Li et al., 2026). To accelerate sampling in our interactive setting, we adapt rollout-based score distillation alongside kinematic adversarial supervision to distill InterFlow into FastInterFlow, obtaining a practical one-step generator for efficient autoregressive block rollout.

3 Method

Motivation. Human–human interaction couples temporal dynamics with interpersonal coordination. Because both persons share the same articulated representation, their states can be jointly modeled as person–frame tokens within a single attention stream. We build on this observation to introduce **InterFlow**, an end-to-end flow-matching model operating directly in motion space, without separate person streams, dedicated interaction modules, or a motion autoencoder. Fixed-window block autoregression supports incremental generation with bounded network cost per block. We then distill InterFlow into **FastInterFlow**, combining rollout-based score distillation with a later stage of forward-kinematics adversarial supervision to generate each block in one network evaluation.

3.1 Problem Formulation and Preliminaries

Motion representation and notation. We use the SMPL-X-based representation of Inter-X (Xu et al., 2024; Pavlakos et al., 2019) to retain articulated hand motion. Let $i \in \mathbb{N}_0$ index a discrete motion frame and $p \in \{a, b\}$ identify a person. The motion vector $x_i^p \in \mathbb{R}^D$ contains 55 joint rotations in the continuous 6D representation (Zhou et al., 2019) and six padded root-velocity channels, giving $D = 336$. Thus,

$$x_i^p = [r_i^p, u_i^p] \in \mathbb{R}^{336}, \qquad r_i^p \in \mathbb{R}^{330}, \quad u_i^p \in \mathbb{R}^6, \qquad x_i = [x_i^a, x_i^b] \in \mathbb{R}^{672}. \tag{1}$$

We consistently use lowercase x_i^p for a single person's motion at one frame and uppercase X and H for prediction and history blocks. A superscript p selects a person's block, as in X_t^p and H^p. Continuous flow time is denoted by $t \in [0, 1]$; it is distinct from the motion-frame index i. The state of a prediction token at frame i and flow time t is $x_i^p(t)$, with $x_i^p(1) = x_i^p$.

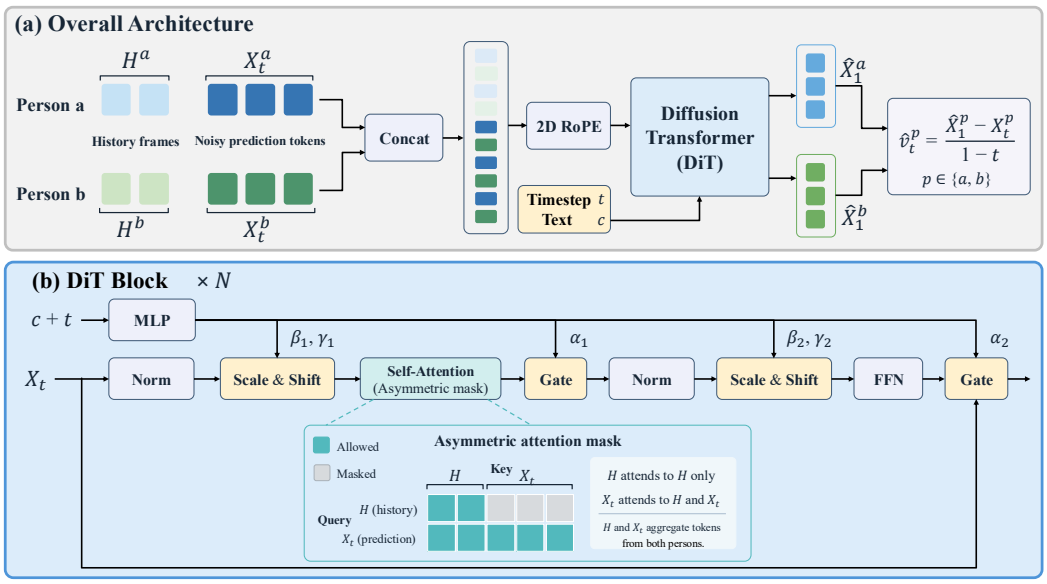


Figure 2: **InterFlow architecture.** (a) History blocks H^a, H^b and noisy prediction blocks X_t^a, X_t^b are projected into person–frame tokens and interleaved in one stream. The transformer predicts endpoints $\widehat{X}_1^p$, from which $\widehat{v}_\theta^p$ is obtained. (b) Each block operates on hidden tokens Z, with 2D rotary encoding applied to attention queries and keys. Text and token-specific flow times modulate the blocks. History queries read only history, while prediction queries read both groups.

Continuous-time flow matching. For a clean target block $X_1 \in \mathbb{R}^{L\times 672}$ and independent Gaussian noise $X_0 \sim \mathcal{N}(0, I)$, we use the linear interpolation path (Lipman et al., 2022)

$$X_t = (1-t)X_0 + tX_1, \qquad v^\star = X_1 - X_0. \tag{2}$$

Here, $v^\star$ is the conditional training target along a sampled path. A learned velocity field averages such targets at each noisy state; it need not be constant along a generated trajectory.

Blockwise autoregressive interface. At prediction boundary s, the model generates the next L frames $X_1 = [x_s, \ldots, x_{s+L-1}]$, conditioned on text c and a history block $H \in \mathbb{R}^{K\times 672}$. The history contains selected frames strictly before s, not the entire past. Fixing both K and L limits the attention input to $2(K+L)$ person–frame tokens. After generation, the block is appended, the boundary advances to $s+L$, and history is selected again. This bounds model computation per block as the sequence grows, while allowing the context to be updated between blocks.

3.2 Single-Stream Interaction Modeling

Person–frame tokenization. To expose temporal and interpersonal dependencies to the same attention operation, we embed each x_i^p separately with a shared linear projection. For window coordinates $i_1, \ldots, i_{K+L}$, where the first K coordinates are sampled history indices and the last L are $s, \ldots, s+L-1$, the stream is

$$Z = [z_{i_1}^a, z_{i_1}^b, \ldots, z_{i_{K+L}}^a, z_{i_{K+L}}^b] \in \mathbb{R}^{2(K+L)\times d}. \tag{3}$$

Each person–frame state is an independently addressable token. All motion projections and transformer weights are shared across persons, allowing a diffusion transformer (Peebles & Xie, 2023) to learn within-person dynamics and cross-person coordination through shared self-attention.

Two-dimensional positional structure. A flattened token index does not directly separate person identity from elapsed time, particularly when history is sparse. We therefore apply factorized 2D rotary encoding (Su et al., 2024) to queries and keys, using coordinates $(i, \pi(p))$, with $\pi(a) = 0$ and $\pi(b) = 1$:

$$\widetilde{q}_i^p = R_{2\mathrm{D}}(i, \pi(p))q_i^p, \qquad \widetilde{k}_j^{p'} = R_{2\mathrm{D}}(j, \pi(p'))k_j^{p'}. \tag{4}$$

Separate rotary subspaces encode temporal and person offsets. Original frame indices retain the temporal gaps between remote samples. The resulting attention scores depend on token content and signed relative coordinates; no separate interaction pathway is required.

Conditioning and asymmetric attention. Adaptive normalization combines pooled text features c (Radford et al., 2021) with token-specific flow-time embeddings. To keep history features independent of the evolving prediction states, we order history tokens first and apply

$$M = \begin{bmatrix} 0_{2K\times 2K} & -\infty_{2K\times 2L} \\ 0_{2L\times 2K} & 0_{2L\times 2L} \end{bmatrix}. \tag{5}$$

History queries access only valid history keys; prediction queries access history and all current-block tokens. Thus, generation is causal between blocks and bidirectional within a prediction block. During inference, clean history uses time 1. With its values, positions, text, and time conditions fixed, history K/V states can be cached across integration steps. Prediction queries still attend to both cached history and current prediction keys.

3.3 Context Design and InterFlow Training

Sparse remote and dense recent history. A fixed context budget must preserve local continuity without discarding all earlier interaction information. We allocate $K = K_{\mathrm{rem}} + K_{\mathrm{rec}}$ slots to sparse remote and dense recent frames. With sufficient past motion, one remote frame is sampled from each of K_{rem} equal temporal segments before the recent window. The joint history is

$$H = [x_{r_1}, \ldots, x_{r_{K_{\mathrm{rem}}}}, x_{s-K_{\mathrm{rec}}}, \ldots, x_{s-1}], \qquad r_1 < \cdots < r_{K_{\mathrm{rem}}} < s - K_{\mathrm{rec}}. \tag{6}$$

Both persons share these indices, preserving temporal alignment. Invalid slots are masked. Sparse samples retain earlier states, while recent frames support local motion continuity.

Endpoint parameterization. Each person–frame token is 336-dimensional. Direct velocity prediction requires representing both structured motion and an unstructured noise component. Following the clean-data parameterization of JiT (Li & He, 2026), we predict the clean motion endpoint to promote stable learning and then convert it to velocity:

$$\widehat{X}_1 = f_\theta(X_t, t, H, c), \qquad \widehat{v}_\theta(X_t, t, H, c) = \frac{\widehat{X}_1 - X_t}{1-t}, \qquad t < 1. \tag{7}$$

We retain velocity supervision,

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\left[\left\| \widehat{v}_\theta(X_t, t, \widetilde{H}, c) - (X_1 - X_0) \right\|^2\right], \tag{8}$$

where the norm averages prediction entries. Endpoint prediction is a parameterization choice, not an additional paired endpoint loss. InterFlow uses this objective without auxiliary kinematic regression.

Diffusion forcing for imperfect context. Clean-history training does not expose the model to the errors present in generated history, whereas obtaining multi-step training rollouts is costly. We therefore corrupt the context following the diffusion-forcing principle (Chen et al., 2024a). With probability p_{df}, each remote slot ℓ receives an independent noise time τ_ℓ, while all recent slots share τ_{rec}, independently of prediction time t:

$$\widetilde{H}_\ell = (1-\tau_\ell)\eta_\ell + \tau_\ell H_\ell, \qquad \eta_\ell \sim \mathcal{N}(0, I). \tag{9}$$

The noise times lie in $[0, 1]$ and condition their corresponding history; uncorrupted history uses $\tau_\ell = 1$. The two persons share each slot's time. Independent remote times expose heterogeneous context reliability, while a common recent time preserves a shared corruption level over the local window. This is a low-cost proxy for imperfect history, not an exact simulation of rollout errors.

Sequence initialization. To unify text-only cold starts and context-conditioned continuation within a single framework, we stochastically drop the history context near the sequence onset with a distance-aware probability $p_{\max}\max(0, 1-s/K)$ and oversample the initial boundary $s = 0$. These operations expose the model to absent context during training, eliminating the need for a dedicated launch network. Implementation details are deferred to Appendix D.3.

