TEXT2INTERACT: HIGH-FIDELITY AND DIVERSE TWO-PERSON INTERACTION GENERATION FROM TEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating realistic and diverse human–human interactions from text remains a fundamental but challenging problem in vision, graphics, and robotics. Current approaches face two main limitations: (i) interaction synthesis requires both high-quality individual motion and precise spatiotemporal coordination, yet existing datasets are too small to support such complexity, limiting generalization; and (ii) complex interactions often demand detailed textual descriptions, but sentence-level embeddings fail to capture fine-grained semantics. We address these issues with two contributions. First, we introduce InterCompose, a scalable data synthesis framework that combines the general knowledge of large language models with strong single-person motion priors to generate high-quality two-person interactions beyond existing distributions. Second, we propose Text2Interact, which employs word-level attention for fine-grained text–motion alignment and an adaptive supervision signal that dynamically weights body parts based on interaction context to enhance realism. Extensive experiments demonstrate that our approach substantially improves motion diversity, semantic alignment, and realism over state-of-the-art baselines. Our code and models will be released for reproducibility.

1 INTRODUCTION

Modeling realistic and controllable two-person interactions is a fundamental challenge in human motion generation, with applications in animation, virtual reality, and human–robot collaboration. Despite rapid progress in single-person motion synthesis (Tevet et al., 2022b; Zhou et al., 2024; Wan et al., 2024; Cong et al., 2024), extending these capabilities to diverse two-person scenarios remains difficult due to two key limitations. *1) Limited data.* High-quality interaction generation requires not only plausible individual motions but also precise spatiotemporal coordination and semantic consistency between agents. Training such models demands large-scale corpora, yet current two-person datasets are markedly smaller than single-person counterparts (e.g., INTERHUMAN (Liang et al., 2024) $<$ 8k sequences vs. HUMANML3D (Guo et al., 2022) $>$ 14k), constraining diversity and generalization. The high cost of capturing interaction data makes scalable synthesis essential. *2) Insufficient interaction modeling.* Two-person interactions are language-rich: INTERHUMAN captions have a median of 21 words vs. 7 for HUMANML3D. Yet prior methods (Liang et al., 2024; Tanaka & Fujiwara, 2023; Javed et al., 2024) compress prompts into a single sentence-level embedding, discarding fine-grained spatial and temporal cues needed for faithful alignment. This bottleneck limits both diversity and text–motion fidelity.

In this paper, we address the data scarcity challenge by proposing InterCompose, which synthesizes diverse and plausible paired texts and two-person interactions from language and single-person motion priors. *Our key insight is that a wide range of interaction patterns can be effectively composed from single-person motions by ensuring the motions' alignment with interaction semantics and spatial-temporal consistencies between the interacting motions.* To synthesize and compose single-person motion primitives, we first generate diverse two-person interaction texts paired with accurate and succinct single-person motion descriptions using an LLM (Liu et al., 2024), utilizing the general world knowledge encoded in the LLM for semantic richness and variety. Then, we sample single-person motions from a state-of-the-art generative model (Guo et al., 2024) and trained a conditional reaction generation model that generates the second party of an interaction given the first party and the interaction description, leveraging the strong language and motion priors while ensuring interaction semantics alignment and spatial-temporal consistency. To ensure the motion quality and

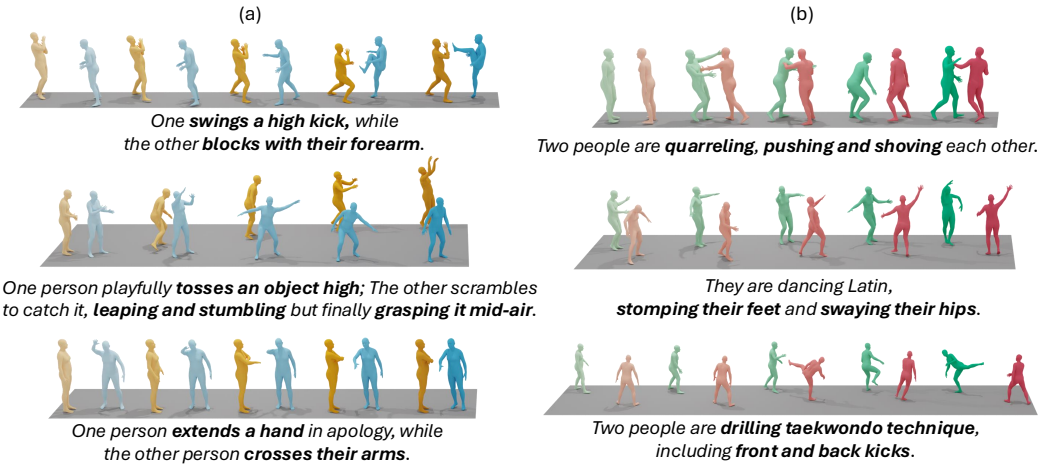

Figure 1: (a) Our generative two-person motion composition framework, InterCompose, synthesizes plausible and diverse interactions from generated textual descriptions and a single-person motion condition (yellow). (b) Our interaction generation framework Text2Interact generates high-quality and plausible interactions faithful to text. A deeper color indicates a later time.

diversity of the synthesized motions, we introduce a neural motion evaluator to measure the quality and text-motion alignment of the synthesized data. As a result, incorporating our synthesized and filtered dataset enhances the model's ability to generate unseen interactions (see Tab. 2), improving generalizability without requiring additional real data.

