Unified Number-Free Text-to-Motion Generation Via Flow Matching

Guanhe Huang
King's College London
guanhe.huang@kcl.ac.uk

Oya Celiktutan
King's College London
oya.celiktutan@kcl.ac.uk

Abstract

Generative models excel at motion synthesis for a fixed number of agents but struggle to generalize with variable agents. Based on limited, domain-specific data, existing methods employ autoregressive models to generate motion recursively, which suffer from inefficiency and error accumulation. We propose ***Unified Motion Flow (UMF)****, which consists of Pyramid Motion Flow (P-Flow) and Semi-Noise Motion Flow (S-Flow). UMF decomposes the number-free motion generation into a single-pass motion prior generation stage and multi-pass reaction generation stages. Specifically, UMF utilizes a unified latent space to bridge the distribution gap between heterogeneous motion datasets, enabling effective unified training. For motion prior generation, P-Flow operates on hierarchical resolutions conditioned on different noise levels, thereby mitigating computational overheads. For reaction generation, S-Flow learns a joint probabilistic path that adaptively performs reaction transformation and context reconstruction, alleviating error accumulation. Extensive results and user studies demonstrate UMF's effectiveness as a generalist model for multi-person motion generation from text. Project page:* https://githubhgh.github.io/umf/.

1. Introduction

Text-to-motion generation, particularly via diffusion models, has advanced rapidly, progressing from single-agent [13, 14, 16, 51, 56] to multi-agent [11, 32, 45, 47, 58, 61] synthesis. However, how to synthesize realistic number-free (*i.e.*, any arbitrary number) human motions with text prompts remains an open challenge. Existing methods struggle to generalize to unseen crowded scenes and are limited by motion data scarcity. These limitations hinder the applications in robotics [23, 35] and virtual reality [19, 63], which often require seamless transitions between independent and collaborative tasks. This gap highlights the need for methods that can effectively utilize available heterogeneous data [12, 46].

To address the problem of text-to-motion generation with varying number of agents, previous methods typically rely on tailored architectures, more specifically, requiring expensive and time-consuming datasets [13, 32] for specific motion generation tasks. Critically, existing multi-person interaction datasets [32, 61] are smaller and less diverse compared to single-person datasets [13, 21, 39], despite the interactive tasks being more complex. On the other hand, there is a significant overlap for basic movements (*e.g.*, walking) across these heterogeneous datasets, suggesting that *single-person motion data can serve as heterogeneous prior for interaction synthesis*.

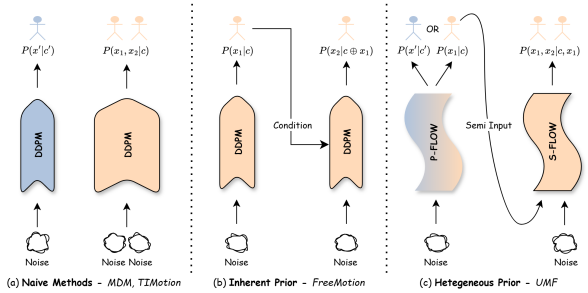

Figure 1. Core contribution of UMF. We show dual-agent cases here for simplicity. (a) Standard methods [51, 57] are restricted to a fixed number of agents. (b) Autoregressive methods [12] decouple generation into a motion prior and subsequent reaction. The reaction is typically guided by the prior using a conditioning network. (c) Our UMF leverages a heterogeneous motion prior as the adaptive start point of the reaction flow path, mitigating error accumulation.

To leverage this overlap, in this paper, we introduce a single-person multi-token tokenizer that supports unified modeling and establishes the foundation for number-free, text-conditional generation. Compared to the noisy raw motion space, the regularized multi-token latent space stabilizes flow matching training on heterogeneous single-agent (*i.e.*, HumanML3D [13]) and multi-agent (*i.e.*, InterHuman [32]) datasets. Based on this latent space, we propose Unified Motion Flow (UMF), a framework for number-free human motion generation from text prompts. UMF features two modules, the Pyramid Motion Flow (P-Flow) and Semi-Noise Motion Flow (S-Flow), which utilize flow matching

to learn the mapping between text, motion prior, and reaction. Specifically, it decouples number-free generation into a single-pass motion prior initialization (P-Flow) and a subsequent multi-pass reaction transformation (S-Flow).

Compared to previous single-token methods [7, 9], our multi-token latent space shows superior reconstruction performance, mitigating heterogeneous domain gaps. However, it also imposes greater computational overhead. Inspired by the fact that samples in early timesteps are noisy and less informative [29, 55], we introduce the P-Flow, which decomposes the motion prior generation into continuous hierarchical stages based on the timestep (noise level). Specifically, P-Flow maintains the original resolution only at later timesteps and applies a lower resolution via downsampling for early stages. Previous works [28, 50, 60] that employ cascade models for these different resolutions are still accompanied by extra model complexity. In contrast, our P-Flow can handle different resolutions within a single transformer [52], improving efficiency for multi-token motion prior generation.

The motion prior generated by P-Flow serves as the input for the iterative synthesis of subsequent agent reactions. However, this autoregressive process often suffers from potential error accumulation [24, 54]. Previous methods [12] rely on deterministic condition mechanisms (*e.g.*, ControlNet [66]) to guide the process, which struggle to capture the causal relationship between interactive agents. Consequently, we propose Semi-Noise Motion Flow (S-Flow) to learn the joint probabilistic path between previously generated motions (the context) and the subsequent agent's motion (the reaction). As shown in Fig. 1, rather than using the generated motions as a static condition, S-Flow integrates them to define the context distribution. This source distribution initializes the reaction generation path, which enables S-Flow to focus directly on learning the dynamic transformation between motion distributions. Concurrently, the S-Flow learns another auxiliary path to reconstruct the integrated context from noise distributions, as a strong regularizer for global interactive dependencies. This joint training of two distinct flow paths balances between the reaction prediction and context awareness, making it less prone to error accumulation.

In summary, our contributions are as follows:

- We propose Unified Motion Flow (UMF), a generalist framework for number-free text-to-motion generation. UMF's core design unifies heterogeneous single-person (e.g., HumanML3D) and multi-person (e.g., InterHuman) datasets within a multi-token latent space.
- For efficient individual motion synthesis, we introduce Pyramid Motion Flow (P-Flow). P-Flow operates on hierarchical resolutions conditioned on the noise level, which alleviates computational overheads of multi-token representations while maintaining high-fidelity generation.
- For reaction and interaction synthesis, we develop Semi-Noise Motion Flow (S-Flow). S-Flow learns a joint probabilistic path by balancing reaction transformation and context reconstruction, thereby alleviating error accumulation.
- Extensive experiments demonstrate UMF achieves state-of-the-art (SoTA) performance for multi-person generation (FID 4.772 on InterHuman) benchmarks. We also validate UMF's zero-shot generalization to unseen group scenarios through a user study.