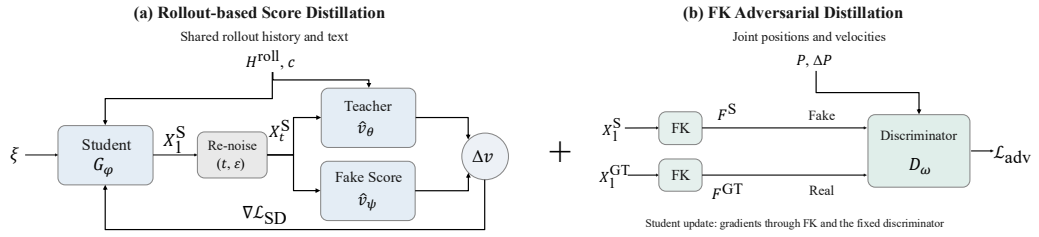


Figure 3: **FastInterFlow distillation.** (a) the student generates from rollout history, and the teacher and fake-score model evaluate the same re-noised student block under the same context. Their field difference supplies the score-distillation update. (b) FK maps motion into joint positions and velocities. GT features are real, and student features are fake. Adversarial supervision is introduced once score-distillation training has stabilized.

3.4 FastInterFlow: One-Step Distillation

Although InterFlow enables incremental generation, it still requires iterative numerical integration per block. We distill it into a one-step student via two complementary objectives: rollout-based score distillation to transfer the teacher's distribution under student context, and forward-kinematics adversarial distillation to supervise joint-space motion quality (Figure 3).

Rollout-based score distillation. Initialized from InterFlow, the student produces a complete motion block in one forward pass:

$$X_1^{\text{S}} = G_\phi(\xi, H^{\text{roll}}, c) := f_\phi(\xi, 0, H^{\text{roll}}, c), \qquad \xi \sim \mathcal{N}(0, I). \tag{10}$$

The low sampling cost allows us to construct H^{roll} from preceding student predictions, following the self-forcing principle (Huang et al., 2026). This exposes the student to its own history errors during training. We progressively increase the rollout budget and detach the generated history, concentrating gradients on the current block.

We adapt score distillation (Wang et al., 2023) to flow matching by directly comparing velocity fields. The teacher field $\widehat{v}_\theta$ is frozen, while an auxiliary field $\widehat{v}_\psi$, termed the fake-score model, is trained by flow matching on detached student samples to estimate the student distribution. We re-noise X_1^{S} using Equation equation 2, substituting X_1^{S} for the data endpoint and independent $\epsilon \sim \mathcal{N}(0, I)$ for the noise endpoint, to obtain X_t^{S}. We then compute

$$\Delta v = \widehat{v}_\theta(X_t^{\text{S}}, t, H^{\text{roll}}, c) - \widehat{v}_\psi(X_t^{\text{S}}, t, H^{\text{roll}}, c). \tag{11}$$

Both fields evaluate the same noisy block with the same rollout history and text, so their difference compares the reference and student distributions under a shared context. The student receives this detached direction through

$$\mathcal{L}_{\text{SD}} = \mathbb{E}\left[\left\|X_t^{\text{S}} - \text{sg}[X_t^{\text{S}} + \Delta v]\right\|_{\text{valid}}^2\right], \tag{12}$$

where sg denotes stop-gradient. We use unit loss weight at all flow times, without time-dependent reweighting. The score-to-velocity conversion cancels the explicit parameterization factors; its derivation and the fake-score training objective are given in Appendix F.1.

Forward-kinematics adversarial distillation. Score distillation operates in the learned motion-feature distribution. To provide explicit joint-space supervision, we exploit the fact that each student output is already a complete motion sample. Direct regression to an arbitrarily paired GT trajectory would impose a noise-to-example correspondence and can suppress diversity. Instead, we match kinematic features adversarially. Let $P = \text{FK}(X)$ denote world-space joint positions and $\Delta P_i = P_i - P_{i-1}$ the corresponding joint velocities. We form

$$\Phi(P) = [P, \Delta P], \qquad F^{\text{S}} = \Phi(\text{FK}(X_1^{\text{S}})). \tag{13}$$

A compact discriminator D_ω uses shared person–frame self-attention with 2D positional encoding to assess these features. GT kinematics F^{GT} provide real examples, while detached student features provide fake examples. The discriminator learns to distinguish their kinematic distributions. With its parameters fixed, the student's adversarial objective is

$$\mathcal{L}_{\text{adv}} = \mathbb{E}\big[\,\text{softplus}(-D_\omega(F^{\text{S}}))\big]. \tag{14}$$

Table 1: **Comparison on Inter-X.** $*$ denotes GT-first-frame initialization. Means and 95% confidence intervals are reported over 20 runs (all other experimental evaluations use this configuration.). Bold/underlined means mark the best/second-best distinct values among generators.

Eval.	Method	R-Precision ↑			FID ↓	MM Dist ↓	Div. →	MMod. ↑
		Top-1	Top-2	Top-3				
Full	*Ground Truth*	$0.431^{\pm.00}$	$0.631^{\pm.00}$	$0.740^{\pm.00}$	$0.002^{\pm.00}$	$3.318^{\pm.01}$	$8.973^{\pm.10}$	–
	T2M	$0.211^{\pm.00}$	$0.339^{\pm.00}$	$0.434^{\pm.00}$	$9.079^{\pm.11}$	$5.766^{\pm.02}$	$7.994^{\pm.08}$	$0.789^{\pm.03}$
	InterGen	$0.411^{\pm.01}$	$0.608^{\pm.00}$	$0.721^{\pm.00}$	$0.874^{\pm.02}$	$3.618^{\pm.01}$	$\mathbf{8.876}^{\pm.07}$	$\mathbf{3.345}^{\pm.07}$
	InterMask	$0.415^{\pm.00}$	$0.607^{\pm.00}$	$0.722^{\pm.00}$	$0.671^{\pm.02}$	$3.487^{\pm.01}$	$8.654^{\pm.06}$	$1.686^{\pm.06}$
	Interact2Ar*	$0.453^{\pm.00}$	$0.661^{\pm.00}$	$0.773^{\pm.00}$	$0.277^{\pm.01}$	$3.095^{\pm.01}$	$9.305^{\pm.07}$	$1.427^{\pm.04}$
	InterFlow	$\underline{0.479}^{\pm.00}$	$\underline{0.689}^{\pm.00}$	$\mathbf{0.799}^{\pm.00}$	$0.208^{\pm.01}$	$\underline{2.969}^{\pm.02}$	$9.500^{\pm.09}$	$1.823^{\pm.08}$
	InterFlow*	$\mathbf{0.480}^{\pm.00}$	$\mathbf{0.690}^{\pm.00}$	$\mathbf{0.799}^{\pm.00}$	$0.192^{\pm.01}$	$\mathbf{2.962}^{\pm.01}$	$9.463^{\pm.08}$	$1.445^{\pm.06}$
	FastInterFlow	$0.445^{\pm.00}$	$0.648^{\pm.00}$	$0.760^{\pm.00}$	$\underline{0.139}^{\pm.01}$	$3.183^{\pm.02}$	$\underline{9.172}^{\pm.08}$	$\underline{2.224}^{\pm.08}$
	FastInterFlow*	$0.464^{\pm.00}$	$0.668^{\pm.00}$	$\underline{0.779}^{\pm.00}$	$\mathbf{0.136}^{\pm.01}$	$3.100^{\pm.02}$	$9.202^{\pm.07}$	$1.548^{\pm.06}$
Body	*Ground Truth*	$0.462^{\pm.01}$	$0.655^{\pm.00}$	$0.758^{\pm.00}$	$0.001^{\pm.00}$	$3.272^{\pm.01}$	$9.002^{\pm.07}$	–
	T2M	$0.222^{\pm.00}$	$0.354^{\pm.00}$	$0.452^{\pm.00}$	$3.416^{\pm.07}$	$5.436^{\pm.02}$	$8.018^{\pm.07}$	$0.905^{\pm.03}$
	InterGen	$0.393^{\pm.00}$	$0.571^{\pm.00}$	$0.671^{\pm.00}$	$5.762^{\pm.30}$	$4.564^{\pm.03}$	$10.45^{\pm.09}$	$\mathbf{4.566}^{\pm.13}$
	InterMask	$0.386^{\pm.00}$	$0.565^{\pm.01}$	$0.664^{\pm.00}$	$6.720^{\pm.28}$	$4.616^{\pm.03}$	$10.61^{\pm.11}$	$\underline{2.493}^{\pm.12}$
	Interact2Ar*	$0.469^{\pm.00}$	$0.672^{\pm.00}$	$0.779^{\pm.00}$	$0.352^{\pm.01}$	$3.173^{\pm.01}$	$9.271^{\pm.08}$	$1.421^{\pm.04}$
	InterFlow	$\underline{0.495}^{\pm.01}$	$\mathbf{0.702}^{\pm.00}$	$\mathbf{0.802}^{\pm.00}$	$0.264^{\pm.01}$	$\underline{2.993}^{\pm.02}$	$9.409^{\pm.10}$	$1.897^{\pm.06}$
	InterFlow*	$\mathbf{0.497}^{\pm.00}$	$\underline{0.697}^{\pm.00}$	$\underline{0.799}^{\pm.00}$	$0.250^{\pm.01}$	$\mathbf{2.992}^{\pm.02}$	$9.299^{\pm.10}$	$1.520^{\pm.06}$
	FastInterFlow	$0.461^{\pm.00}$	$0.658^{\pm.00}$	$0.764^{\pm.00}$	$\underline{0.154}^{\pm.01}$	$3.240^{\pm.01}$	$\underline{9.087}^{\pm.08}$	$2.353^{\pm.07}$
	FastInterFlow*	$0.483^{\pm.00}$	$0.685^{\pm.00}$	$0.785^{\pm.00}$	$\mathbf{0.144}^{\pm.01}$	$3.080^{\pm.01}$	$\mathbf{9.066}^{\pm.10}$	$1.584^{\pm.06}$
Hands	*Ground Truth*	$0.399^{\pm.00}$	$0.597^{\pm.00}$	$0.713^{\pm.00}$	$0.002^{\pm.00}$	$3.312^{\pm.01}$	$8.370^{\pm.05}$	–
	T2M	$0.196^{\pm.00}$	$0.317^{\pm.00}$	$0.405^{\pm.01}$	$9.424^{\pm.12}$	$5.398^{\pm.02}$	$6.996^{\pm.11}$	$0.791^{\pm.04}$
	InterGen	$0.357^{\pm.00}$	$0.536^{\pm.01}$	$0.645^{\pm.00}$	$1.389^{\pm.03}$	$3.806^{\pm.02}$	$8.006^{\pm.07}$	$\mathbf{3.736}^{\pm.11}$
	InterMask	$0.360^{\pm.00}$	$0.538^{\pm.00}$	$0.647^{\pm.00}$	$1.960^{\pm.05}$	$3.794^{\pm.02}$	$7.939^{\pm.07}$	$1.895^{\pm.09}$
	Interact2Ar*	$0.422^{\pm.00}$	$0.629^{\pm.00}$	$0.745^{\pm.00}$	$0.257^{\pm.01}$	$3.111^{\pm.01}$	$8.614^{\pm.08}$	$1.439^{\pm.05}$
	InterFlow	$\mathbf{0.449}^{\pm.00}$	$\mathbf{0.654}^{\pm.00}$	$\mathbf{0.765}^{\pm.00}$	$0.216^{\pm.01}$	$\underline{3.051}^{\pm.02}$	$8.858^{\pm.09}$	$1.949^{\pm.07}$
	InterFlow*	$\underline{0.442}^{\pm.00}$	$\underline{0.648}^{\pm.00}$	$\underline{0.761}^{\pm.00}$	$\mathbf{0.164}^{\pm.01}$	$\mathbf{3.039}^{\pm.01}$	$8.737^{\pm.07}$	$1.504^{\pm.05}$
	FastInterFlow	$0.409^{\pm.00}$	$0.609^{\pm.01}$	$0.725^{\pm.01}$	$\underline{0.195}^{\pm.01}$	$3.261^{\pm.02}$	$\underline{8.578}^{\pm.07}$	$\underline{2.385}^{\pm.08}$
	FastInterFlow*	$0.421^{\pm.00}$	$0.624^{\pm.00}$	$0.740^{\pm.00}$	$0.197^{\pm.01}$	$3.181^{\pm.01}$	$\mathbf{8.520}^{\pm.07}$	$1.620^{\pm.06}$

We introduce adversarial supervision after score-distillation training stabilizes, retaining the score objective throughout subsequent training. During student updates, gradients propagate through the fixed discriminator and differentiable FK layers, while the teacher remains frozen and both the auxiliary score model and discriminator receive separate updates. Detailed training objectives and update schedules are provided in Appendix F.2. At inference, only G_ϕ is deployed.