To faithfully capture the nuanced semantics in text-to-interaction generation, we revisit the language conditioning mechanism and propose Text2Interact, a new text-to-interaction generation framework equipped with a novel word-level text conditioning module that injects fine-grained semantic information throughout the generation process. This is motivated by the observation that *while single person motions can often be described with abstract phrases or words ("dances", "playing golf", "sidesteps left"), two-person motions often requires a sequence of consecutive phrases to accurately describe, with spatial and temporal cues embedded in the language for synchronizing the interaction*. In contrast to methods that inject a single sentence-level embedding, our approach preserves detailed textual semantics and avoids information bottlenecks caused by compressing rich interaction descriptions into a single vector. To take advantage of this rich representation, we leverage the cross-attention mechanism, in which each motion token dynamically attends to all individual tokens in the textual prompt, leading to better semantic alignment. To enhance interaction plausibility, we design an adaptive interaction loss that dynamically weights joint-pair distances based on their spatial relevance, promoting tighter physical and contextual coupling between agents during training. Unlike existing diffusion-based methods (Liang et al., 2024; Ruiz-Ponce et al., 2024) that treat all inter-person joint pairs equally, our loss emphasizes spatially proximate joints, such as hands or arms in handshakes and sparring, thereby encouraging tighter contextual coupling between the two agents during training.

Extensive experiments demonstrate that our method achieves state-of-the-art performance in two-person motion generation, outperforming prior art in terms of motion fidelity, faithfulness, and generalizability. Moreover, our ablation studies validate the effectiveness of each component in the proposed framework, especially in scenarios where real interaction data is sparse. In summary, our contribution is three-fold:

- A scalable synthesis-and-filtering strategy (InterCompose) that constructs high-quality, diverse two-person interactions from LLM text priors and single-person motion priors.
- A word-level attention conditioning module (Text2Interact) with an adaptive interaction loss for semantically faithful and spatiotemporally coherent two-person generation.
- State-of-the-art results on standard benchmarks and superior performance in challenging out-of-distribution settings via a broad user study.

2 Related Works

2.1 Text-to-Human Motion Generation

Text-to-motion generation aims to synthesize human motion sequences from natural language descriptions (Fan et al., 2024; Tanke et al., 2023; Jeong et al., 2024; Jiang et al., 2023; Guo et al., 2022; Zhang et al., 2023a; Wan et al., 2024; Lu et al., 2024; Guo et al., 2024). Early methods such as Text2Action (Ahn et al., 2018) and Language2Pose (Ahuja & Morency, 2019) utilized GANs and sequence-to-sequence architectures to map text to motion, laying foundational work in this area. Subsequent approaches leveraged variational autoencoders (VAEs) for probabilistic generation, including Guo et al. (Guo et al., 2022) and TEMOS (Petrovich et al., 2022), which improved motion diversity and fluency. More recent advancements have focused on powerful generative models. Diffusion-based approaches such as MDM (Tevet et al., 2022b) and latent diffusion via MLD (Chen et al., 2023) significantly improved motion realism and sample efficiency. T2M-GPT (Zhang et al., 2023a) employed autoregressive transformers for fine-grained motion synthesis, while MoMask (Guo et al., 2024) introduced generative masked transformers to enhance fidelity under the autoregressive paradigm. ReMoDiffuse (Zhang et al., 2023b) further enhanced generation quality by retrieving reference motions from a motion database. Parallel to improving generation quality, increasing attention has been given to controllable text-to-motion generation. Techniques have explored conditioning on spatial trajectories (Shafir et al., 2023; Karunratanakul et al., 2023; Wan et al., 2024; Xie et al., 2023) and linguistic constraints (Wan et al., 2024; Huang et al., 2024) to provide more precise control over generated outputs. Additionally, MotionCLIP (Tevet et al., 2022a) aligned motion and language embeddings in a shared space, enabling zero-shot text-to-motion generation. Despite stellar results in single-person motion generation, extending them to two-person interactions introduces additional challenges such as modeling inter-agent coordination and handling semantically richer text descriptions. Our work builds on these foundations by proposing a scalable framework that composes diverse and semantically aligned two-person interactions from single-person motion priors and language models.

2.2 Human-Human Interaction Generation

Although some progress has been achieved in multi-human interaction modeling (Fan et al., 2024; Tanke et al., 2023; Jeong et al., 2024), prior works on human interaction modeling have been mostly focused on the two-person interaction problem. A pioneer work, ComMDM (Shafir et al., 2023), explores two-person motion generation by using a bridge network to compose the outputs of two single-person motion diffusion models (Tevet et al., 2022b). RIG (Tanaka & Fujiwara, 2023) and InterGen (Liang et al., 2024) first trained dedicated networks to directly model two-person interaction. in2IN (Ruiz-Ponce et al., 2024) explores the simultaneous use of individual and interaction descriptions to enhance textual alignment and generation quality. MoMat-MoGen (Cai et al., 2024) proposes to enhance generation quality by retrieving from a motion database and a generative framework that models interactive behaviors between agents, considering personality, motivations, and interpersonal relationships. InterMask (Javed et al., 2024) utilizes the generative masked transformer architecture and spatial-temporal attention to enhance generation quality and text-motion alignment. TIMotion (Wang et al., 2024), a contemporaneous work, proposes to model the human interaction sequence in a causal sequence, leveraging the temporal and causal properties of human motions. Although these methods have achieved impressive results, there remains significant possibilities of improvement due to their common flaw of limited training corpus and inadequate text modeling granularity. In this paper, we aim to tackle these two key issues with our generative interaction composition framework and fine-grained word-level conditioning module.