2. Related Work

2.1. Text-conditioned Human Motion Synthesis

Generative models have shown promising results on human motion synthesis [7, 9, 14, 51, 53, 67, 68], though most works focus on single-agent or dual-agent scenarios. Most recently, MaskControl [42] introduces accurate single-person controllability to the generative masked motion model [14], while maintaining high-quality generation. Dual-agent motion synthesis has also seen rapid advancements [32, 43, 61]. Ma et al. [38] employs an interleaved learning strategy to capture the dynamic interactions and nuanced coordination, exhibiting higher text-to-motion alignment, and improved diversity. Wang et al. [57] subsequently introduces TIMotion, a parameter-efficient approach utilizing temporal modeling and interaction mixing. Synthesizing human-like reactions [48] is another active area of research. Xu et al. [62] establishes one of the earliest multi-setting benchmarks for this task, supported by three dedicated annotated datasets. Similar to us, Jiang et al. [27] propose direct noise-free action-to-reaction mappings through flow matching, while they ignore the error accumulation for autoregressive multi-person generation.

2.2. Unified Motion Synthesis

The recent success of Large Language Models [1, 2, 6, 15], particularly their strong generative and zero-shot transfer capabilities, has inspired new generalist approaches in motion synthesis. Research in unified motion generation has focused on several aspects, including: 1) unifying generation with understanding [25, 70], 2) integrating diverse input modalities [31, 40], and 3) handling a variable number of actors [12, 17, 69]. An early work [25] proposed MotionGPT to address diverse motion-relevant tasks, which treats human motion as a foreign language to unify tasks like motion generation and understanding. Then, Petrov et al. [40] proposed TriDi for human-object interaction, a unified model capturing the joint 3D distribution of humans, objects, and their interactions. To unify motion generation across different conditioning modalities (*e.g.*, text, video), Li et al. [31] introduced GENMO, a generalist model conditioned on videos, music, text, 2D keypoints, and 3D keyframes. [17] introduced dualFlow, a flow-based model for interactive and reactive text-to-motion, though

it is limited to dual-agent scenarios. Most related to our work, FreeMotion [12] proposes a decoupled generation and interaction module for number-free motion generation, while it suffers from inefficiency and error accumulation in multi-person scenarios. Recently, Zhao et al. [69] proposed FreeDance, a unified, number-free music-to-motion framework based on masked modeling of 2D discrete tokens, whereas our UMF focuses on the text-to-motion task.

3. Preliminaries

Flow Matching. Flow generative models [3, 33, 34] aim to learn a velocity field v_t that maps source distribution $x_0 \sim p$ to target distribution $x_1 \sim q$ via an ordinary differential equation (ODE):

$$\frac{dx_t}{dt} = v_t(x_t). \tag{1}$$

Recently, Lipman et al. [33] proposed the flow matching framework, which offers a simulation-free training objective by directly regressing the model's velocity field v_t on a conditional vector field $u_t(\cdot|x_1)$:

$$\mathbb{E}_{t,q(x_1),p_t(x_t|x_1)} \left\| v_t(x_t) - u_t(x_t|x_1) \right\|^2, \tag{2}$$

where $u_t(\cdot|x_1)$ uniquely determines a conditional probability path $p_t(\cdot|x_1)$ toward data sample x_1. An effective choice of the conditional probability path is linear interpolation [37] of data and noise:

$$x_t = tx_1 + (1-t)x_0, \tag{3}$$

$$x_t \sim \mathcal{N}(tx_1, (1-t)^2 I), \tag{4}$$

and $u(x_t|x_1) = x_1 - x_0$. Notably, flow matching can be flexibly extended to interpolate between distributions other than Gaussians. This enables us to employ the flow matching for both motion prior and reaction generation.

4. Proposed Method

4.1. Unified Latent Space

A key challenge in building a generalist motion model is that generative frameworks like flow matching require a consistent data format, a condition not met by heterogeneous motion datasets. For instance, individual motion datasets [13] often use canonical representations, while interaction datasets [32] use non-canonical representations. To bridge this gap, we first convert individual motions to a unified non-canonical SMPL skeleton representation with 22 joints. Then we split the interaction sample into multiple individual motion sequences (see Appendix A for details).

As shown in Fig. 2(A), the single motion tokenizer learns a continuous latent space for individual motion sequences. Similar to TEMOS [41], we utilize transformers [52] as the encoder and decoder, enhanced with skip connections and layer norms. The individual encoder takes an individual motion sequence $x_I^{1:N} \in \mathbb{R}^{N\times D}$ as input and compresses it into a latent representation $z \in \mathbb{R}^{p\times r}$. Using the reparameterization trick [30], we sample a latent vector $z \in \mathbb{R}^{p\times r}$ from the learned Gaussian distribution. Then, the individual decoder reconstructs the latent vector z into motion sequences $\hat{x}_I^{1:N}$. Different from existing number-free methods [12] that are trained on raw motion space, which suffer from performance degradation on heterogeneous datasets, our multi-token latent space shows better stability.

Multiple latent tokens. Previous latent motion diffusion works [7, 70] employ *single latent token* learning (*e.g.* 1×256), imposing a bottleneck on the VAE's reconstruction performance. While naively increasing the number of tokens can improve reconstruction, it often degrades the generative performance [65]. Inspired by Dai et al. [8], we utilize a latent adapter to decouple the internal token representation from the final latent dimension. The VAE encoder first captures complex motion details using a larger token (e.g., 16×256) and then projects them to a compact, semantically dense space (e.g., 16×32) for the motion generation. This design achieves a better trade-off between reconstruction capacity and generative quality (See Sec. 3).

Regularized latent space. In a typical VAE training process, motion reconstruction $x^{1:N}$ is constrained by the Mean Squared Error (MSE) and Kullback-Leibler (KL) losses. We further adapt the geometric loss [51], which enhances the physical plausibility within involved individuals and preserves the original interaction relationships between individuals. The training loss of VAE is:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{geometric}} + \mathcal{L}_{\text{reconstruction}} + \lambda_{\text{KL}}\,\mathcal{L}_{\text{KL}}\,. \tag{5}$$

4.2. Unified Motion Flow Matching

As shown in Fig. 2, based on the multi-token latent space, we decouple the number-free motion generation process into two stages: (1) Motion Prior Generation: An individual motion prior is generated via the **Pyramid Motion Flow (P-Flow)**, a hierarchical flow matching process conditioned on the timestep. Unlike Denoising Diffusion Probabilistic Models (DDPMs) [18] operating in the raw motion space, this design offers better scalability [10] and efficiency within multi-token latent spaces [29, 44]. (2) Reaction Motion Generation: Given the motion prior (or preceding reaction), **Semi-Noise Motion Flow (S-Flow)** learns a joint path for context reconstruction and reaction transformation for the next person. Instead of fine-tuning complex ControlNet [66], S-Flow learns an adaptive, context-aware motion transition, alleviating potential error accumulation.