4 Experiments

4.1 Experimental Setup

Dataset and evaluation. We evaluate on Inter-X (Xu et al., 2024), comprising approximately 11K two-person motion capture sequences with articulate hand poses and fine-grained text descriptions. To evaluate motion fidelity at different granularities, we adopt the multi-level evaluation proposed by Interact2Ar (Ruiz-Ponce et al., 2026), evaluating across three structural tiers: *Full*, *Body*, and *Hands*. Following standard protocols (Guo et al., 2022), each evaluator reports **R-Precision**, **MM Dist** (text–motion feature distance), **FID** (distributional fidelity to real motion), **Diversity** (sample variation across generation), and **MModality** (variation under identical text conditions). We report means and 95% confidence intervals over 20 runs; more details are in Appendix G.

Implementation details. InterFlow employs an 8-block single-stream DiT (41.7M parameters) with a hidden dimension of 512. It predicts 30-frame motion blocks conditioned on a fixed 15-frame history (10 sparse remote and 5 dense recent frames). Generation utilizes an Euler solver with 10 uniform steps and a classifier-free guidance scale of 1.5. FastInterFlow is initialized from the pre-trained InterFlow checkpoint and trained via rollout-based score distillation (200 epochs) followed by joint forward-kinematics adversarial distillation (100 epochs). At inference, FastInter-

Table 2: **Architectural ablations.** Evaluating positional encodings, stream interaction paradigms, and attention masking schemes across all three evaluation tiers.

	Full		Body		Hands	
Variant	R@1 ↑	MM Dist ↓	R@1 ↑	MM Dist ↓	R@1 ↑	MM Dist ↓
InterFlow	**0.479**	**2.969**	**0.495**	**2.993**	**0.449**	**3.051**
1D + person emb	0.459	3.116	0.468	3.230	0.426	3.152
Dual-stream + cross attn	0.467	2.994	0.483	3.092	0.437	3.080
Unrestricted attn	0.476	2.993	0.490	3.071	0.442	3.070

Table 3: **Initialization, context, runtime, and distillation.** (a) Launch probability/history dropout. (b) Remote/recent history frames. (c) Full Inter-X test-set generation time, using the same server and batch size. (d) One-step generation: Score distillation and Score + FK adversarial distillation.

(a) Initialization

Launch / dropout	Full R@1 ↑	Body FID ↓
0/off	0.438	0.461
0/on	0.462	0.348
0.25/on	0.479	0.264

(b) History

Remote / recent	Full R@1 ↑	Body FID ↓
0/15	0.458	0.340
5/10	0.474	0.287
10/5	0.479	0.264

(c) Runtime

Method	Time (s) ↓
Interact2Ar	550
InterFlow	97
FastInterFlow	15

(d) Distillation

Training	Full FID ↓	Hands FID ↓
None	0.741	0.752
Score	0.143	0.220
Score+Adv	0.139	0.195

Flow synthesizes each block in a single forward evaluation without guidance. Training schedules, optimization hyperparameters, and network architectures are detailed in Appendix D.

Compared methods and initialization. We compare against T2M (Guo et al., 2022), InterGen (Liang et al., 2024), InterMask (Javed et al., 2024), and Interact2Ar (Ruiz-Ponce et al., 2026). A superscript $*$ identifies methods initialized with the GT first frame. We distinguish this setting from generation without a GT motion seed when interpreting Table 1.

4.2 Comparison with Prior Methods

Interaction modeling with a shared stream. Across Full, Body, and Hands evaluations, InterFlow consistently achieves the highest retrieval accuracy (R-Precision) and lowest text–motion distance (MM Dist). Under matched GT first-frame conditioning ($*$), InterFlow* significantly outperforms the multi-branch Interact2Ar*, achieving lower FID on Full (0.192 vs. 0.277), Body (0.250 vs. 0.352), and Hands (0.164 vs. 0.257), while improving Top-1 R-Precision by 2.0–2.8 percentage points. Unseeded InterFlow likewise exceeds all unseeded baselines across evaluators. These results demonstrate that a unified single-stream transformer effectively captures complex whole-body coordination and articulate hand interactions without requiring dedicated per-person branches.

One-step generation and initialization. FastInterFlow achieves the overall lowest FID on Full (0.136) and Body (0.144), outperforming competitive multi-step baselines in a single network evaluation while bringing Diversity closest to real motion, accompanied by a minor trade-off in text alignment. Even without a GT seed, it further lowers Full and Body FID compared to InterFlow (0.139 vs. 0.208 and 0.154 vs. 0.264, respectively). Across both models, GT first-frame conditioning ($*$) consistently improves distributional realism (lower FID) while moderately reducing MModality due to constrained starting poses.

Inference efficiency. We benchmark inference efficiency against Interact2Ar, as it represents the state-of-the-art fixed-window autoregressive baseline and shares our sequential blockwise rollout protocol. Under matched hardware and batch configurations on the complete Inter-X test set (Table 3(c)), InterFlow and FastInterFlow achieve $5.7\times$ and $36.7\times$ speedups over Interact2Ar, respectively. Furthermore, one-step distillation provides an additional $6.47\times$ acceleration over the multi-step teacher, substantially mitigating rollout latency for interaction synthesis.

4.3 Ablation Studies

Architectural ablations. Table 2 evaluates key design choices across positional encodings, stream structures, and attention masking. Factorized 2D RoPE consistently outperforms 1D positions with additive person embeddings, verifying its effectiveness in modeling joint spatiotemporal coordinates. Crucially, InterFlow surpasses the dual-stream baseline coupled via mutual cross-attention,

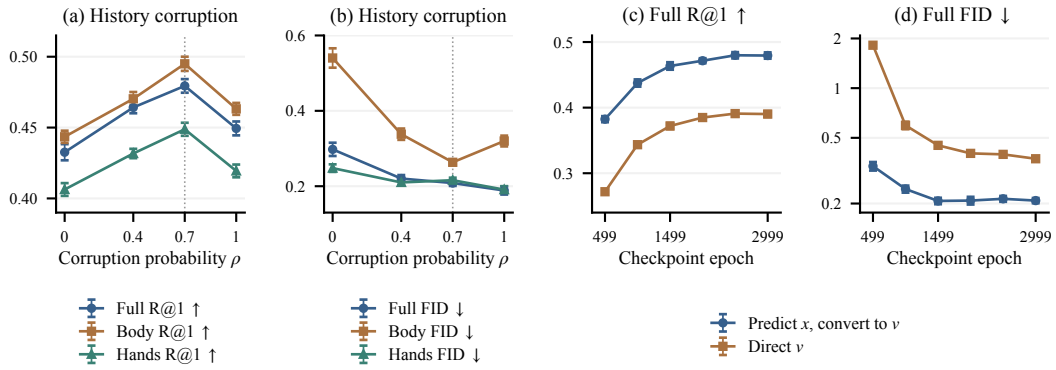


Figure 4: **History corruption and prediction parameterization.** (a,b) Diffusion-forcing probability versus R@1 and FID for Full, Body, and Hands. (c,d) Full-evaluator learning curves for endpoint prediction ($x \to v$) and direct velocity prediction; panel (d) uses a logarithmic FID axis.

demonstrating that standard self-attention over a unified token stream captures interpersonal dynamics directly, rendering handcrafted interaction modules unnecessary. Finally, our asymmetric attention mask matches unrestricted attention in generation quality, while preventing history from attending to the prediction block; this allows history key–value states to be computed once and cached during iterative ODE integration, saving approximately 30% inference compute.

Training for sequence initialization. Table 3(a) separates launch oversampling from history dropout. With launch oversampling disabled, enabling history dropout increases Full R@1 from 0.4382 to 0.4624 and reduces Body FID from 0.4607 to 0.3479. Adding launch oversampling at probability 0.25 further improves these values to 0.4794 and 0.2635. Full and Hands FID are not monotonic along this sequence of changes, as detailed in the Appendix H. These results support explicit exposure to sequence starts. The available experiments isolate dropout only when launch oversampling is absent.

Allocating a fixed history budget. At a fixed budget of 15 frames, increasing the remote allocation from 0 to 10 improves Full R@1 from 0.4576 to 0.4794 and lowers Body FID from 0.3396 to 0.2635 (Table 3(b)). However, the intermediate $5/10$ allocation gives lower Full and Hands FID than $10/5$ ($0.1997/0.1949$ versus $0.2084/0.2159$). Remote context therefore benefits alignment and Body FID in the tested configurations, while its allocation remains a trade-off. These aggregate metrics do not directly measure long-horizon consistency.

Diffusion forcing. Figure 4(a,b) varies the probability of corrupting history. Moving from $\rho = 0$ to 0.7 raises Full R@1 from 0.4326 to 0.4794 and reduces Body FID from 0.5402 to 0.2635. Corrupting history on every training sample ($\rho = 1$) instead gives lower Full and Hands FID, but weaker retrieval across all evaluators and higher Body FID. Among the tested probabilities, 0.7 provides the strongest retrieval scores and Body FID. These results support mixing clean and corrupted history for the selected operating point, rather than a monotonic benefit from increasing corruption.

Distillation objectives. Directly reducing InterFlow to one sampling step gives Full FID 0.7411 and Hands FID 0.7519. Score distillation reduces these to 0.1426 and 0.2195 (Table 3(d)), while Full R@1 increases from 0.4130 to 0.4454. Adding FK adversarial supervision yields Full/Body/Hands FID of $0.1394/0.1543/0.1948$, compared with $0.1426/0.1632/0.2195$ for score-only training. The largest relative FID change is on Hands (11.3% lower); Full and Body intervals overlap, and alignment remains close.