3 Method

Problem Formulation. Given a text prompt c_t, the task of *human-human interaction generation from text* involves generating a two-person interaction sequence $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2] \in \mathbb{R}^{2\times T\times N\times 3}$ that is both semantically and spatially coherent and faithful to the original text prompt, where $\mathbf{x}_i$ denotes the i-th person's motion sequence, T is the sequence length in frames and N is the number of joints. Following standard practice in single human and interaction generation (Guo et al., 2022; Liang et al., 2024; Ponce et al., 2024), we use a extended representation formulated as: $\mathbf{x}_i^{(t)} = [\mathbf{j}_g^p, \mathbf{j}_g^v, \mathbf{j}^r, \mathbf{c}_f]$,

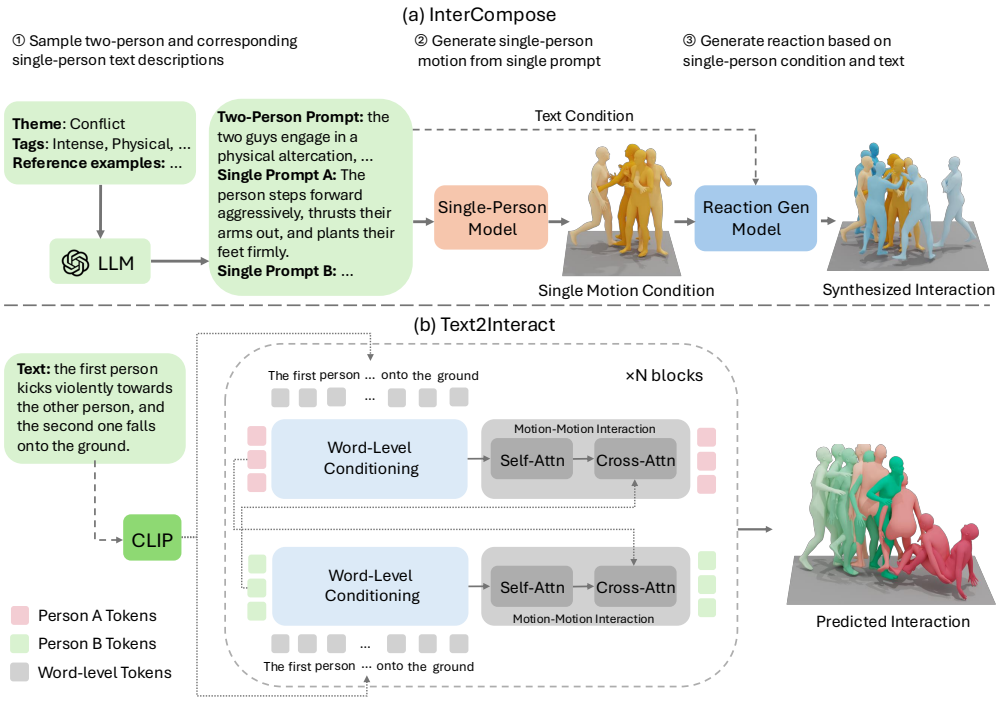

Figure 2: Overview of the proposed frameworks. (a) **InterCompose**: sample interaction and single-person descriptions via an LLM, generate a single-person motion from a motion prior (Guo et al., 2024), then compose the second agent with a reaction model conditioned on the two-person prompt and the motion prior. (b) **Text2Interact**: an N-block generator with word-level conditioning and motion–motion interaction. Each block cross-attends motion tokens to CLIP word tokens (Radford et al., 2021), followed by self-attention and inter-agent cross-attention to model individual motion and interactions.

where motion state of the i-th person at time t, $\mathbf{x}_i^{(t)}$, is defined as a collection of global joint positions $\mathbf{j}_g^p \in \mathbb{R}^{3N}$, velocities $\mathbf{j}_g^v \in \mathbb{R}^{3N}$ in the world frame, 6D representation of local rotations $\mathbf{j}^r \in \mathbb{R}^{6N}$ in the root frame, and binary foot-ground contact $\mathbf{c}_f \in \mathbb{R}^4$.