Scalability to Group Scenarios ($N > 2$). Due to the scarcity of SMPL-based [36] datasets featuring $\geq$3 interacting agents, our framework is mainly trained and evaluated on dual-agent scenarios, while UMF is not limited to this setting. For $N > 2$ people, the S-Flow module is applied autoregressively, using the synthesized motions

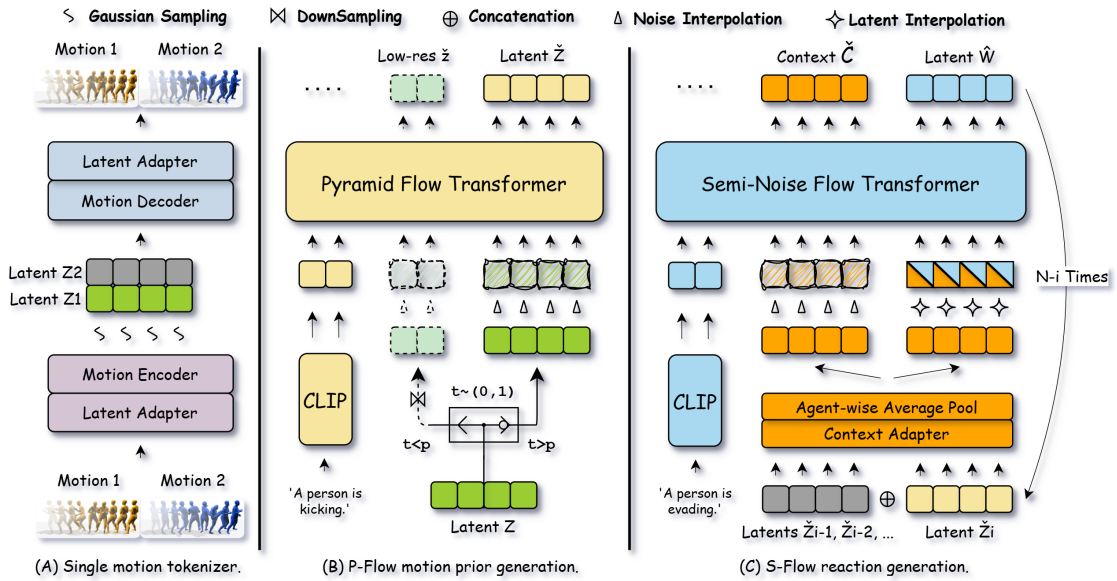

Figure 2. **Overview of the Unified Motion Flow (UMF) architecture.** The UMF framework consists of three stages. **(A) Unified motion VAE:** A motion VAE with latent adapters encodes raw motions from heterogeneous datasets (e.g., HumanML3D [13], InterHuman [32]) into a regularized multi-token latent representation (Z). **(B) P-Flow motion prior generation:** The Pyramid Flow Transformer synthesizes the latent motion prior ($\check{Z}$) based on noisy latent motion and text conditions. The P-Flow operates hierarchically based on the timestep $t \sim (0, 1)$: it processes downsampled, low-resolution latents for $t < p$ and switches to original-resolution latents for $t > p$, mitigating multi-token computational overheads. **(C) S-Flow reaction generation:** Based on the previously generated latent $\{\check{Z}_i, \ldots, \check{Z}_1\}$, the context adapter generates the context motion C. Then the Semi-Noise Flow transformer predicts the reaction latent ($\check{W}$) by jointly modeling context reconstruction and reaction transformation, alleviating the error accumulation from previously generated motion.

of preceding agents as input to generate the next agent's motion. We demonstrate its zero-shot capability via a user study (Sec. 5.3).

4.2.1. Motion Prior

Compared to single-token approaches, the multi-token latent space unlocks better motion generation conditioned on text prompt c, but it also imposes more computational demands. A key observation is that initial generation steps [55] often operate on noisy and less informative variables, suggesting that the entire full resolution is not necessary. Previous works address this by training multiple models with different resolutions [26, 60] based on the timestep, which still introduces extra model complexity. We introduce the Pyramid Motion Flow (P-Flow) [29], which reinterprets the Gaussian flow matching trajectory as hierarchical stages within one transformer model. Each stage operates at a resolution corresponding to the timestep, where only the final stage uses the original resolution, enabling efficient flow matching inference.

P-Flow forward process. Unlike standard Gaussian flow matching [20, 33] that evolves between full-resolution noise and data, P-Flow starts with a coarser interpolation between downsampled latent motion, and progressively yields finer-grained, higher-resolution endpoints. To handle the varying dimensions of z_t, we decompose the trajectory into a piecewise flow [64]. It divides $[0, 1]$ into K time windows, each interpolating between successive resolutions with a unique start and end point. For the k-th time window $[s_k, e_k]$, we jointly compute the endpoints $(\hat{z}_{s_k}, \hat{z}_{e_k})$ with noise $\epsilon \sim \mathcal{N}(\mathbf{0}, I)$ and data point z_1 as:

$$\text{Start Point:} \quad \hat{z}_{s_k} = s_k Up(Down(z_1, 2^k)) + (1 - s_k)\epsilon, \tag{6}$$

$$\text{End Point:} \quad \hat{z}_{e_k} = e_k Down(z_1, 2^{k-1}) + (1 - e_k)\epsilon, \tag{7}$$

where $k \in [K, 1]$, $Up(\cdot)$ and $Down(\cdot)$ are standard resampling functions and irreversible between them. Notably, $Up(Down(z, 2^1))$ is a lossy approximation of z, which forces the flow model to learn the correlation between resolutions. The path spans from pure noise ϵ (at $k = K, s_k = 0, \hat{z}_{s_k} = \epsilon$) to the data point z_1 (at $k = 1, e_k = 1, \hat{z}_{e_k} = Down(z_1, 2^0) = z_1$).

To enhance the straightness of the flow trajectory, we couple the sampling of its endpoints by enforcing the noise ϵ to be in the same direction. Let $t' = (t - s_k)/(e_k - s_k)$ denote the rescaled timestep, then the flow within it follows:

$$\hat{z}_t = t'\hat{z}_{e_k} + (1 - t')\hat{z}_{s_k}, \tag{8}$$

Here, the trajectory at k-th stage starts at $\hat{z}_{s_k}$ and ends at $\hat{z}_{e_k}$. This pyramidal structure, applicable to spatial or temporal dimensions, concentrates computation at lower resolutions, reducing the cost by a factor of $\approx 1/K$ in theory.

Thereafter, we can regress the flow model G_θ^P on the conditional vector field $u_t(\hat{z}_t|z_1) = \hat{z}_{e_k} - \hat{z}_{s_k}$ with the following objective to unify different stages:

$$\mathcal{L}_{\text{P-Flow}} = \mathbb{E}_{k,t,\hat{z}_{e_k},\hat{z}_{s_k}} \left\| G_\theta^P(\hat{z}_t; t, c) - (\hat{z}_{e_k} - \hat{z}_{s_k}) \right\|^2. \tag{9}$$

P-Flow sampling process. Using Euler ODE solvers, each pyramid stage is discretized into $M = T_{P_k}$ steps:

$$\hat{z}_{t_{m+1}} \leftarrow \hat{z}_{t_m} + (t_{m+1} - t_m) G_\theta^P(\hat{z}_{t_m}, t_m, c), \quad (10)$$

where $t_1 = s_k, \cdots, t_M = e_k$ are the discrete timesteps. However, we must carefully handle the jump points [5] between successive pyramid stages of different resolutions to ensure continuity of the probability path.