Endpoint versus velocity prediction. Figure 4(c,d) compares endpoint prediction followed by velocity conversion with direct velocity prediction at 6 training checkpoints. Both use the velocity training target and identical inference settings. Endpoint parameterization gives higher Full R@1 and lower Full FID at every evaluated checkpoint. At epoch 2999, the values are $0.4794/0.2084$ versus $0.3902/0.3735$, respectively. The advantage is visible early in training: endpoint prediction at epoch 999 already surpasses direct velocity prediction at epoch 2999 on both metrics. This supports more effective learning with endpoint parameterization; the evaluation curves alone do not identify high-frequency noise as its cause.

AI USE STATEMENT

In this work, we used generative AI tools for [assist with translation]. We have not used generative AI tools for [assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, interpret results], and [generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, clean and reformat dataset, support qualitative and thematic data analysis] are not applicable to this work. Additionally, we used generative AI tools for [suggest experimental parameters, create or edit software code, creation of draft parts of a research paper, summarize or analyse existing literature, sourcing/searching for information, edit a research paper to improve readability]. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure complete reproducibility of our framework and empirical results, we have documented extensive methodological, algorithmic, and experimental details throughout the manuscript and its appendix. Specifically, the network architectures, hyperparameter configurations, and motion representation settings are thoroughly elaborated in Appendix D. The complete algorithmic procedures for training and sequential blockwise sampling are outlined in Appendix E. Furthermore, the detailed theoretical derivations, objective formulations, and alternating optimization schedules for FastInterFlow's rollout score distillation and forward-kinematics adversarial supervision are provided in Appendix F. All evaluation protocols, multi-tier metric formulations, and joint coordinate definitions are rigorously detailed in Appendix G. To further facilitate reproducible research and support the community, we commit to releasing our complete codebase, including training scripts, distillation pipelines, and pretrained model checkpoints, upon publication.

REFERENCES

Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

German Barquero, Sergio Escalera, and Cristina Palmero. Seamless human motion composition with blended positional encodings. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 457–469. IEEE, 2024.

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 *International Conference on Learning Representations*, volume 2025, pp. 81935–81954, 2025.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024a.

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*, pp. 1–10, 2024b.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18000–18010. IEEE, 2023.

Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. In *European Conference on Computer Vision*, pp. 390–408. Springer, 2024.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 2978–2988, 2019.

Ke Fan, Junshu Tang, Weijian Cao, Ran Yi, Moran Li, Jingyu Gong, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. Freemotion: A unified framework for number-free text-to-motion synthesis. In *European Conference on Computer Vision*, pp. 93–109. Springer, 2024.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *Advances in Neural Information Processing Systems*, 38:75460–75482, 2026.

Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Remos: 3d motion-conditioned reaction synthesis for two-person interactions. In *European conference on computer vision*, pp. 418–437. Springer, 2024.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5142–5151. IEEE, 2022.

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

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026.

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.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023.

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

Haobo Li, Yanhong Zeng, Yunhong Lu, Jiapeng Zhu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Yujun Shen, and Zhipeng Zhang. Aad-1: Asymmetric adversarial distillation for one-step autoregressive video generation. *arXiv preprint arXiv:2606.03972*, 2026.

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*, 2024a.

Siyao Li, 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. In *International Conference on Learning Representations*, volume 2024, pp. 810–829, 2024b.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026.

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.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Huakun Liu, Qing Yu, Kent Fujiwara, Hideaki Uchiyama, and Kiyoshi Kiyokawa. Arms: Anchor-relational motion streaming for seamless solo-social motion transitions. *arXiv preprint arXiv:2607.05733*, 2026.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Cheng Luo, Bizhu Wu, Bing Li, Jianfeng Ren, Ruibin Bai, Rong Qu, Linlin Shen, and Bernard Ghanem. Reactmotion: Generating reactive listener motions from speaker utterance. *arXiv preprint arXiv:2603.15083*, 2026.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Zichong Meng, Zeyu Han, Xiaogang Peng, Yiming Xie, and Huaizu Jiang. Absolute coordinates make motion generation easy. *arXiv preprint arXiv:2505.19377*, 2025a.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27859–27871. IEEE, 2025b.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10975–10985, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

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*, pp. 8748–8763. PmLR, 2021.

Pablo Ruiz-Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and José García-Rodríguez. in2in: Leveraging individual information to generate human interactions. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1941–1951. IEEE, 2024.

Pablo Ruiz-Ponce, Sergio Escalera, José García-Rodríguez, Jiankang Deng, and Rolandos Alexandros Potamias. Interact2ar: Full-body human-human interaction generation via autoregressive diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23559–23569, 2026.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Kewei Sui, Anindita Ghosh, Inwoo Hwang, Bing Zhou, Jian Wang, and Chuan Guo. A survey on human interaction motion generation. *International Journal of Computer Vision*, 134(3):113, 2026.

Julian Tanke, Linguang Zhang, Amy Zhao, Chengcheng Tang, Yujun Cai, Lezi Wang, Po-Chen Wu, Juergen Gall, and Cem Keskin. Social diffusion: Long-term multiple human motion anticipation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9567–9577. IEEE, 2023.

Julian Tanke, Takashi Shibuya, Kengo Uchida, Koichi Saito, and Yuki Mitsufuji. Dyadic mamba: Long-term dyadic human motion synthesis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2859–2868. IEEE, 2025.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

Guy Tevet, Sigal Raab, Setareh Cohan, Daniele Reda, Zhengyi Luo, Xue Bin Peng, Amit Bermano, and Michiel Van de Panne. Closd: Closing the loop between simulation and diffusion for multi-task character control. In *International Conference on Learning Representations*, volume 2025, pp. 46506–46520, 2025.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36:8406–8441, 2023.

Zhenzhi Wang, Jingbo Wang, Yixuan Li, Dahua Lin, and Bo Dai. Intercontrol: Zero-shot human interaction generation by controlling every joint. *Advances in Neural Information Processing Systems*, 37:105397–105424, 2024.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Lixing Xiao, Shunlin Lu, Huaijin Pi, Ke Fan, Liang Pan, Yueer Zhou, Ziyong Feng, Xiaowei Zhou, Sida Peng, and Jingbo Wang. Motionstreamer: Streaming motion generation via diffusion-based autoregressive model in causal latent space. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10086–10096. IEEE, 2025.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22260–22271. IEEE, 2024.

Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14882–14894. IEEE, 2023.

Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, et al. Interact: Advancing large-scale versatile 3d human-object interaction generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7048–7060. IEEE, 2025a.

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*, pp. 1–11, 2025b.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. IEEE, 2024b.

Heng Yu, Juze Zhang, Changan Chen, Tiange Xiang, Yusu Fang, Juan Carlos Niebles, and Ehsan Adeli. Socialgen: Modeling multi-human social interaction with language models. In *2026 International Conference on 3D Vision (3DV)*, pp. 1–17. IEEE, 2026a.

Qing Yu, Akihisa Watanabe, and Kent Fujiwara. Causal motion diffusion models for autoregressive motion generation. *arXiv preprint arXiv:2602.22594*, 2026b.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14730–14740, 2023.

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*, 46(6):4115–4128, 2024.

Yan Zhang, Yao Feng, Alpár Cseke, Nitin Saini, Nathan Bajandas, Nicolas Heron, and Michael J Black. Primal: Physically reactive and interactive motor model for avatar learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12725–12736. IEEE, 2025.

Kaifeng Zhao, Gen Li, and Siyu Tang. Dartcontrol: A diffusion-based autoregressive motion model for real-time text-driven motion control. In *International Conference on Learning Representations*, volume 2025, pp. 23569–23592, 2025.

Kaifeng Zhao, Mathis Petrovich, Haotian Zhang, Tingwu Wang, Siyu Tang, and Davis Rempe. Ardy: Autoregressive diffusion with hybrid representation for interactive human motion generation. *arXiv preprint arXiv:2607.08741*, 2026.

Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. Emdm: Efficient motion diffusion model for fast and high-quality motion generation. In *European Conference on Computer Vision*, pp. 18–38. Springer, 2024.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5738–5746. IEEE, 2019.

Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2430–2449, 2023.

A Qualitative Visualizations

Since 3D human motion is best evaluated through video rather than static images, we provide video demonstrations on our anonymous project page:

`https://anonymous.4open.science/r/Visualization-of-InterFlow-and-FastInterFlow--C910/README.md`

The webpage includes videos of InterFlow and FastInterFlow, covering both successful generations and typical failure cases.

B Conclusion

In this paper, we revisited the foundational modeling of 3D human-human interaction and introduced **InterFlow**, a streamlined framework that eliminates the need for separate per-person branches or learned motion autoencoders. By representing interactants as a unified sequence of person-frame tokens, InterFlow demonstrates that a standard single-stream diffusion transformer equipped with 2D rotary embeddings can effectively capture deeply coupled interpersonal dynamics end-to-end within raw motion space. To overcome the rollout latency bottleneck, we further presented **FastInterFlow**, distilling the iterative teacher into a one-step blockwise generator via rollout-based score distillation and forward-kinematics adversarial supervision. Comprehensive experiments on the Inter-X benchmark demonstrate that our approach achieves state-of-the-art interaction fidelity while delivering substantial inference speedups. We hope this unified, efficient architecture serves as a solid baseline and inspires future research in real-time interactive motion synthesis.

C Limitations and Future Work

While InterFlow and FastInterFlow provide a unified and efficient framework for two-person interaction generation, several limitations remain to be addressed:

Evaluation Metric Misalignment and Teacher-Student Gap. Standard evaluation protocols (e.g., feature-space FID and R-Precision) predominantly reflect distribution statistics rather than fine-grained physical plausibility and subtle interactive coordination. Although the one-step FastInterFlow achieves competitive or even superior FID scores over the multi-step teacher, visual inspection reveals that its fidelity in nuanced coordination and delicate dynamics still lags behind InterFlow. Effectively mining ground-truth distributions and closing the perceptual gap under a single network evaluation remain an open challenge.

Physical Constraints and Adversarial Stability. While forward-kinematics adversarial distillation explicitly supervises joint-space realism, adversarial training remains inherently sensitive and requires delicate balancing against the score distillation objective. Integrating differentiable physics priors or lightweight collision-resolving modules into the one-step rollout is a promising avenue for future work.

Scalability to Multi-Person and Scene Interactions. The current person-frame tokenization and 2D RoPE coordinates are designed around dyadic ($p \in \{a, b\}$) interactions. Extending this single-stream paradigm to scalable multi-person groups ($N > 2$) or embodied interactions involving complex 3D environments warrants further exploration of dynamic token routing and scene context conditioning.

D Shared Representation and Implementation

InterFlow, the FastInterFlow student, and the fake-score model share the motion representation, generator backbone, and context interface described below. We first collect their common settings, then detail InterFlow training and sampling, FastInterFlow distillation, and the evaluation protocol. Discriminator-specific settings are given separately in Appendix F.2.