3.1 Generative Two-person Motion Composition

3.1.1 Data Generation from Single-Person Motion and Language Priors

To address the limited diversity of existing two-person motion datasets, we propose a modular pipeline, InterCompose, that synthesizes realistic two-person interactions by sampling coherent two-person and single-person motion descriptions and composing individual motion sequences generated from the descriptions. Specifically, we first use an LLM (Liu et al., 2024) to annotate the text descriptions in InterHuman (Liang et al., 2024), classifying them into a discrete space of coarse-grained *themes* (e.g. greeting, dancing, conflict) and fine-grained *tags* (e.g. excited, synchronized, disarm) that further describes the interaction. By systematically combining plausible theme–tag combinations, we can generate interaction descriptions that remain stylistically consistent with InterHuman but span a broader range of behaviors by sampling from the LLM in the joint theme-tag space: $c_t \sim \mathcal{T}_{\text{LLM}}(\text{theme}, \text{tags})$. Then, given a generated interaction text c_t, we decompose it into two role-specific sub-descriptions (c_t^1, c_t^2) using an additional LLM prompt. Each c_t^i describes the motion of person i independent of the other, while taking the context information into account. Please refer to Fig. 2 (a) for an illustration. We use (c_t^1, c_t^2) to generate corresponding single-person motions $(\mathbf{x}_1, \mathbf{x}_2)$ via a pre-trained single-person text-to-motion generator MoMask (Guo et al., 2024) trained on single-person motion datasets (Guo et al., 2022), enabling it to generate motions beyond the single-person motion distribution of InterHuman (Liang et al., 2024).

To model dependencies between the interactants, we train a conditional diffusion model $\mathcal{D}_\theta$ that synthesizes the second agent's motion $\mathbf{x}_2$ given the first agent's motion $\mathbf{x}_1$ and the shared interaction description c_t. Formally, we model the conditional distribution $p_\theta(\mathbf{x}_2 \mid \mathbf{x}_1, c_t)$ using a denoising diffusion probabilistic model (DDPM) with an 8-layer Transformer architecture. At training time,

$\mathcal{D}_\theta$ aims to recover an interaction sequence $(\mathbf{x}_1, \mathbf{x}_2)$ sampled from InterHuman (Liang et al., 2024) from one ground-truth and one noised interactant $(\mathbf{x}_1, \mathbf{x}_2')$. During inference, we sample $\mathbf{x}_1$ using MoMask then generate $\mathbf{x}_2$ using $\mathcal{D}_\theta$ conditioned on $\mathbf{x}_1$, producing a complete interaction $(\mathbf{x}_1, \mathbf{x}_2)$ that is semantically aligned with c_t and physically coordinated.

This compositional approach significantly enlarges the diversity of two-person interactions compared to existing datasets, as it decouples single-person motion priors and recombines them under guided conditions. Unlike direct generation approaches, which must learn joint coordination from sparse data, our formulation leverages both rich single-person priors and role-specific semantics to scaffold plausible and varied interactions from structured textual prompts. In addition, the inference-based nature of our data composition process allows it to be extremely scalable and cost-efficient compared to the traditional MoCap-based data collection process.

3.1.2 High-Quality and Diverse Data Filtering.

To ensure the quality and diversity of the synthesized two-person motions, we propose a two-stage filtering pipeline that considers text-motion alignment and distributional regularization. We first train a contrastive encoder using the InterHuman (Liang et al., 2024) two-person interaction dataset to project both text and motion into a shared embedding space. Specifically, we freeze a pretrained text encoder (CLIP (Radford et al., 2021)) with a trainable Transformer (Vaswani et al., 2017) feature extractor head f_{head}, and learn a motion encoder f_ϕ based on the Transformer architecture. The training objective is a symmetric cross-entropy (CE) loss over cosine similarities between normalized embeddings. A held-out subset of the InterHuman dataset is reserved to provide a reference embedding distribution for diversity filtering.

After training, we apply the encoder to the synthetic dataset $\mathcal{D}_{\text{syn}} = \{(\mathbf{x}, c_t)\}$ and compute the cosine similarity between each motion and its paired text. We discard samples with similarity scores below a threshold $\delta = 0.58$, empirically chosen based on performance on a validation split. This step eliminates low-quality or semantically misaligned samples.

To further enforce motion diversity and promote high-quality samples that are underrepresented in the original two-person dataset, we perform a distributional filtering step using the two-person motion embeddings from the held-out InterHuman subset $\mathcal{E}_{\text{real}} = \{f_\phi(\mathbf{x}_r)\}$ as reference. For each synthetic motion embedding $f_\phi(\mathbf{x})$, we compute its Euclidean distance to the k nearest neighbors in $\mathcal{E}_{\text{real}}$, and retain only those whose average distance falls within a predefined annulus: $r_{\text{min}} \leq d(f_\phi(\mathbf{x}), \mathcal{E}_{\text{real}}) \leq r_{\text{max}}$. This preserves synthesized motions that are novel (outside the inner radius r_{min}) but not far from the real data distribution (inside the outer radius r_{max}).

This dual-stage filtering framework ensures that the final synthetic dataset exhibits both semantic fidelity and distributional diversity. Detailed analysis of the effects of δ, r_{min}, and r_{max} is in Sec. 4.3.

3.2 Fine-grained Interaction Modeling

3.2.1 Word-level Attention Modeling of Language and Interaction Dynamics

Having diversified our training distribution with synthetic data, we now address the issue of insufficient granularity in two-person text semantics modeling. To tackle the issue and improve semantic alignment between natural language and generated motion, we design a cross-attention-based word-level text-motion conditioning architecture that injects fine-grained text information throughout the generation process. Unlike prior methods that inject a sentence-level embedding into motion tokens via AdaLN (Liang et al., 2024; Javed et al., 2024; Ponce et al., 2024) or sentence-level cross-attention (Tanaka & Fujiwara, 2023), our architecture allows each motion token to dynamically attend to individual word-level tokens, preserving the nuanced motion semantics and spatial-temporal alignment cues in semantic-rich interaction prompts.