As shown in Algorithm 1, for the transition from stage k to $k-1$, we first upsample the previous endpoint $\hat{z}_{e_k}$ via nearest-neighbor interpolation. The inference has to match the Gaussian distributions at each jump point by a linear transformation of the upsampled result. Specifically, the following rescaling and renoising scheme suffices:

$$\hat{z}_{s_{k-1}} = \frac{s_{k-1}}{e_k} Up(\hat{z}_{e_k}) + \alpha n', \quad \text{s.t. } n' \sim \mathcal{N}(0, \Sigma'), \quad (11)$$

where Σ' is a blockwise diagonal covariance matrix (e.g., 4×4 blocks). The coefficient s_{k-1}/e_k matches the means, and the corrective noise $\alpha n'$ matches the covariances. To ensure continuity after upsampling (see Appendix B for derivation), we set $e_k = 2s_{k-1}/(1 + s_{k-1})$ and $\alpha = \frac{\sqrt{3}(1-s_{k-1})}{2}$ for a consistent mean and covariance.

4.2.2. Reaction Motion Generation

For number-free motion generation, we generate the reaction W conditioned on an arbitrary action (*i.e.*, $\hat{Z}_i$) and text prompt c. This process is applied iteratively to synthesize interactions involving more than two agents. Based on the set $\mathcal{Z}_{gen}$ of previously generated motions, Semi-Noise Flow (S-Flow) learns a **joint transformation** to generate reaction motion W for subsequent characters, which is trained exclusively on the multi-person dataset.

As shown in Fig. 2 (C), S-Flow reformulates reaction generation with context C_i by adaptively optimizing two probability paths simultaneously: (1) **reaction transformation** (the path from C_i to W) via context interpolation, and (2) **context reconstruction** (the path from ϵ to C_i) via Gaussian noise interpolation. Instead of relying on complex conditional mechanisms like ControlNet [12, 59, 62], we first employ a context adapter to generate context motion, which is used as *the direct input into flow matching*. This design provides a more flexible starting point for learning the reaction transformation paths, allowing the adaptive adjustment for possibly sub-optimal motion from other characters. The auxiliary context reconstruction path also helps S-Flow understand context at a global level, balancing its context-awareness and reaction forecasting, thereby alleviating overall error accumulation in autoregressive models.

Adaptive Context Formulation. The adapter first produces the context motion C_i by encoding the set of previously generated motions $\mathcal{Z}_{gen}$ with a transformer encoder:

$$C_i = TranEnc(\mathcal{Z}_{gen}). \quad (12)$$

Algorithm 1 UMF Inference Algorithm

Input: c (text prompt); N (agent number); K (P-Flow stage count); Models (G_θ^P, G_θ^S, Dec, $TranEnc$)
Parameters: $\{T_{P_k}\}$ (P-Flow steps); T_S (S-Flow steps)
 # P-Flow Motion Prior Generation
$\hat{z}_{s_K} \sim \mathcal{N}(0, I), 0 \le s_k < e_k \le 1, s_K = 0, e_1 = 1$
for $k = K$ **down to** 1 **do**
  $\hat{z}_{e_k} \leftarrow \text{SolveODE}(G_\theta^P, \hat{z}_{s_k}, c; T_{P_k})$ ▷ Eq. 10
  if $k \ge 2$ **then**
    $\hat{z}_{s_{k-1}} \leftarrow \text{JumpUpdate}(\hat{z}_{e_k}, s_{k-1}, e_k)$ ▷ Eq. 11
  end if
end for
$\hat{Z}_1 \leftarrow \hat{z}_{e_1}, \mathcal{Z}_{gen} \leftarrow \{\hat{Z}_1\}$
 # S-Flow Reaction Generation
for $i = 2$ to N **do**
  $C_i = TranEnc(\mathcal{Z}_{gen})$ ▷ Context Adapter
  $\hat{Z}_i \leftarrow \text{SolveODE}(G_\theta^S, C_i, c; T_S)$ ▷ Eq. 17
  $\mathcal{Z}_{gen} \leftarrow \mathcal{Z}_{gen} \cup \{\hat{Z}_i\}$
end for
 # VAE Decoding
$\{x_1, \ldots, x_N\} \leftarrow Dec(\mathcal{Z}_{gen})$
Return $\{x_1, \ldots, x_N\}$

Subsequently, if $i > 2$, agent-wise average pooling is applied to match the latent dimension of $\hat{Z}_i$. This design adaptively refines $\mathcal{Z}_{gen}$ into a concise global context, which alleviates error accumulation (See cases in Fig. 3).

S-Flow forward process. Similar to previous works [3, 33], we use the rectified flow as the backbone, which is parameterized by a neural network G_θ^S to predict vector fields, i.e., $v = w_1 - w_0$. S-Flow is trained by jointly modeling two probabilistic paths for reaction transformation and context reconstruction as follows:
(1) For the reaction path, we interpolate between the previously generated motion (context) $w_0 = C$ and the target reaction motion $w_1 = W$, w_t^{react} at timestep t is:

$$w_t^{\text{react}} = tw_1 + [1-t]w_0. \quad (13)$$

The training objective of the reaction transformation is:

$$\mathcal{L}_{\text{trans}} = \mathbb{E}_{t,w_1,w_0} \| G_\theta^S(w_t^{\text{react}}, t, c) - (W - C) \|_2^2, \quad (14)$$

where c refers to the text prompt.
(2) For the context path, we interpolate between Gaussian noise $w_0' = \epsilon$ and context motion $w_1' = C$, w_t^{cont} at timestep t is:

$$w_t^{\text{cont}} = tw_1' + [1-t]w_0'. \quad (15)$$

The training objective of the context reconstruction is:

$$\mathcal{L}_{\text{recon}} = \mathbb{E}_{t,w_0,\epsilon} \| G_\theta^S(w_t^{\text{cont}}, t, c) - (C - \epsilon) \|_2^2, \quad (16)$$

where c refers to the text prompt.

Finally, the S-Flow training objective is a weighted sum of these two losses $\mathcal{L}_{\text{S-Flow}} = \mathcal{L}_{\text{trans}} + \lambda_{\text{recon}} \mathcal{L}_{\text{recon}}$. Thus

G_θ^S learns to predict reaction for the next agent while being aware of the current context, balanced by λ_{recon}.

S-Flow sampling process. As detailed in Algorithm 1, the sampling process mirrors P-Flow by using an Euler ODE solver. The discretization process involves dividing the procedure into $M = T_S$ steps, as follows:

$$\hat{w}_{t_{m+1}} \leftarrow \hat{w}_{t_m} + (t_{m+1} - t_m) G_\theta^S(\hat{w}_{t_m}, t_m, c), \tag{17}$$

where the integer time steps $t_1 = 0 < t_2 < \cdots < t_M = 1$. The trajectory starts from the motion context C from the context adapter layer, and ends with the reaction motion W.

4.3. Justification of design choices

Asymmetric Inference Budget for UMF Efficiency. Generating motion for N agents requires one P-Flow execution and $N-1$ S-Flow executions. This structure motivates an asymmetric inference budget, as the quality of the motion prior determines the upper bound for all subsequent reactions. We therefore allocate a substantial budget to P-Flow (*e.g.*, 50 steps), which remains computationally feasible due to its pyramid structure. We find the performance of P-Flow is sensitive to the total number of steps, but far less sensitive to the ratio of low-to-high resolution steps. This allows us to assign more inference steps at low resolution (*e.g.*, 45 steps), minimizing the overhead from the multi-token representation. Furthermore, this dedicated motion prior enables the S-Flow to generate reactions with a minimal inference budget (*e.g.*, 10 steps), keeping UMF computationally tractable when N becomes large.