D.1 Notation and Shared Implementation Details

To ensure mathematical precision across the paper and visual figures, we explicitly distinguish the two decoupled temporal axes governing our framework: the discrete sequence frame index $i, j \in \mathbb{N}_0$ (with s marking the start frame of the target prediction block) along the physical motion timeline, and the continuous flow time $t, \tau_\ell \in [0, 1]$ parameterizing probability trajectories and history corruption levels. Throughout the manuscript, lowercase variables ($x_i^p \in \mathbb{R}^{336}$, $x_i \in \mathbb{R}^{672}$) designate single-frame poses, while uppercase symbols (X, H) denote spatiotemporal blocks spanning multiple frames. Superscripts $p \in \{a, b\}$ specify the interactant identity mapped to rotary coordinate $\pi(p) \in \{0, 1\}$, and subscripts indicate the corresponding time coordinate. Parameter sets θ, ϕ, ψ, and ω index the teacher model, the one-step student generator, the auxiliary fake-score estimator, and the kinematic discriminator, respectively. A complete summary of notations is provided in Table 4.

In Table 5, we summarize the shared generator backbone and context specifications. At each frame, the joint 672-dimensional two-person motion vector is projected into two separate 336-dimensional person-frame tokens, each comprising 55 continuous 6D joint rotations (330 channels) and 6 root-velocity channels. The underlying backbone is a shared single-stream Diffusion Transformer (DiT) parameterized by 8 layers with a hidden dimension of 512, 8 attention heads (head dimension 64), and SwiGLU feed-forward networks (hidden size 682, intermediate capacity 1024), totaling 41.7M trainable parameters. Queries and keys incorporate factorized 2D Rotary Position Embeddings (2D RoPE), allocating 32 channels to motion-frame indices to preserve original temporal intervals and 32 channels to interactant identities. During autoregressive rollouts, the context interface conditions on a fixed history budget of $K = 15$ frames ($K_{\text{rem}} = 10$ sparse remote frames and $K_{\text{rec}} = 5$ dense recent frames) to predict an $L = 30$ frame block. With two interactant tokens per frame, the maximum sequence length inside self-attention is strictly bounded to $2(K + L) = 90$ tokens, guaranteeing constant per-block computational complexity regardless of the total rollout duration.

Table 4: Motion-frame indices and flow times describe different axes. Uppercase X and H denote blocks, not single frames.

Symbol	Meaning
i, j; s	Motion-frame indices; first frame of the current prediction block
$p \in \{a, b\}$; $\pi(p)$	Person label; its numeric rotary coordinate
t; τ_ℓ	Continuous prediction flow time; history-slot corruption time
x_i^p; x_i	One person's clean frame; concatenated two-person frame
X_t; X_t^p	L-frame prediction block at flow time t; person slice
$x_i^p(t)$	One person–frame entry of X_t
H; H^p	Selected K-frame history; person slice
X_1^{gen}	Block generated by integrating InterFlow's field
$\widehat{X}_1$	Endpoint estimate used to parameterize InterFlow's field
X_1^{S}	FastInterFlow's one-step generated block
H^{roll}	Shared rollout history for student, teacher, and auxiliary field
$\theta, \phi, \psi, \omega$	Teacher, student, auxiliary score, and discriminator parameters
P, ΔP, F	FK joint positions, joint velocities, and discriminator features
Z	Hidden person–frame tokens inside the transformer

D.2 Generator Backbone and Context

Motion tokens and transformer blocks. The two-person representation has 672 features per frame and is unpacked into two 336-dimensional person vectors before the shared input projection. Outputs are projected back to the same motion representation. The model uses RMS normalization, SwiGLU feed-forward layers, and gated residual attention and feed-forward branches. The modulation projection and final motion readout are zero-initialized. There is no learned motion compression stage. Dataset feature normalization is disabled in this configuration.

Two-dimensional position encoding. In each attention head, half of the query/key channels encode motion-frame position and half encode person identity. Temporal rotations use the original frame coordinates supplied with the sampled context; person coordinates are 0 and 1. We omit ad-

Table 5: Shared generator-backbone and context settings. History counts refer to motion frames, each containing two person tokens. The token budget is the maximum before masking unavailable history.

Setting	Value
Persons / motion features per person	2 / 336
Transformer width / layers / attention heads	512 / 8 / 8
Head dimension	64
Configured feed-forward size / SwiGLU hidden width	1024 / 682
Transformer dropout	0.1
Prediction block L	30 frames
History budget K	15 frames
Remote / recent history K_{rem} / K_{rec}	10 / 5 frames
Maximum motion tokens $2(K+L)$	90
2D rotary dimensions per head: time / person	32 / 32
Rotary frequency base	10,000
Text feature dimension	768
Trainable Parameters	$41.7M$

ditional additive temporal and person embeddings. Shared person projections do not by themselves imply invariance to exchanging the two coordinate labels.

History construction and masking. Remote sampling selects one random frame per equal temporal segment, rather than using a fixed downsampling stride. The recent component contains the five immediately preceding frames. Both components preserve paired person observations. Invalid history entries are masked, and history queries cannot read the prediction block. The attention implementation also supplies an always-visible zero key/value entry to avoid an entirely masked key set when valid history is absent.

D.3 Common Data and Numerical Conventions

Both training stages use Inter-X text–motion pairs with motion crops of at most 150 frames and a configured batch size of 128. The effective global batch size depends on the distributed training setup. Prediction boundaries are sampled within crops. Weight decay is 0 in both stages, and generator EMA uses decay 0.999. Model-specific optimization and sampling settings are summarized in Table 6.

Text conditioning uses precomputed frozen CLIP features. The projected text feature is combined with each token's flow-time embedding. Valid clean history uses time 1, and prediction tokens share time t; InterFlow's corrupted history instead uses the assigned slot or group time. Flow time t is distinct from motion-frame index i throughout.

Shared endpoint-to-velocity conversion. For network parameters $\vartheta \in \{\theta, \phi, \psi\}$, the numerically guarded field is

$$\widehat{v}^{\delta}_{\vartheta}(X_t, t, H, c) = \frac{f_{\vartheta}(X_t, t, H, c) - X_t}{\max(1-t, \delta)}, \qquad \delta = 0.05. \tag{15}$$

This conversion is used whenever an endpoint network is queried as a velocity field. The one-step student instead directly outputs its endpoint at $t = 0$. The floor limits error amplification near $t = 1$ and agrees with the ideal conversion for $t \le 1 - \delta$. Ideal score identities below use the unguarded conversion; implementation losses use the regularized fields. Squared motion losses average over valid prediction entries, excluding masked padding.

History Dropout and Launch Oversampling. In autoregressive motion generation, the initial block synthesized at sequence onset ($s = 0$) anchors the entire rollout; any initial artifacts or unnatural postures will cascade and amplify across subsequent blocks. However, there is a fundamental train-test discrepancy at the launch phase: during test-time cold starts, the generator must synthesize the initial motion block conditioned purely on the text prompt c without any preceding motion

context. In contrast, standard training randomly samples the prediction boundary s uniformly across the entire sequence ($s \sim \mathcal{U}\{0, T_{\max} - L\}$), which suffers from two distinct challenges:

1. **Scarce 0-Start Exposure:** In a typical training crop of length $T_{\max} \approx 150$ with prediction block length $L = 30$, the uniform probability of hitting the exact global start $s = 0$ is merely $1/(T_{\max} - L + 1) \approx 0.8\%$. The network is thus overwhelmingly trained to condition on rich historical context, leaving its text-only launch capability under-trained.
2. **Context Incompleteness Near Onset:** When the sampled start frame is near the sequence beginning ($0 < s < K$), the available past frames are insufficient to satisfy the full history budget K. Simply padding the missing slots can cause the network to overfit to incomplete or sparse prefixes rather than learning to rely on text semantics.

To bridge this gap without introducing a separate launch network, we incorporate two complementary sequence initialization mechanisms:

1. DISTANCE-AWARE HISTORY DROPOUT. When the prediction start boundary s falls within the initial budget window ($s < K$), the generator has access to an incomplete history of length s. To progressively decouple the model from fragile initial contexts and force it to learn text-driven motion initiation, we apply an adaptive, linearly decaying history dropout:

$$P_{\text{drop}}(s) = \begin{cases} p_{\max} \cdot \left(1 - \frac{s}{K}\right), & \text{if } s < K, \\ 0, & \text{if } s \geq K, \end{cases} \tag{16}$$

where $p_{\max} \in [0, 1]$ is the maximum drop probability (set to 1.0 by default). When a dropout event occurs, the entire history token sequence H is masked out via key padding masks, compelling the single-stream transformer to synthesize the future block X purely from the text embedding c and flow timestep t. As the boundary s approaches K, $P_{\text{drop}}(s)$ decays smoothly to zero, naturally transitioning the model from pure launch to context-conditioned continuation.

2. COLD-START BOUNDARY OVERSAMPLING. To explicitly boost the exposure of the network to the text-only initial rollout phase, we introduce an oversampling strategy with probability $p_{\text{cold}} = 0.25$. Before sampling the prediction boundary for a training crop, the boundary selection rule is defined as:

$$s \sim \begin{cases} \delta(s - 0), & \text{with probability } p_{\text{cold}}, \\ \mathcal{U}\{0, T_{\max} - L\}, & \text{with probability } 1 - p_{\text{cold}}, \end{cases} \tag{17}$$

where $\delta(\cdot)$ denotes the Dirac delta distribution centering the start frame at index 0. Whenever $s = 0$ is sampled, history context is empty by definition and forced to be dropped, providing the network with frequent, high-gradient supervision on text-to-motion launch blocks.

Together, distance-aware dropout and cold-start oversampling enable a single shared transformer to seamlessly unify zero-context launch and multi-block autoregressive continuation within the same parameter space.

Table 6: Model-specific training and inference settings. Shared data, architecture, weight decay, and EMA settings are stated once above. InterFlow values describe the supplied configuration; sampling overrides used for a final evaluation must be recorded separately.

Setting	InterFlow	FastInterFlow
Training duration	3000 epochs	200 score-only + 100 with adversarial supervision
Prediction/score time	$\text{sigmoid}(z)$	$0.01 + 0.98\,\text{sigmoid}(z)$
Learning rate	10^{-4}	Student/fake score: 2×10^{-5}; discriminator: 10^{-6}
Gradient-norm clipping	1.0	0.5
History during training	GT with diffusion forcing	Bounded, detached student rollout
Launch oversampling	0.25	0.25
Inference per block	10 uniform Euler steps	One endpoint-network pass
Inference guidance scale	1.5	1 (no CFG)

Here, $z \sim \mathcal{N}(0, 1)$. Time sampling distributions are distinct from explicit loss weights.