Formally, given a tokenized interaction description $c_t = \{w_1, \ldots, w_L\}$, we extract word-level embeddings $\mathbf{T} = \{\mathbf{t}^{(1)}, \ldots, \mathbf{t}^{(L)}\}$ using a frozen CLIP text encoder. The architecture is composed of alternating processing modules, each consisting of two types: 1) **Word-level Conditioning Module $\mathcal{M}_w$:** A Transformer block with cross-attention between a single agent's motion tokens $\mathbf{x}_i$ and the full text embedding sequence $\mathbf{T}$, enabling each motion token to focus on semantically relevant parts of the prompt. This block preserves temporal resolution and injects lexical cues aligned with event structure.

2) **Motion-Motion Interaction Module $\mathcal{M}_m$:** A two-stage module where motion tokens $\mathbf{x}_i$ first perform self-attention over their own sequence (intra-agent context), followed by cross-attention over the other agent's motion tokens $\mathbf{x}_j$ ($j \neq i$), which models inter-agent physical and temporal dependencies such as push-pull or synchronization. Please see Fig. 2 (b) for an illustration.

Each update step consists of a word-level conditioning module followed by a motion-motion interaction module; these two modules together form a full block that is applied in an alternating fashion: first to one agent, conditioning on the text and the other agent's motion, and then to the other agent in the next step. Leveraging the symmetry of two-person interactions, the blocks $\mathcal{B}_w$ and $\mathcal{B}_m$ are shared across agents, ensuring architectural symmetry and parameter efficiency. The alternating structure allows each agent to respond adaptively to both the linguistic description and the dynamic behavior of their partner, while preserving causal and temporal coherence.

Overall, the network design enables high-fidelity generation that is both semantically grounded and interaction-aware, allowing nuanced conditioning through the word-level representation and fostering motion patterns that are faithful to the described scenario.

3.2.2 Adaptive Interaction Supervision

We use the standard velocity loss $\mathcal{L}_{\text{vel}}$, foot contact loss $\mathcal{L}_{\text{foot}}$, bone-length loss $\mathcal{L}_{\text{BL}}$, and relative orientation loss $\mathcal{L}_{\text{RO}}$. For these objective functions, refer to InterGen (Liang et al., 2024) for details.

In addition to the above objective functions, we designed a new objective $\mathcal{L}_{\text{AdaInteract}}$ to enhance the generation of plausible interaction semantics, a crucial element of text-to-interaction generation. Motivated by the insight that joint pairs that are closer to each other carry more importance in the interaction semantics, we propose a novel adaptive interaction loss that supervises the pairwise distances between human-human joint pairs with an adaptive weighting:

$$\mathcal{L}_{\text{AdaInteract}} = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{1}{d_{ij} + \epsilon} \| d_{ij} - \hat{d}_{ij} \|_2 \tag{1}$$

Where d_{ij}, $\hat{d}_{ij}$ are the ground-truth and predicted distances between the joints i and j respectively, and $\epsilon = 0.1$ is an empirically set constant. By putting more emphasis on spatially proximate inter-agent joint pairs, our adaptive interaction objective function provides strong guidance for the model to adhere to the interaction semantics.

4 Experiments

4.1 Experimental Setup

Dataset. We use the InterHuman (Liang et al., 2024) dataset for training and evaluating our model. InterHuman contains 6,022 two-person interacting motions and 3 textural descriptions per motion in the training split, and 1,177 two-person interacting motions in the test split. Additionally, a synthesized dataset of 25,000 text-motion pairs before filtering and 1,200 text-motion pairs after filtering is used for fine-tuning. All models are first trained on the InterHuman training split. For fine-tuning, the model is fine-tuned on the InterHuman training split augmented by the filtered synthetic dataset. All metrics are calculated using the InterHuman test split.

Metrics. Following standard practice in human-human interaction generation (Liang et al., 2024; Ruiz-Ponce et al., 2024; Javed et al., 2024; Cai et al., 2024), we use the R-Precision (Top-1, 2, 3), Frechet Inception Distance (FID), Multimodal Distance (MM Dist), Diversity, and Multimodality (MModality) for evaluation our models. Please refer to InterGen (Liang et al., 2024) for the detailed definition of these metrics.

Implementation Details. Our model consists of 12 attention blocks and 12 word-level conditioning blocks, positioned in an interleaved manner. We utilize a frozen CLIP-ViT-L/14 (Radford et al., 2021) model for extracting as the text encoder. We set the number of diffusion (Ho et al., 2020) steps to 1,000 and use a cosine noise schedule (Nichol & Dhariwal, 2021). The model is trained with 8 NVIDIA A100 GPUs for 200,000 steps, with a 5e-5 learning rate and a batch size of 16 with the AdamW (Loshchilov & Hutter, 2017) optimizer, cosine learning rate scheduling, and 1000-step