Shared transformer between P-Flow and S-Flow. Sharing the transformer backbone between P-Flow and S-Flow would reduce the overall parameter count [12]. However, we found that a shared backbone struggles to converge and yields degraded performance (See Tab. 5). We attribute this to two factors: 1) P-Flow focuses on mapping noise to motion, while S-Flow learns both motion-to-motion and noise-to-motion paths. These tasks are incompatible and challenging to optimize. 2) The continuity guarantees [29] at the pyramid jump points are difficult to maintain, which assume tractable distributions (e.g., Gaussian noise), while S-Flow operates on complex motion distributions with intractable means and variances. Therefore, UMF employs separate P-Flow and S-Flow modules. The S-Flow transformer is shared autoregressively to generate the reaction for the subsequent agent, using all previously generated motions as context.

5. Experiments

5.1. Datasets, Metrics & Implementation Details

We utilize InterHuman [32] and HumanML3D [13] datasets for the evaluation of text-conditioned motion generation performance. The InterHuman and HumanML3D datasets contain 7,779 interaction sequences and 14,616 individual sequences, respectively, where each sequence is illustrated with 3 textual annotations. The InterHuman-AS dataset [62] is essentially the same as InterHuman, but includes additional actor–reactor order annotations. We employ the evaluation metrics following previous studies [13, 32]. Fidelity is assessed using Frechet Inception Distance (FID), R-precision, and Multimodal Distance (MM Dist), and diversity is evaluated with Diversity and Multimodality scores. All our models are trained with the AdamW optimizer using an initial learning rate of 10^{-4} and a cosine decay schedule. Our mini-batch size is set to 128 during the VAE training stage and 64 during the flow matching training stage. Each model was trained for 6K epochs during the VAE stage, 2K epochs during the P-Flow and 2K epochs during S-Flow stage. See Appendix C for details.

5.2. Quantitative Results

As shown in Table 1, on the InterHuman benchmark, UMF substantially outperforms the generalist baseline, FreeMotion [12], improving Top3 R-Precision by 28% and reducing FID by 29%. Furthermore, its Diversity score closely matches the ground truth, indicating a highly realistic output. Against specialist methods tailored for dual-agent scenarios, UMF demonstrates competitive performance, outperforming the strongest baseline, InterMask [22], by 7% in FID. It also achieves the second-best results on R-Precision and MM-Distance, demonstrating competitive text-following ability. In Table 2, we compare UMF with existing approaches on the InterHuman-AS dataset, where we observe a similar trend. Specifically, UMF improves Top3 R-Precision by over 30% and reduces MM-Distance by 27% compared to ReGenNet [62], significantly improving the reactive motion quality.

5.3. Qualitative Results & User Study

Fig. 3 demonstrates UMF's ability to generate more realistic human interactions compared to FreeMotion [12]. In the "kick" (dual-agent) scenario, UMF generates a plausible kicking motion with correct leg assignment. In contrast, FreeMotion fails to produce a coherent generation, only attempting a poorly directed kick in the end. In the "stroll" (three-agent) scenario, UMF correctly positions the third agent (green) between the other two (yellow, blue), maintaining plausible proximity, while FreeMotion's output suffers from severe interpenetration. In the complex, multi-agent ($N > 3$) "fight" scenario, FreeMotion fails to animate all participants, resulting in artifacts such as the static poses of agents. In contrast, UMF generalizes effectively to this zero-shot number-free task, producing dynamic and plausible interactions.

Due to the scarcity of motion databases for group scenarios, we conducted a user study for assessing UMF's zero-shot generalization capability (see Appendix D). The pro-

Figure 3. Qualitative comparison (zoom into see it better) between FreeMotion [12] and UMF. Red circles demonstrate successful cases, while Blue circles show failure cases.

Table 1. Quantitative evaluation on the InterHuman test sets. $\pm$ indicates a 95% confidence interval and $\rightarrow$ means the closer to ground truth the better. **Boldface** indicates the best result, while underline refers to the second best.

Method	R Top 1↑	R Top 2↑	R Top 3↑	FID↓	MM Dist↓	Diversity→	MModality↑
Ground Truth	$0.452^{\pm0.008}$	$0.610^{\pm0.009}$	$0.701^{\pm0.008}$	$0.273^{\pm0.007}$	$3.755^{\pm0.008}$	$7.948^{\pm0.064}$	-
TEMOS [41]	$0.224^{\pm0.010}$	$0.316^{\pm0.013}$	$0.450^{\pm0.018}$	$17.375^{\pm0.043}$	$6.342^{\pm0.015}$	$6.939^{\pm0.071}$	$0.535^{\pm0.014}$
T2M [13]	$0.238^{\pm0.012}$	$0.325^{\pm0.010}$	$0.464^{\pm0.014}$	$13.769^{\pm0.072}$	$5.731^{\pm0.013}$	$7.046^{\pm0.022}$	$1.387^{\pm0.076}$
MDM [51]	$0.153^{\pm0.012}$	$0.260^{\pm0.009}$	$0.339^{\pm0.012}$	$9.167^{\pm0.056}$	$7.125^{\pm0.018}$	$7.602^{\pm0.045}$	$\mathbf{2.350^{\pm0.080}}$
ComMDM [46]	$0.223^{\pm0.009}$	$0.334^{\pm0.008}$	$0.466^{\pm0.010}$	$7.069^{\pm0.054}$	$6.212^{\pm0.021}$	$7.244^{\pm0.038}$	$1.822^{\pm0.052}$
InterGen [32]	$0.371^{\pm0.010}$	$0.515^{\pm0.012}$	$0.624^{\pm0.010}$	$5.918^{\pm0.079}$	$5.108^{\pm0.014}$	$7.387^{\pm0.029}$	$2.141^{\pm0.063}$
MoMat-MoGen [4]	$0.449^{\pm0.004}$	$0.591^{\pm0.003}$	$0.666^{\pm0.004}$	$5.674^{\pm0.085}$	$3.790^{\pm0.001}$	$8.021^{\pm0.035}$	$1.295^{\pm0.023}$
in2IN [43]	$0.425^{\pm0.008}$	$0.576^{\pm0.008}$	$0.662^{\pm0.009}$	$5.535^{\pm0.120}$	$3.803^{\pm0.002}$	$7.953^{\pm0.047}$	$1.215^{\pm0.023}$
TIMotion [57]	$\mathbf{0.491^{\pm0.005}}$	$\mathbf{0.648^{\pm0.004}}$	$\mathbf{0.724^{\pm0.004}}$	$5.433^{\pm0.080}$	$\mathbf{3.775^{\pm0.001}}$	$8.032^{\pm0.030}$	$0.952^{\pm0.032}$
InterMask [22]	$0.449^{\pm0.004}$	$0.599^{\pm0.005}$	$0.683^{\pm0.004}$	$5.154^{\pm0.061}$	$3.790^{\pm0.002}$	$\mathbf{7.944^{\pm0.033}}$	$1.737^{\pm0.020}$
FreeMotion	$0.326^{\pm0.003}$	$0.462^{\pm0.006}$	$0.544^{\pm0.006}$	$6.740^{\pm0.130}$	$3.848^{\pm0.002}$	$7.828^{\pm0.130}$	$1.226^{\pm0.046}$
UMF	$0.467^{\pm0.004}$	$0.620^{\pm0.004}$	$0.694^{\pm0.005}$	$\mathbf{4.772^{\pm0.079}}$	$3.784^{\pm0.001}$	$8.039^{\pm0.032}$	$1.398^{\pm0.012}$

Table 2. Comparison to state-of-the-art for human action-reaction synthesis on the InterHuman-AS dataset. $\pm$ indicates 95% confidence interval, $\rightarrow$ means that closer to Real is better. **Bold** indicates best result and underline indicates second best.