E INTERFLOW TRAINING AND SAMPLING

Training objective and conditioning. InterFlow uses velocity MSE without auxiliary motion or FK regression losses. Training uses FP32, a three-epoch learning-rate warm-up, gradient clipping and learning rate as in Table 6, and EMA from the first update. Text features are replaced by a zero vector with probability 0.1 for classifier-free guidance training. First-frame anchoring, bootstrap generation, motion-peak weighting, and person-order augmentation are disabled in the supplied training configuration.

Diffusion forcing. History corruption is applied with probability $\rho = 0.7$. Remote slots receive independent corruption times drawn from p_H, whereas recent slots share one draw. Both persons share each slot's time. Corruption changes motion values while retaining original frame and person coordinates. Clean history uses time 1; corrupted slots carry their corresponding flow times. This provides imperfect training context without running the multi-step generator in rollout.

Numerical sampling. For $t_n = n/S$, $S = 10$, the sampling update is

$$X^{\text{gen}}_{t_{n+1}} = X^{\text{gen}}_{t_n} + \frac{1}{S} v^{\text{sample}}_\theta (X^{\text{gen}}_{t_n}, t_n, H, c). \tag{18}$$

Classifier-free guidance combines conditional and null-text fields using the scale in Table 6. Field evaluations occur at $t = 0, 0.1, \ldots, 0.9$, where the denominator floor in Equation equation 15 is inactive. After each block, generated frames are appended and history indices are selected again. The bounded-cost claim concerns model evaluation on a fixed token budget; storing and selecting from the full history is a separate implementation consideration. History key/value caching is optional and requires fixed history, positional coordinates, token times, and text conditions.

E.1 TRAINING AND GENERATION PROCEDURES

The procedures below omit padding and text dropout for readability. $\mathrm{History}$ returns paired person observations, their original frame indices, and validity masks. Field_θ applies the endpoint-to-velocity conversion in Equation equation 15; $\mathrm{SampleField}_\theta$ additionally applies the configured guidance rule.

Procedure 1: One InterFlow training update

1. Sample a text–motion pair and prediction boundary s; extract the clean target block X_1.
2. Set $H \leftarrow \mathrm{History}(x_{<s}; K_{\text{rem}}, K_{\text{rec}})$, retaining original frame coordinates. Apply configured prefix dropout.
3. Sample $z \sim \mathcal{N}(0, 1)$ and $\epsilon \sim \mathcal{N}(0, I)$; set $t \leftarrow \mathrm{sigmoid}(z)$ and $X_t \leftarrow (1 - t)\epsilon + tX_1$.
4. Initialize $\widetilde{H} \leftarrow H$ and every history time to 1.
5. With probability ρ, sample remote times independently from p_H and one recent time from p_H. For each valid slot, set $\widetilde{H}_\ell \leftarrow (1 - \tau_\ell)\eta_\ell + \tau_\ell H_\ell$, with independent Gaussian η_ℓ.
6. Compute $\widehat{v} \leftarrow \mathrm{Field}_\theta(X_t, t, \widetilde{H}, c)$ using person–time coordinates, history times, and the block mask.
7. Update θ with the valid-entry mean of $\|\widehat{v} - (X_1 - \epsilon)\|^2$; update the exponential moving average.

Procedure 2: Autoregressive InterFlow generation

1. Initialize the motion sequence with an optional seed and set the next prediction boundary s.
2. Until the requested sequence length is reached:
   (a) Build $H \leftarrow \mathrm{History}(x_{<s}; K_{\mathrm{rem}}, K_{\mathrm{rec}})$; assign valid history time 1 and retain its coordinates and mask.
   (b) Sample $X \sim \mathcal{N}(0, I)$ for both persons over the next L frames.
   (c) For $n = 0, \ldots, S-1$, set $t \leftarrow n/S$ and update $X \leftarrow X + S^{-1}\,\mathrm{SampleField}_\theta(X, t, H, c)$.
   (d) Append X to the sequence and advance $s \leftarrow s + L$.
3. Trim the final block to the requested length and return the sequence.

F FastInterFlow Distillation

The student and fake-score model inherit InterFlow's architecture and checkpoint weights. The teacher remains frozen. Both score fields use the same student rollout history and text. Score distillation is active throughout training; FK adversarial supervision is introduced after score training stabilizes.

F.1 Rollout-Based Score Distillation

We adapt score distillation to the linear flow-matching path in Equation equation 2. In this subsection, all fields are conditioned on the same detached rollout history H^{roll} and text c; these arguments are suppressed where unambiguous. The teacher parameters θ remain frozen. The auxiliary parameters ψ model the current student distribution, while ϕ parameterizes the one-step student.

Training the fake-score model. The teacher cannot estimate the changing student distribution by itself. We therefore train a separate fake-score model on freshly generated student blocks. Let $\overline{X}_1^{\mathrm{S}} = \mathrm{sg}[G_\phi(\xi, H^{\mathrm{roll}}, c)]$. We independently sample t and Gaussian noise ϵ, and construct $\overline{X}_t^{\mathrm{S}}$ using Equation equation 2 with endpoints ϵ and $\overline{X}_1^{\mathrm{S}}$. The auxiliary objective is

$$\mathcal{L}_{\mathrm{fake}}(\psi) = \mathbb{E}\left[\left\|\widehat{v}_\psi(\overline{X}_t^{\mathrm{S}}, t, H^{\mathrm{roll}}, c) - (\overline{X}_1^{\mathrm{S}} - \epsilon)\right\|_{\mathrm{valid}}^2\right]. \tag{19}$$

Only ψ is updated: both the student endpoint and rollout history are detached. The norm averages squared errors over valid prediction entries. This is flow matching on student samples, rather than regression to GT motion. The auxiliary network inherits the teacher's endpoint parameterization and converts its output to a velocity field.

From scores to velocity fields. For the ideal linear path, the conditional score associated with a velocity field satisfies

$$\mathcal{S}(X_t, t \mid H^{\mathrm{roll}}, c) = -\frac{\mathbb{E}[\epsilon \mid X_t, H^{\mathrm{roll}}, c]}{1-t} = \frac{t\,v(X_t, t, H^{\mathrm{roll}}, c) - X_t}{1-t}. \tag{20}$$

At the same re-noised student block, the X_t^{S} terms cancel when subtracting the teacher and fake scores. Consequently, for $0 < t < 1$,

$$\underbrace{\frac{1-t}{t}(\widehat{\mathcal{S}}_\theta - \widehat{\mathcal{S}}_\psi)}_{\text{score difference in velocity coordinates}} = \frac{1-t}{t}\frac{t}{1-t}(\widehat{v}_\theta - \widehat{v}_\psi) = \widehat{v}_\theta - \widehat{v}_\psi = \Delta v. \tag{21}$$

The reciprocal factors arise from changing parameterization and cancel algebraically. We compute Δv directly from the two velocity fields; no explicit score conversion or additional time-dependent loss multiplier is used. This specifies the velocity-space adaptation of score distillation, rather than asserting that an arbitrarily weighted score objective is unchanged by reparameterization.

Training the student. For the student update, X_1^{S} retains its gradient, and X_t^{S} is formed with Equation equation 2. Teacher and fake-score evaluations are fixed and detached. Using the direction from Equation equation 21, the final surrogate is

$$\mathcal{L}_{\mathrm{SD}}(\phi) = \mathbb{E}\left[\left\|X_t^{\mathrm{S}} - \mathrm{sg}[X_t^{\mathrm{S}} + \Delta v]\right\|_{\mathrm{valid}}^2\right], \tag{22}$$

with unit weight for every sampled t. The pseudo-target supplies a detached field direction, not a paired GT trajectory. In particular, its squared-error appearance does not make it supervised reconstruction of a specific motion sample.

For n_{valid} valid scalar entries in a sample, the induced gradient is

$$\nabla_\phi \mathcal{L}_{\mathrm{SD}} = \mathbb{E}\left[\frac{2t}{n_{\mathrm{valid}}}\left(\frac{\partial G_\phi}{\partial \phi}\right)^\top \left(\widehat{v}_\psi - \widehat{v}_\theta\right)_{\mathrm{valid}}\right]. \tag{23}$$

The remaining t comes from $\partial X_t^{\mathrm{S}}/\partial X_1^{\mathrm{S}} = tI$; it is not an explicit loss weight. Removing this chain-rule factor would change the implemented student update. Rollout histories are detached, so the objective does not differentiate through the full generated trajectory. The teacher and auxiliary fields are learned estimates, and the surrogate is not claimed to be an exact full-trajectory KL gradient.

Numerical implementation. Score queries and fake-score training use the time distribution in Table 6 and the shared denominator safeguard in Equation equation 15. Equations equation 20–equation 21 describe the ideal conversion; near the endpoint, the implementation compares regularized fields directly. The explicit score-distillation weight remains 1, with adaptive reweighting disabled.

F.2 Forward-Kinematics Adversarial Distillation

Once score-distillation training stabilizes, we introduce a discriminator to supervise joint-space motion quality without assigning a single GT trajectory to each noise sample. GT motion provides real samples, and student-generated motion provides fake samples.

Kinematic representation. Using the feature map in Equation equation 13, let

$$F^{\mathrm{GT}} = \Phi(\mathrm{FK}(X_1^{\mathrm{GT}})), \qquad F^{\mathrm{S}} = \Phi(\mathrm{FK}(X_1^{\mathrm{S}})), \qquad \Phi(P) = [P, \Delta P]. \tag{24}$$

Here, P contains world-space joint positions and $\Delta P_i = P_i - P_{i-1}$ denotes joint velocities in per-frame units. A 30-frame window contains both persons, with 55 joints per person. Position and velocity features give 330 input channels per person. The compact discriminator shares token projections across persons and uses 2D rotary self-attention, RMS normalization, SwiGLU feed-forward layers, mean token pooling, and a linear logit head. Its width, depth, and number of heads are 512, 3, and 8, respectively. It receives neither text nor flow time, and no preceding history is included in its window.

Discriminator update. Writing $\mathrm{sp}(z) = \log(1 + e^z)$, we use the non-saturating logistic discriminator objective

$$\mathcal{L}_D(\omega) = \mathbb{E}_{X_1^{\mathrm{GT}}}\left[\mathrm{sp}(-D_\omega(F^{\mathrm{GT}}))\right] + \mathbb{E}_{X_1^{\mathrm{S}}}\left[\mathrm{sp}(D_\omega(\mathrm{sg}[F^{\mathrm{S}}]))\right]. \tag{25}$$

Real windows are drawn independently from the training data. The student and fake-score model are fixed during this update; detaching F^{S} prevents discriminator training from updating the generator. Positive logits correspond to real motion.

Student adversarial update. With D_ω fixed, the student minimizes

$$\mathcal{L}_{\mathrm{adv}}(\phi) = \mathbb{E}\left[\mathrm{sp}(-D_\omega(\Phi(\mathrm{FK}(G_\phi(\xi, H^{\mathrm{roll}}, c)))))\right]. \tag{26}$$

The discriminator parameters are frozen, but its input derivative is retained. Gradients therefore pass through the discriminator, joint velocities, and differentiable FK to the student. This distribution-level supervision complements the score objective without a pointwise GT regression target. Score distillation remains active after adversarial supervision is introduced; only the student is needed at inference.