Method	R Precision↑			FID↓	MM Dist↓	Diversity→	MModality↑
	Top 1	Top 2	Top 3				
Ground Truth	$0.452^{\pm.008}$	$0.610^{\pm.009}$	$0.701^{\pm.008}$	$0.273^{\pm.007}$	$3.755^{\pm.008}$	$7.948^{\pm.064}$	-
T2M (Guo et al., 2022)	$0.238^{\pm.012}$	$0.325^{\pm.010}$	$0.464^{\pm.014}$	$13.769^{\pm.072}$	$5.731^{\pm.013}$	$7.046^{\pm.022}$	$1.387^{\pm.076}$
MDM (Tevet et al., 2022b)	$0.153^{\pm.012}$	$0.260^{\pm.009}$	$0.339^{\pm.012}$	$9.167^{\pm.056}$	$7.125^{\pm.018}$	$7.602^{\pm.045}$	$\mathbf{2.350}^{\pm.080}$
ComMDM (Shafir et al., 2023)	$0.223^{\pm.009}$	$0.334^{\pm.008}$	$0.466^{\pm.010}$	$7.069^{\pm.054}$	$6.212^{\pm.021}$	$7.244^{\pm.038}$	$1.822^{\pm.052}$
RIG (Tanaka & Fujiwara, 2023)	$0.285^{\pm.010}$	$0.409^{\pm.014}$	$0.521^{\pm.013}$	$6.775^{\pm.069}$	$5.876^{\pm.002}$	$7.311^{\pm.043}$	$2.096^{\pm.065}$
InterGen (Liang et al., 2024)	$0.371^{\pm.010}$	$0.515^{\pm.012}$	$0.624^{\pm.010}$	$5.918^{\pm.079}$	$5.108^{\pm.014}$	$7.387^{\pm.029}$	$2.141^{\pm.063}$
MoMat-MoGen (Cai et al., 2024)	$0.449^{\pm.004}$	$0.591^{\pm.003}$	$0.666^{\pm.004}$	$5.674^{\pm.085}$	$3.790^{\pm.001}$	$8.021^{\pm.350}$	$1.295^{\pm.023}$
in2IN (Ruiz-Ponce et al., 2024)	$0.425^{\pm.008}$	$0.576^{\pm.008}$	$0.662^{\pm.009}$	$5.535^{\pm.120}$	$3.803^{\pm.002}$	$7.953^{\pm.047}$	$1.215^{\pm.023}$
InterMask (Javed et al., 2024)	$0.449^{\pm.004}$	$0.599^{\pm.005}$	$0.683^{\pm.004}$	$\mathbf{5.154}^{\pm.061}$	$3.790^{\pm.002}$	$\mathbf{7.944}^{\pm.033}$	$1.737^{\pm.020}$
Ours	$\mathbf{0.483}^{\pm.005}$	$\mathbf{0.638}^{\pm.005}$	$\mathbf{0.717}^{\pm.005}$	$5.191^{\pm.055}$	$\mathbf{3.778}^{\pm.001}$	$7.900^{\pm.030}$	$1.051^{\pm.031}$

Table 1: Performance on the InterHuman (Liang et al., 2024) test sets. $\pm$ indicates a 95% confidence interval and $\rightarrow$ means the closer to ground truth the better. Boldface indicates the best result.

warm-up. During sampling, we use the DDIM (Song et al., 2020) sampling with 50 timesteps, with a classifier-free guidance (Ho & Salimans, 2022) weight of 3.5.

4.2 Comparison with the State-of-the-arts

Quantitative Comparison. Tab. 1 contains the quantitative comparison between Text2Interact and state-of-the-art methods. Each experiment is repeated 20 times, after which the mean and 95% confidence interval of each metric is recorded. Text2Interact achieves state-of-the-art results on all three R-precision metrics, surpassing the previous state-of-the-art, InterMask (Javed et al., 2024), by a significant margin, highlighting the effectiveness of the word-level conditioning design choice in text-motion alignment. In terms of motion quality, our Text2Interact also achieves the best MM Distance and the second-best FID, with a small FID margin (0.037) from the state-of-the-art, InterMask, and surpassing all other prior arts. Notably, our FID is within the 95% confidence interval of InterMask's FID, highlighting an equal level of generation quality from a statistical perspective.

Figure 3: Qualitative comparisons of interaction generation results from Text2Interact and InterMask (Javed et al., 2024). Our method produces results with better text-motion alignment and is more robust to implausible poses. A deeper color indicates a later time.

Qualitative Comparison. We provide a qualitative comparison between our method and the current state-of-the-art, InterMask (Javed et al., 2024). As shown in Fig. 3, our model exhibits stronger adherence to text, higher robustness, and more plausible interaction semantics.