Methods	RTop3↑	FID ↓	MM Dist↓	Diversity→	MModality ↑
Real	$0.722^{\pm0.004}$	$0.002^{\pm0.0002}$	$3.503^{\pm0.011}$	$5.390^{\pm0.058}$	-
T2M [13]	$0.224^{\pm0.003}$	$32.482^{\pm0.098}$	$7.299^{\pm0.016}$	$4.350^{\pm0.073}$	$0.719^{\pm0.041}$
MDM [51]	$0.370^{\pm0.006}$	$3.397^{\pm0.052}$	$8.640^{\pm0.045}$	$4.780^{\pm0.015}$	$2.288^{\pm0.039}$
MDM-GRU [51]	$0.328^{\pm0.012}$	$6.397^{\pm0.214}$	$8.884^{\pm0.040}$	$4.851^{\pm0.081}$	$2.076^{\pm0.040}$
RAIG [49]	$0.363^{\pm0.008}$	$2.915^{\pm0.029}$	$7.294^{\pm0.027}$	$4.736^{\pm0.099}$	$2.203^{\pm0.049}$
InterGen [32]	$0.374^{\pm0.005}$	$13.237^{\pm0.035}$	$10.929^{\pm0.026}$	$4.376^{\pm0.042}$	$\mathbf{2.793^{\pm0.014}}$
ReGenNet [62]	$0.407^{\pm0.003}$	$\mathbf{2.265^{\pm0.097}}$	$6.860^{\pm0.040}$	$\mathbf{5.214^{\pm0.139}}$	$2.391^{\pm0.023}$
FreeMotion [12]	$0.409^{\pm0.006}$	$3.896^{\pm0.029}$	$7.632^{\pm0.036}$	$6.089^{\pm0.027}$	$2.496^{\pm0.036}$
UMF	$\mathbf{0.530^{\pm0.006}}$	$2.577^{\pm0.024}$	$\mathbf{4.987^{\pm0.011}}$	$7.764^{\pm0.024}$	$2.116^{\pm0.041}$

posed UMF and FreeMotion [12] are compared according to the aspects of text alignment, physical realism, interaction quality, and overall quality. 30 unique users participated in the user study, with 20 randomly sampled multi-person generations ($N > 2$). The zero-shot results in Fig. 4 show that the number-free motions generated by UMF were clearly preferred over those generated by FreeMotion.

5.4. Ablation Studies

Heterogeneous Priors and Latent Space. Table 3 investigates the impact of individual priors from HumanML3D [13] and our latent space design. The results demonstrate that models trained with the HumanML3D prior outperform those without, improving both text adherence and motion fidelity. This highlights the potential of leveraging single-agent datasets to enhance multi-agent in-

Table 3. Ablation study of individual priors on the HumanML3D and InterHuman datasets. **HP**: **H**eterogeneous **P**riors; **LA**: **L**atent **A**dapter; **MT**: **M**ulti-token **T**okenizer.

	Components			HumanML3D		InterHuman	
Method	HP	LA	MT	RTop3 ↑	FID ↓	RTop3↑	FID↓
Real Data	-	-	-	0.797	0.002	0.701	0.273
FreeMotion	✓	-	-	0.612	3.539	0.503	7.984
w/o HP	✗	-	-	0.128	13.155	0.544	6.740
UMF (Full)	✓	✓	✓	0.729	0.486	0.694	4.772
w/o HP	✗	✓	✓	0.134	10.806	0.651	4.933
w/o LA	✓	✗	✓	0.708	0.646	0.627	5.473
w/o MT	✓	✓	✗	0.713	0.534	0.655	5.231
w/o (HP + LA)	✗	✗	✓	0.131	12.914	0.579	5.561
w/o (HP + MT)	✗	✓	✗	0.139	12.748	0.616	5.493
w/o (LA + MT)	✓	✗	✗	0.682	0.845	0.605	5.668
w/o (HP + LA + MT)	✗	✗	✗	0.125	13.844	0.511	6.036

Table 4. Ablation study of Pyramid Flow on the InterHuman dataset. UMF has a 2-stage temporal pyramid structure. We report FLOPs(G) and AITS (Average Inference Time in Seconds). T_{P2}, T_{P1}, and T_S refer to the corresponding inference step for P-Flow low-res stage, P-Flow full-res stage, and S-Flow, respectively. UMF-PFK1 and UMF-PFS refer to P-Flow with the original resolution and with the spatial pyramid structure, respectively.

Methods	T_{P_2}	T_{P_1}	T_S	FID ↓	RTop3↑	FLOPs(G)↓	AITS↓
FreeMotion	-	-	-	6.740	0.544	217.8	3.059
UMF	45	5	10	4.772	0.694	140.3	0.623
UMF-PFK1	-	50	10	4.761	0.674	320.2	1.119
UMF-PFS	45	5	10	7.238	0.527	135.9	0.581
UMF-Fast	5	5	10	4.937	0.687	74.7	0.439
UMF-Symmetric	25	25	10	4.784	0.697	206.0	0.874

teraction generation. We attribute the modest improvement to the complexity gap between the single-agent and multi-agent generation targets, which manifests as a challenging cross-dataset transfer effect. Furthermore, we compare UMF against variants without the Latent Adapter (w/o. LA) and with a single-token latent space (1×256). The results indicate that the Latent Adapter is crucial for multi-token flow matching, whereas the single-token variant lacks sufficient capacity to model number-free generation effectively.

Efficiency Analysis of Pyramid Flow. Table 4 ablates the Pyramid Flow (PF) structure and its inference step allocation. First, we compare UMF with FreeMotion under the same inference steps (*i.e.*, 60 steps), where UMF achieves lower FLOPs and is nearly 5× faster, which demonstrates the efficiency of P-Flow for complex interaction generation. Next, we compare two variants, where UMF-PFK1 achieves a slightly better FID but with extra computational cost. Conversely, the UMF-PFS variant shows severe performance degradation. We also find that reducing the P-Flow budget from 50 to 10 steps can nearly halve the FLOPs but degrade performance. Notably, allocating asymmetric steps (T_{P2} = 45, T_{P1} = 5) achieves the best speed-quality trade-off, yielding competitive FID with fewer FLOPs compared to a symmetric allocation ($T_{P2} = T_{P1} = 25$).