F.3 Training Schedule and Alternating Updates

Table 7 lists settings specific to distillation. Common settings are given in Appendix D.3; the discriminator architecture is specified in Appendix F.2.

Table 7: Distillation-specific optimization and rollout settings.

Setting	Value
Teacher score-query guidance scale	1 (conditional field)
Score-distillation weight / adaptive reweighting	1 / disabled
Rollout budget	1 to 3 blocks over 300 epochs
History gradients / frames outside rollout budget	Detached / GT
Adversarial activation	Epoch 201 (one-based)
Score scope / adversarial scope	All starts / $s \geq 30$
Student adversarial weight	0.003
Auxiliary / discriminator updates per student update	5 / 5 after activation
Optimizer / Adam moments	AdamW / $(0.9, 0.999)$

Rollout construction. A bounded number of preceding blocks are generated without gradients and replace the corresponding GT frames; older frames outside the rollout budget can remain GT. The same selected history, frame indices, and masks are used for both score queries. At inference, history instead comes from the model's own generated sequence. Launch oversampling and history dropout follow Appendix D.3.

Adversarial continuation and checkpoints. After 200 score-only epochs, training continues for 100 epochs with the score objective retained and adversarial supervision enabled. The continuation resumes the full training checkpoint, including optimizer and auxiliary-network states. The rollout schedule follows the global epoch and is not reset. The discriminator uses only positions and joint velocities; acceleration and hand-to-body descriptors are disabled. Real windows come from an independent shuffled data-loader draw, and fake windows come exclusively from the student. Student EMA weights are exported for one-pass evaluation; no auxiliary model or discriminator is used during generation.

Alternating update procedure.

1. Sample a text–motion crop and boundary s. Apply launch oversampling and history dropout, then construct detached bounded rollout history.
2. Generate a student block. For each auxiliary update, detach the endpoint, sample fresh t, ϵ, and update ψ with Equation equation 19.
3. After adversarial activation, draw independent real GT windows and detached student fake windows, and update ω with Equation equation 25.
4. Regenerate the current student block with gradients. Re-noise it and query both fixed score fields under the same rollout context. Update ϕ using Equation equation 22, including Equation equation 26 for eligible boundaries after adversarial activation. Update the student EMA.

Auxiliary and discriminator updates precede each student update at the ratios in Table 7; samples are regenerated on their respective turns. Only the network being optimized is updated.

G Evaluation Protocol and Metrics

Evaluation settings. Both models are evaluated on Inter-X with the same metric definitions. Access to a GT first frame is indicated by $*$ in the main-results table and is distinct from training-time launch oversampling. Unless explicitly marked, generation does not use a GT motion seed. The reported uncertainty summarizes 20 evaluation runs, not independent training seeds. Exact evaluator checkpoints and sampling counts must match across compared methods.

Evaluator inputs. The Full, Body, and Hands protocol uses separate pretrained text–motion evaluators (Ruiz-Ponce et al., 2026). Motion features are converted to world-space SMPL-X joints, including root orientation and translation. The two persons' coordinates are concatenated for each joint before flattening the frame. Table 8 specifies the joint subsets in the released evaluation implementation. Evaluators remain fixed, receive valid sequence lengths, and use joint coordinates without additional dataset standardization. These evaluators are separate from the generator's text-conditioning network and the FK discriminator.

Table 8: Input representations for the joint-space evaluators. Joint indices are zero-based SMPL-X indices; dimensions include both persons.

Evaluator	Joint indices	Channels/frame	Content
Full	0–54	330	All 55 joints per person
Body	0–21	132	22 body joints per person
Hands	25–54	180	30 finger joints per person

For one evaluator, let a_n and b_n denote the frozen text and motion embeddings for the nth paired example. All distances below are Euclidean distances in that evaluator's embedding space. We compute each metric separately for Full, Body, and Hands; their numerical scales are not interchangeable.

R-Precision and matching distance. Let $\mathcal{B}_b$ be evaluation batch b and $N = \sum_b |\mathcal{B}_b|$. Within each batch, form $d_{nm}^{(b)} = \|a_n^{(b)} - b_m^{(b)}\|_2$. Each text retrieves motions from its own batch, with the paired motion at the same local index. We compute

$$\mathrm{R@}k = \frac{1}{N} \sum_b \sum_{n=1}^{|\mathcal{B}_b|} \mathbf{1}\Big[n \in \mathrm{TopK}_{\min}(d_{n,:}^{(b)}, k)\Big], \tag{27}$$

$$\text{MM Dist} = \frac{1}{N} \sum_b \sum_{n=1}^{|\mathcal{B}_b|} d_{nn}^{(b)}. \tag{28}$$

We report $k \in \{1, 2, 3\}$ and accumulate counts and diagonal distances over all batches before normalization. Candidate-batch size must be held fixed because it controls retrieval difficulty. MM Dist uses unsquared distances between paired embeddings.

FID. We pool motion embeddings across evaluation batches and estimate generated and GT means and sample covariances, $(\mu_{\mathrm{gen}}, \Sigma_{\mathrm{gen}})$ and $(\mu_{\mathrm{GT}}, \Sigma_{\mathrm{GT}})$. The reported feature-space Fréchet distance is

$$\mathrm{FID} = \|\mu_{\mathrm{gen}} - \mu_{\mathrm{GT}}\|_2^2 + \mathrm{Tr}\Big(\Sigma_{\mathrm{gen}} + \Sigma_{\mathrm{GT}} - 2(\Sigma_{\mathrm{gen}}\Sigma_{\mathrm{GT}})^{1/2}\Big). \tag{29}$$

The covariance estimate uses the $N - 1$ denominator. The reference implementation computes the matrix square root numerically; if it is non-finite, it adds $10^{-6}I$ to both covariances before recomputing it. Negligible imaginary residuals are discarded. This FID is computed on learned motion embeddings, not image features or raw joint coordinates.

Diversity and multimodality. Diversity averages Q_{div} sampled pairwise distances across generated motion embeddings. Multimodality averages distances between repeated generations conditioned on the same caption:

$$\text{Diversity} = \frac{1}{Q_{\mathrm{div}}} \sum_{q=1}^{Q_{\mathrm{div}}} \|b_{u_q} - b_{v_q}\|_2, \tag{30}$$

$$\text{MModality} = \frac{1}{C_{\mathrm{mm}} Q_{\mathrm{mm}}} \sum_{n=1}^{C_{\mathrm{mm}}} \sum_{q=1}^{Q_{\mathrm{mm}}} \|b_{n,u_q} - b_{n,v_q}\|_2. \tag{31}$$

Here, C_{mm} captions each have R_{mm} generated repeats, and $b_{n,r}$ denotes repeat r for caption n. Each side's indices are sampled without replacement, independently of the other side; corresponding indices can therefore coincide. Multimodality uses repeat indices within the same caption, whereas

Diversity samples across the generated set. Diversity is ranked by absolute distance to its GT reference. Higher MModality indicates more conditional variation, but does not alone establish fidelity or text alignment.

Repeated evaluation. For a metric value q_r from evaluation run r, we report

$$\overline{q} \pm 1.96\frac{\sigma_q}{\sqrt{R}}, \qquad \overline{q} = \frac{1}{R}\sum_{r=1}^{R} q_r, \qquad \sigma_q^2 = \frac{1}{R}\sum_{r=1}^{R}(q_r - \overline{q})^2, \qquad R = 20. \tag{32}$$

This uses the reference implementation's normal-approximation interval and population-standard-deviation convention across evaluation repeats. Metrics are first computed separately per run, then summarized; embeddings are not pooled across repeats before FID computation. These intervals quantify evaluation variability, not variation across independently trained models.

H Additional Ablation Results

All ablation configurations use 20 evaluation runs, 10 uniform sampling steps, CFG scale 1.5, and no GT first frame. Reference settings are $p_{\text{launch}} = 0.25$, diffusion-forcing probability $\rho = 0.7$, and remote/recent history $10/5$. Each named variant changes the stated component; removing history dropout is evaluated only with launch oversampling disabled.

Tables below provide additional metrics for attention masking, sequence initialization, history allocation, and diffusion forcing, including Top-2/3 retrieval. The intermediate launch probability 0.15 is retained here. Values are means with reported 95% confidence intervals. Divergent GT evaluations are preserved per configuration rather than replaced by the reference run's GT. We report $\Delta_{\text{Div}} = |\text{Div} - \text{Div}_{\text{GT}}|$ as a descriptive difference of means; its uncertainty cannot be recovered from the marginal intervals alone. A GT MModality reference is unavailable, so MModality is reported without a directional ranking. Evaluation-repeat intervals do not measure variability across independent training seeds.

Table 9: Full evaluator: text–motion alignment.

Variant	R@1 ↑	R@2 ↑	R@3 ↑	MM Dist ↓
Reference	$0.4794_{\pm 0.0048}$	$0.6886_{\pm 0.0045}$	$0.7990_{\pm 0.0042}$	$2.9693_{\pm 0.0177}$
Unrestricted attn.	$0.4758_{\pm 0.0055}$	$0.6851_{\pm 0.0049}$	$0.7941_{\pm 0.0037}$	$2.9926_{\pm 0.0137}$
Launch 0; drop off	$0.4382_{\pm 0.0057}$	$0.6392_{\pm 0.0046}$	$0.7503_{\pm 0.0039}$	$3.2523_{\pm 0.0162}$
Launch 0; drop on	$0.4624_{\pm 0.0046}$	$0.6717_{\pm 0.0052}$	$0.7819_{\pm 0.0038}$	$3.0694_{\pm 0.0159}$
Launch 0.15; drop on	$0.4822_{\pm 0.0046}$	$0.6956_{\pm 0.0042}$	$0.8019_{\pm 0.0039}$	$2.9348_{\pm 0.0143}$
DF 0.0	$0.4326_{\pm 0.0057}$	$0.6395_{\pm 0.0062}$	$0.7507_{\pm 0.0046}$	$3.2532_{\pm 0.0202}$
DF 0.4	$0.4642_{\pm 0.0041}$	$0.6697_{\pm 0.0037}$	$0.7815_{\pm 0.0047}$	$3.0668_{\pm 0.0170}$
DF 1.0	$0.4494_{\pm 0.0049}$	$0.6546_{\pm 0.0043}$	$0.7656_{\pm 0.0032}$	$3.1632_{\pm 0.0146}$
Remote/recent 0/15	$0.4576_{\pm 0.0055}$	$0.6628_{\pm 0.0044}$	$0.7749_{\pm 0.0038}$	$3.0991_{\pm 0.0163}$
Remote/recent 5/10	$0.4745_{\pm 0.0041}$	$0.6803_{\pm 0.0049}$	$0.7881_{\pm 0.0041}$	$3.0140_{\pm 0.0194}$

Table 10: Full evaluator: distribution metrics. Δ_{Div} uses the GT mean from the same run.