Specifically, in the first generation result of our method, the agent marked in red successfully pulls the agent marked in green, and the red agent emerges as the winner (frame 5). In contrast, the motion generated by InterMask only exhibits pulling, and does not reflect this final part of the motion. In the second row, the InterMask result exhibits implausible human pose outputs in frames 1 and 2, and does not reflect the final bowing action, while our model generates plausible results faithful to the complete semantic meanings of the text. In the third row, the kneeling human generated by InterMask

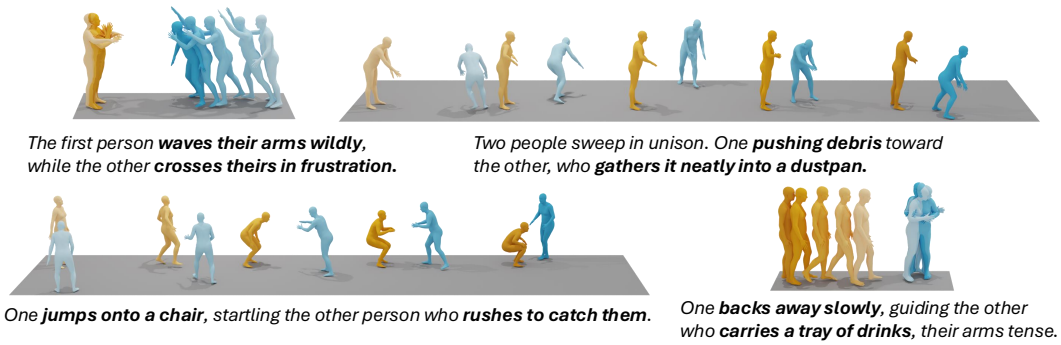

Figure 4: Qualitative samples of InterCompose. Prompts are synthesized by an LLM (Liu et al., 2024). The yellow is synthesized by the single-person motion generator, while the blue is generated by the reaction model with the yellow as the condition. A deeper color indicates a later time.

again exhibits an implausible human pose in frames 2, 3, 4, and 5. These results highlight our Text2Interact's pose robustness over InterMask, an aspect not adequately measured by the evaluator and the FID metric, while confirming Text2Interact's lead in text to motion alignment.

4.3 Further Evaluation

Method	R Precision↑			FID↓	MM Dist↓	Diversity→	MModality↑
	Top 1	Top 2	Top 3				
Before Fine-tuning	$0.485^{\pm.010}$	$0.644^{\pm.007}$	$0.721^{\pm.009}$	$5.701^{\pm.065}$	$3.777^{\pm.001}$	$7.904^{\pm.033}$	$1.081^{\pm.019}$
Finetune ($0.25 < d < 0.6$)	$0.485^{\pm.004}$	$0.641^{\pm.004}$	$0.717^{\pm.004}$	$5.981^{\pm.056}$	$3.778^{\pm.001}$	$7.946^{\pm.028}$	$1.080^{\pm.026}$
Finetune ($0.3 < d < 0.6$)	$0.480^{\pm.007}$	$0.635^{\pm.004}$	$0.715^{\pm.004}$	$5.682^{\pm.100}$	$3.779^{\pm.002}$	$7.909^{\pm.030}$	$1.058^{\pm.030}$
Finetune ($0.35 < d < 0.6$)	$0.483^{\pm.005}$	$0.638^{\pm.005}$	$0.717^{\pm.005}$	$5.191^{\pm.055}$	$3.778^{\pm.001}$	$7.900^{\pm.030}$	$1.051^{\pm.031}$

Table 2: Quantitative Results of Text2Interact after fine-tuning on synthetic data generated by InterCompose. d denotes the Euclidean distance between a synthetic data sample point and its closest held-out data point in the embedding space of the neural evaluator.

Quantitative Evaluation of Fine-Tuning and Filtering. We present the results of fine-tuning the model on a combined dataset consisting of InterHuman (Liang et al., 2024) training split and filtered synthetic data for 50k steps, with a learning rate of 5e-6. As shown in Tab. 2, the model exhibits a similar level of text-to-motion matching (R-Precision) after fine-tuning and a significantly improved FID in the best case, highlighting the improvement in generalizability. Notably, the FID exhibits a clear increasing trend when the minimum Euclidean distance d of the filtering process is increased within a reasonable range, confirming the effectiveness of the proposed filtering pipeline in achieving synthetic data quality and diversity at the same time. The results also indicates that the increased dataset diversity by synthetic data improves the model's generalizability.

Qualitative Visualization of Synthetic Data. In Fig. 4, we present exemplar results of our data synthesis pipeline, InterCompose. The agent marked yellow is generated by the single-person motion generator (Guo et al., 2024) while the agent marked blue is generated by the reaction generation model. As shown in the figure, our pipeline synthesizes high-quality and diverse motions from single-person motion and text descriptions, with close adherence to the text. Moreover, the textual descriptions resemble real-life human-human interaction situations with emotional interactions or inter-person collaboration instilled into the text prompts.

User Study Results on Fine-Tuning. We present the user preference study results conducted with 51 participants on 10 samples generated with out-of-distribution texts using our data generation pipeline. Fig. 5 shows the users' strong preference for the model after fine-tuning, in terms of both motion quality and text-motion matching. This result confirms our model's improved generalizability to out-of-distribution samples after fine-tuning.

User Study Results on InterCompose vs Text2Interact. Fig. 6 presents a user preference study between motions synthesized with InterCompose and generated by Text2Interact. Two studies are

conducted for InterCompose (a) without and (b) with distributional filtering, where only high-quality and novel motions are retained after the filtering process. The results show that: (a) Text2Interact shows stronger consistency in motion quality compared to InterCompose, rendering the former suitable for general text-to-interaction tasks; (b) with the distributional filtering step, motions from InterCompose have higher quality compared to motions generated by Text2Interact, confirming the quality of the synthesized and filtered motion dataset used for fine-tuning.