Semi-Noise Flow Component Analysis. Table 5 analyzes the key components of S-Flow. Sharing the transformer backbone between S-Flow and P-Flow, while parameter-efficient, results in significantly worse fidelity, likely due to their incompatible learning paths. We also compare the semi-noise flow with noise-free flow in [27], which only learns the reaction transformation path. This variant shows degraded performance without considering the error accumulation. Removing the reconstruction loss also harms generation quality. Similarly, removing the Context Adapter entirely or replacing it with a ControlNet [66] both lead to a significant performance drop, underscoring the importance of context reconstruction. In contrast, the transformer-based adapter in UMF preserves a global view of the entire context.

Table 5. Ablation study of Semi-Noise Flow on the InterHuman dataset.

Methods	FID ↓	RTop3↑	Diversity→
UMF	4.772	0.694	8.039
UMF w. Shared Transformer	6.206	0.644	8.088
UMF w. Noise-Free path [27]	5.617	0.646	8.112
UMF w. ControlNet [66]	6.868	0.637	8.061
UMF w/o. Context Adapter	7.038	0.642	8.087
UMF w/o. L_{recons}	5.765	0.649	8.124

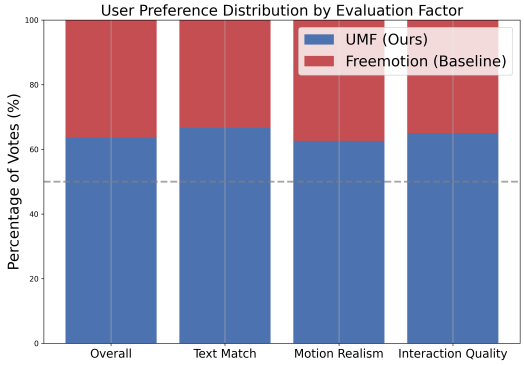

Figure 4. The UMF number-free zero-shot generation user study. We asked users to compare our UMF (Blue Bar) to the FreeMotion (Red Bar) in a side-by-side view. The dashed line marks 50%. UMF outperforms FreeMotion in all three aspects of generation.

6. Conclusion

We introduce Unified Motion Flow (UMF), a generalist framework for number-free, text-conditioned motion generation, which consists of Pyramid Motion Flow (P-Flow) and Semi-Noise Motion Flow (S-Flow). Based on a unified heterogeneous latent space, UMF achieves number-free motion generation via P-Flow for mitigating computational overheads and S-Flow for alleviating error accumulation. Extensive results show UMF achieves state-of-the-art performance for multi-person generation, and exhibits robust zero-shot generalization to challenging group scenarios. While the $1 + N$ paradigm enhances generalization, UMF remains constrained to medium sized group interactions ($\approx$ 10 agents) centered on a primary agent. Future will explore leveraging visual priors from large-scale video diffusion models to scale synthesis to dense crowd dynamics ($\approx$ 100 agents).

Acknowledgments

This work was supported by the K-CSC funding. The authors acknowledge the use of King's CREATE HPC. Retrieved March 24, 2026, from https://doi.org/10.18742/rnvf-m076.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 2

[3] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022. 3, 5

[4] Zhongang Cai, Jianping Jiang, Zhongfei Qing, Xinying Guo, Mingyuan Zhang, Zhengyu Lin, Haiyi Mei, Chen Wei, Ruisi Wang, Wanqi Yin, et al. Digital life project: Autonomous 3d characters with social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 582–592, 2024. 7

[5] Andrew Campbell, William Harvey, Christian Weilbach, Valentin De Bortoli, Thomas Rainforth, and Arnaud Doucet. Trans-dimensional generative modeling via jump diffusion models. *Advances in Neural Information Processing Systems*, 36:42217–42257, 2023. 5

[6] Ling-Hao Chen, Wenxun Dai, Xuan Ju, Shunlin Lu, and Lei Zhang. Motionclr: Motion generation and training-free editing via understanding attention mechanisms. *arXiv e-prints*, pages arXiv–2410, 2024. 2

[7] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 2, 3

[8] Wenxun Dai, Ling-Hao Chen, Yufei Huo, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Real-time controllable motion generation via latent consistency model. 3

[9] Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. *arXiv preprint arXiv:2404.19759*, 2024. 2

[10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 3

[11] Ke Fan, Shunlin Lu, Minyue Dai, Runyi Yu, Lixing Xiao, Zhiyang Dou, Junting Dong, Lizhuang Ma, and Jingbo Wang. Go to zero: Towards zero-shot motion generation with million-scale data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13336–13348, 2025. 1

[12] 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*, pages 93–109. Springer, 2025. 1, 2, 3, 5, 6, 7

[13] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 1, 3, 4, 6, 7

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

[15] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2

[16] Ruoxi Guo, Huaijin Pi, Zehong Shen, Qing Shuai, Zechen Hu, Zhumei Wang, Yajiao Dong, Ruizhen Hu, Taku Komura, Sida Peng, et al. Motion-2-to-3: Leveraging 2d motion data for 3d motion generations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14305–14316, 2025. 1

[17] Prerit Gupta, Shourya Verma, Ananth Grama, and Aniket Bera. Unified multi-modal interactive & reactive 3d motion generation via rectified flow. *arXiv preprint arXiv:2509.24099*, 2025. 2

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[19] Fangzhou Hong, Vladimir Guzov, Hyo Jin Kim, Yuting Ye, Richard Newcombe, Ziwei Liu, and Lingni Ma. Egolm: Multi-modal language model of egocentric motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5344–5354, 2025. 1

[20] Vincent Tao Hu, Wenzhe Yin, Pingchuan Ma, Yunlu Chen, Basura Fernando, Yuki M Asano, Efstratios Gavves, Pascal Mettes, Bjorn Ommer, and Cees GM Snoek. Motion flow matching for human motion synthesis and editing. *arXiv preprint arXiv:2312.08895*, 2023. 4

[21] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 1

[22] Muhammad Gohar Javed, Chuan Guo, Li Cheng, and Xingyu Li. Intermask: 3d human interaction generation via collaborative masked modelling. *arXiv preprint arXiv:2410.10010*, 2024. 6, 7

[23] Kaiyang Ji, Ye Shi, Zichen Jin, Kangyi Chen, Lan Xu, Yuexin Ma, Jingyi Yu, and Jingya Wang. Towards immersive

human-x interaction: A real-time framework for physically plausible motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10173–10183, 2025. 1

[24] Yatai Ji, Teng Wang, Yuying Ge, Zhiheng Liu, Sidi Yang, Ying Shan, and Ping Luo. From denoising to refining: A corrective framework for vision-language diffusion model. *arXiv preprint arXiv:2510.19871*, 2025. 2

[25] 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. 2

[26] Lei Jiang, Ye Wei, and Hao Ni. Motionpcm: Real-time motion synthesis with phased consistency model. *arXiv preprint arXiv:2501.19083*, 2025. 4