Variant	FID ↓	Div	GT Div	Δ_{Div} ↓	MMod
Reference	$0.2084_{\pm 0.0076}$	$9.4997_{\pm 0.0925}$	$9.1067_{\pm 0.0795}$	0.3930	$1.8228_{\pm 0.0805}$
Unrestricted attn.	$0.2005_{\pm 0.0077}$	$9.4607_{\pm 0.0716}$	$9.1067_{\pm 0.0795}$	0.3540	$2.0560_{\pm 0.0763}$
Launch 0; drop off	$0.2192_{\pm 0.0129}$	$9.3642_{\pm 0.0867}$	$9.1067_{\pm 0.0795}$	0.2575	$2.8752_{\pm 0.1028}$
Launch 0; drop on	$0.2272_{\pm 0.0095}$	$9.5132_{\pm 0.0764}$	$9.1067_{\pm 0.0795}$	0.4065	$2.3250_{\pm 0.0885}$
Launch 0.15; drop on	$0.2132_{\pm 0.0078}$	$9.4725_{\pm 0.0864}$	$9.1067_{\pm 0.0795}$	0.3658	$1.9087_{\pm 0.0771}$
DF 0.0	$0.2978_{\pm 0.0176}$	$9.3598_{\pm 0.0863}$	$9.1067_{\pm 0.0795}$	0.2531	$2.4317_{\pm 0.0938}$
DF 0.4	$0.2203_{\pm 0.0093}$	$9.4175_{\pm 0.0902}$	$9.1067_{\pm 0.0795}$	0.3108	$2.1029_{\pm 0.0712}$
DF 1.0	$0.1887_{\pm 0.0108}$	$9.3271_{\pm 0.0792}$	$9.1067_{\pm 0.0795}$	0.2204	$2.1751_{\pm 0.0815}$
Remote/recent 0/15	$0.2287_{\pm 0.0090}$	$9.3758_{\pm 0.0843}$	$9.1067_{\pm 0.0795}$	0.2691	$2.1062_{\pm 0.0726}$
Remote/recent 5/10	$0.1997_{\pm 0.0118}$	$9.4303_{\pm 0.0653}$	$9.1067_{\pm 0.0795}$	0.3236	$2.0722_{\pm 0.0747}$

Table 11: Body evaluator: text–motion alignment.

Variant	R@1 ↑	R@2 ↑	R@3 ↑	MM Dist ↓
Reference	$0.4950_{\pm 0.0050}$	$0.7018_{\pm 0.0044}$	$0.8022_{\pm 0.0040}$	$2.9933_{\pm 0.0167}$
Unrestricted attn.	$0.4898_{\pm 0.0038}$	$0.6893_{\pm 0.0049}$	$0.7908_{\pm 0.0039}$	$3.0705_{\pm 0.0144}$
Launch 0; drop off	$0.4310_{\pm 0.0052}$	$0.6292_{\pm 0.0052}$	$0.7389_{\pm 0.0047}$	$3.4297_{\pm 0.0212}$
Launch 0; drop on	$0.4763_{\pm 0.0045}$	$0.6768_{\pm 0.0031}$	$0.7816_{\pm 0.0044}$	$3.1340_{\pm 0.0153}$
Launch 0.15; drop on	$0.4955_{\pm 0.0046}$	$0.6987_{\pm 0.0046}$	$0.8019_{\pm 0.0035}$	$3.0117_{\pm 0.0173}$
DF 0.0	$0.4435_{\pm 0.0043}$	$0.6404_{\pm 0.0054}$	$0.7448_{\pm 0.0040}$	$3.3971_{\pm 0.0183}$
DF 0.4	$0.4703_{\pm 0.0048}$	$0.6664_{\pm 0.0049}$	$0.7732_{\pm 0.0047}$	$3.1633_{\pm 0.0153}$
DF 1.0	$0.4632_{\pm 0.0042}$	$0.6608_{\pm 0.0041}$	$0.7634_{\pm 0.0047}$	$3.2559_{\pm 0.0177}$
Remote/recent 0/15	$0.4679_{\pm 0.0048}$	$0.6718_{\pm 0.0037}$	$0.7754_{\pm 0.0039}$	$3.1874_{\pm 0.0187}$
Remote/recent 5/10	$0.4795_{\pm 0.0037}$	$0.6808_{\pm 0.0053}$	$0.7846_{\pm 0.0051}$	$3.1055_{\pm 0.0181}$

Table 12: Body evaluator: distribution metrics. Δ_{Div} uses the GT mean from the same run.

Variant	FID ↓	Div	GT Div	Δ_{Div} ↓	MMod
Reference	$0.2635_{\pm 0.0078}$	$9.4085_{\pm 0.0954}$	$9.0211_{\pm 0.0807}$	0.3874	$1.8965_{\pm 0.0610}$
Unrestricted attn.	$0.3068_{\pm 0.0141}$	$9.3922_{\pm 0.0931}$	$9.0211_{\pm 0.0807}$	0.3711	$2.1154_{\pm 0.0610}$
Launch 0; drop off	$0.4607_{\pm 0.0255}$	$9.2784_{\pm 0.0764}$	$9.0211_{\pm 0.0807}$	0.2573	$3.0483_{\pm 0.0896}$
Launch 0; drop on	$0.3479_{\pm 0.0145}$	$9.4318_{\pm 0.0907}$	$9.0211_{\pm 0.0807}$	0.4107	$2.3664_{\pm 0.0693}$
Launch 0.15; drop on	$0.2714_{\pm 0.0122}$	$9.3823_{\pm 0.0897}$	$9.0211_{\pm 0.0807}$	0.3612	$1.9715_{\pm 0.0703}$
DF 0.0	$0.5402_{\pm 0.0257}$	$9.2677_{\pm 0.0931}$	$9.0211_{\pm 0.0807}$	0.2466	$2.5722_{\pm 0.0823}$
DF 0.4	$0.3380_{\pm 0.0150}$	$9.3133_{\pm 0.0862}$	$9.0211_{\pm 0.0807}$	0.2922	$2.2401_{\pm 0.0704}$
DF 1.0	$0.3199_{\pm 0.0143}$	$9.2126_{\pm 0.0937}$	$9.0211_{\pm 0.0807}$	0.1915	$2.2803_{\pm 0.0763}$
Remote/recent 0/15	$0.3396_{\pm 0.0127}$	$9.2686_{\pm 0.0892}$	$9.0211_{\pm 0.0807}$	0.2475	$2.2080_{\pm 0.0750}$
Remote/recent 5/10	$0.2875_{\pm 0.0195}$	$9.3092_{\pm 0.1019}$	$9.0211_{\pm 0.0807}$	0.2881	$2.1528_{\pm 0.0664}$

Table 13: Hands evaluator: text–motion alignment.

Variant	R@1 ↑	R@2 ↑	R@3 ↑	MM Dist ↓
Reference	$0.4488_{\pm 0.0046}$	$0.6544_{\pm 0.0043}$	$0.7650_{\pm 0.0045}$	$3.0509_{\pm 0.0154}$
Unrestricted attn.	$0.4416_{\pm 0.0042}$	$0.6474_{\pm 0.0039}$	$0.7585_{\pm 0.0036}$	$3.0700_{\pm 0.0123}$
Launch 0; drop off	$0.3991_{\pm 0.0048}$	$0.5960_{\pm 0.0041}$	$0.7137_{\pm 0.0044}$	$3.3328_{\pm 0.0139}$
Launch 0; drop on	$0.4226_{\pm 0.0057}$	$0.6286_{\pm 0.0056}$	$0.7457_{\pm 0.0043}$	$3.1685_{\pm 0.0183}$
Launch 0.15; drop on	$0.4459_{\pm 0.0037}$	$0.6532_{\pm 0.0046}$	$0.7663_{\pm 0.0038}$	$3.0473_{\pm 0.0137}$
DF 0.0	$0.4063_{\pm 0.0045}$	$0.6058_{\pm 0.0042}$	$0.7202_{\pm 0.0036}$	$3.2817_{\pm 0.0174}$
DF 0.4	$0.4317_{\pm 0.0033}$	$0.6320_{\pm 0.0037}$	$0.7464_{\pm 0.0044}$	$3.1364_{\pm 0.0136}$
DF 1.0	$0.4194_{\pm 0.0045}$	$0.6196_{\pm 0.0034}$	$0.7354_{\pm 0.0032}$	$3.2216_{\pm 0.0111}$
Remote/recent 0/15	$0.4298_{\pm 0.0047}$	$0.6273_{\pm 0.0038}$	$0.7382_{\pm 0.0039}$	$3.1697_{\pm 0.0132}$
Remote/recent 5/10	$0.4349_{\pm 0.0038}$	$0.6398_{\pm 0.0036}$	$0.7537_{\pm 0.0031}$	$3.1047_{\pm 0.0160}$

Table 14: Hands evaluator: distribution metrics. Δ_{Div} uses the GT mean from the same run.

Variant	FID ↓	Div	GT Div	Δ_{Div} ↓	MMod
Reference	$0.2159_{\pm 0.0067}$	$8.8575_{\pm 0.0884}$	$8.4624_{\pm 0.0736}$	0.3951	$1.9485_{\pm 0.0712}$
Unrestricted attn.	$0.2148_{\pm 0.0082}$	$8.7937_{\pm 0.0824}$	$8.4624_{\pm 0.0736}$	0.3313	$2.1578_{\pm 0.0582}$
Launch 0; drop off	$0.2181_{\pm 0.0103}$	$8.7081_{\pm 0.0606}$	$8.4624_{\pm 0.0736}$	0.2457	$3.0063_{\pm 0.0943}$
Launch 0; drop on	$0.2335_{\pm 0.0075}$	$8.8051_{\pm 0.0733}$	$8.4624_{\pm 0.0736}$	0.3427	$2.4549_{\pm 0.0861}$
Launch 0.15; drop on	$0.2223_{\pm 0.0071}$	$8.8085_{\pm 0.0770}$	$8.4624_{\pm 0.0736}$	0.3461	$2.0356_{\pm 0.0662}$
DF 0.0	$0.2480_{\pm 0.0100}$	$8.7140_{\pm 0.0746}$	$8.4624_{\pm 0.0736}$	0.2516	$2.5276_{\pm 0.0927}$
DF 0.4	$0.2100_{\pm 0.0074}$	$8.8176_{\pm 0.0673}$	$8.4624_{\pm 0.0736}$	0.3552	$2.2681_{\pm 0.0717}$
DF 1.0	$0.1909_{\pm 0.0092}$	$8.7285_{\pm 0.0803}$	$8.4624_{\pm 0.0736}$	0.2661	$2.2998_{\pm 0.0725}$
Remote/recent 0/15	$0.2133_{\pm 0.0090}$	$8.7627_{\pm 0.0917}$	$8.4624_{\pm 0.0736}$	0.3003	$2.2558_{\pm 0.0692}$
Remote/recent 5/10	$0.1949_{\pm 0.0091}$	$8.7753_{\pm 0.0794}$	$8.4624_{\pm 0.0736}$	0.3129	$2.2189_{\pm 0.0705}$