[27] Wentao Jiang, Jingya Wang, Haotao Lu, Kaiyang Ji, Baoxiong Jia, Siyuan Huang, and Ye Shi. Arflow: Human action-reaction flow matching with physical guidance. *arXiv preprint arXiv:2503.16973*, 2025. 2, 8

[28] Peng Jin, Yang Wu, Yanbo Fan, Zhongqian Sun, Wei Yang, and Li Yuan. Act as you wish: Fine-grained control of motion diffusion model with hierarchical semantic graphs. *Advances in Neural Information Processing Systems*, 36: 15497–15518, 2023. 2

[29] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024. 2, 3, 4, 6

[30] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[31] Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. Genmo: A generalist model for human motion. *arXiv preprint arXiv:2505.01425*, 2025. 2

[32] 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*, pages 1–21, 2024. 1, 2, 3, 4, 6, 7

[33] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3, 4, 5

[34] 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. 3

[35] Zuhong Liu, Junhao Ge, Minhao Xiong, Jiahao Gu, Bowei Tang, Wei Jing, and Siheng Chen. It takes two: Learning interactive whole-body control between humanoid robots. *arXiv preprint arXiv:2510.10206*, 2025. 1

[36] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 3

[37] 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*, pages 23–40. Springer, 2024. 3

[38] Yiyi Ma, Yuanzhi Liang, Xiu Li, Chi Zhang, and Xuelong Li. Intersyn: Interleaved learning for dynamic motion synthesis in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12832–12841, 2025. 2

[39] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 1

[40] Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Tridi: Trilateral diffusion of 3d humans, objects, and interactions. *arXiv preprint arXiv:2412.06334*, 2024. 2

[41] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. *arXiv preprint arXiv:2204.14109*, 2022. 3, 7

[42] Ekkasit Pinyoanuntapong, Muhammad Saleem, Korrawe Karunratanakul, Pu Wang, Hongfei Xue, Chen Chen, Chuan Guo, Junli Cao, Jian Ren, and Sergey Tulyakov. Maskcontrol: Spatio-temporal control for masked motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9955–9965, 2025. 2

[43] Pablo Ruiz Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and Jose Garcia-Rodriguez. in2in: Leveraging individual information to generate human interactions. *arXiv preprint arXiv:2404.09988*, 2024. 2, 7

[44] Lingmin Ran and Mike Zheng Shou. Tpdiff: Temporal pyramid video diffusion model. *arXiv preprint arXiv:2503.09566*, 2025. 3

[45] Pablo Ruiz-Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and José García-Rodríguez. Mixermdm: Learnable composition of human motion diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12380–12390, 2025. 1

[46] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 1, 7

[47] Mengyi Shan, Lu Dong, Yutao Han, Yuan Yao, Tao Liu, Ifeoma Nwogu, Guo-Jun Qi, and Mitch Hill. Towards open domain text-driven synthesis of multi-person motions. In *European Conference on Computer Vision*, pages 67–86. Springer, 2024. 1

[48] Wenhui Tan, Boyuan Li, Chuhao Jin, Wenbing Huang, Xiting Wang, and Ruihua Song. Think-then-react: Towards unconstrained human action-to-reaction generation. *arXiv preprint arXiv:2503.16451*, 2025. 2

[49] Mikihiro Tanaka and Kent Fujiwara. Role-aware interaction generation from textual description. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15999–16009, 2023. 7

[50] Jiayan Teng, Wendi Zheng, Ming Ding, Wenyi Hong, Jianqiao Wangni, Zhuoyi Yang, and Jie Tang. Relay diffusion: Unifying diffusion process across resolutions for image synthesis. *arXiv preprint arXiv:2309.03350*, 2023. 2

[51] 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. 1, 2, 3, 7

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3

[53] Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. Tlcontrol: Trajectory and language control for human motion synthesis. In *European Conference on Computer Vision*, pages 37–54. Springer, 2024. 2

[54] Jing Wang, Fengzhuo Zhang, Xiaoli Li, Vincent Y. F. Tan, Tianyu Pang, Chao Du, Aixin Sun, and Zhuoran Yang. Error analyses of auto-regressive video diffusion models: A unified framework, 2025. 2

[55] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 133(5):3059–3078, 2025. 2, 4

[56] Yatian Wang, Haoran Mo, and Chengying Gao. Difusion: Flexible stylized motion generation using digest-and-fusion scheme. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 1

[57] Yabiao Wang, Shuo Wang, Jiangning Zhang, Ke Fan, Jiafu Wu, Zhucun Xue, and Yong Liu. Timotion: Temporal and interactive framework for efficient human-human motion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7169–7178, 2025. 1, 2, 7

[58] Qingxuan Wu, Zhiyang Dou, Chuan Guo, Yiming Huang, Qiao Feng, Bing Zhou, Jian Wang, and Lingjie Liu. Text2interact: High-fidelity and diverse text-to-two-person interaction generation. *arXiv preprint arXiv:2510.06504*, 2025. 1

[59] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. *arXiv preprint arXiv:2310.08580*, 2023. 5

[60] Zhenyu Xie, Yang Wu, Xuehao Gao, Zhongqian Sun, Wei Yang, and Xiaodan Liang. Towards detailed text-to-motion synthesis via basic-to-advanced hierarchical diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6252–6260, 2024. 2, 4

[61] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-x: Towards versatile human-human interaction analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22260–22271, 2024. 1, 2

[62] Liang Xu, Yizhou Zhou, Yichao Yan, Xin Jin, Wenhan Zhu, Fengyun Rao, Xiaokang Yang, and Wenjun Zeng. Regennet: Towards human action-reaction synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1759–1769, 2024. 2, 5, 6, 7

[63] Liang Xu, Chengqun Yang, Zili Lin, Fei Xu, Yifan Liu, Congsheng Xu, Yiyi Zhang, Jie Qin, Xingdong Sheng, Yunhui Liu, et al. Perceiving and acting in first-person: A dataset and benchmark for egocentric human-object-human interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12535–12548, 2025. 1

[64] Hanshu Yan, Xingchao Liu, Jiachun Pan, Jun Hao Liew, Qiang Liu, and Jiashi Feng. Perflow: Piecewise rectified flow as universal plug-and-play accelerator. *arXiv preprint arXiv:2405.07510*, 2024. 4

[65] Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15703–15712, 2025. 3

[66] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 3, 8

[67] Zeyu Zhang, Akide Liu, Ian Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. Motion mamba: Efficient and long sequence motion generation. In *European Conference on Computer Vision*, pages 265–282. Springer, 2025. 2

[68] Kaifeng Zhao, Gen Li, and Siyu Tang. Dartcontrol: A diffusion-based autoregressive motion model for real-time text-driven motion control. *arXiv preprint arXiv:2410.05260*, 2024. 2

[69] Yiwen Zhao, Yang Wang, Liting Wen, Hengyuan Zhang, and Xingqun Qi. Freedance: Towards harmonic free-number group dance generation via a unified framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10560–10569, 2025. 2, 3

[70] Bingfan Zhu, Biao Jiang, Sunyi Wang, Shixiang Tang, Tao Chen, Linjie Luo, Youyi Zheng, and Xin Chen. Motiongpt3: Human motion as a second modality. *arXiv preprint arXiv:2506.24086*, 2025. 2, 3