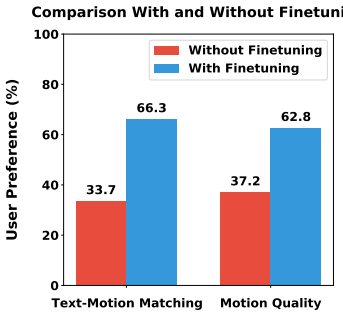

Figure 5: User preference study results of Text2Interact with and without fine-tuning on synthetic data.

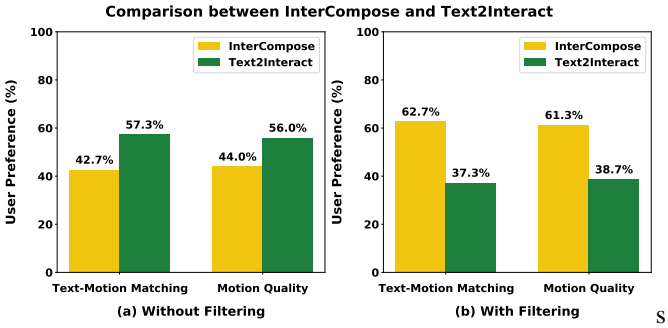

Figure 6: Comparison of motion generation results using InterCompose and Text2Interact, (a) without filtering, and (b) with filtering. The motion quality and text-motion matching of InterCompose surpass Text2Interact only after filtering.

Ablation Study. After thoroughly investigating the effectiveness of the data synthesis and fine-tuning pipeline, we now analyze the effectiveness of the proposed sub-components: the word-level conditioning module (WLC), the adaptive interaction loss (AIL), and the synthetic data fine-tuning process (FT). When removing the word-level conditioning module, we replace it with sentence-level condition injection by AdaLN; for ablation of adaptive interaction loss, we replace it with the flat-weight distance map function $\mathcal{L}_{DM}$ proposed by InterGen (Liang et al., 2024); for ablation of fine-tuning, the model was trained only on InterHuman (Liang et al., 2024) without the fine-tuning process. Tab. 3 shows significant improvement of our model after adding each proposed component, in terms of text-motion matching (R-Precision), FID, Multimodal Distance, and diversity.

Method	R Precision↑			FID↓	MM Dist↓	Diversity→	MModality↑
	Top 1	Top 2	Top 3				
w.o. All Proposed Components	$0.441^{\pm.006}$	$0.608^{\pm.005}$	$0.681^{\pm.005}$	$6.237^{\pm.071}$	$3.781^{\pm.001}$	$7.959^{\pm.035}$	$1.068^{\pm.022}$
w.o. AIL, FT	$0.484^{\pm.005}$	$0.632^{\pm.005}$	$0.710^{\pm.005}$	$6.192^{\pm.069}$	$3.779^{\pm.001}$	$7.853^{\pm.033}$	$1.081^{\pm.019}$
w.o. WLC, FT	$0.484^{\pm.005}$	$0.629^{\pm.005}$	$0.711^{\pm.005}$	$5.877^{\pm.061}$	$3.779^{\pm.001}$	$7.851^{\pm.034}$	$0.996^{\pm.027}$
w.o. FT	$0.485^{\pm.010}$	$0.644^{\pm.007}$	$0.721^{\pm.009}$	$5.701^{\pm.065}$	$3.777^{\pm.001}$	$7.904^{\pm.033}$	$1.046^{\pm.022}$
Ours	$0.483^{\pm.005}$	$0.638^{\pm.005}$	$0.717^{\pm.005}$	$5.191^{\pm.055}$	$3.778^{\pm.001}$	$7.900^{\pm.030}$	$1.051^{\pm.031}$

Table 3: Ablation Study: Effect of removing one or more of the proposed components: Adaptive Interaction Loss (AIL), Synthetic Data Fine-Tuning (FT), Word-Level Conditioning (WLC).

5 Conclusion

In this paper, we presented InterCompose, a novel and effective framework that composes single-person motions into two-person interaction from LLM-generated text descriptions, and Text2Interact, a high-quality and fine-grained two-person interaction generation framework equipped with word-level conditioning. The effectiveness of InterCompose has been confirmed by an ablation study, user study, qualitative results, and latent visualizations. Utilizing data generated by InterCompose, Text2Interact achieves a significant FID boost, achieving SoTA-level R-precision and FID, setting a new state-of-the-art for the two-person motion generation task.

Limitations and Future Work. While Text2Interact demonstrates strong motion fidelity and faithfulness, it does not account for physical plausibility during generation, which can result in artifacts such as floating motions and ground penetration. Incorporating physics priors offers a promising avenue for future work. Additionally, although InterCompose provides an effective synthesis framework, extending it to learn motions directly from video remains an interesting and relatively unexplored direction.

REFERENCES

Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5915–5920. IEEE, 2018.

Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International conference on 3D vision (3DV)*, pp. 719–728. IEEE, 2019.

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*, pp. 582–592, 2024.

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*, pp. 18000–18010, 2023.

Peishan Cong, Ziyi Wang, Zhiyang Dou, Yiming Ren, Wei Yin, Kai Cheng, Yujing Sun, Xiaoxiao Long, Xinge Zhu, and Yuexin Ma. Laserhuman: Language-guided scene-aware human motion generation in free environment. *arXiv preprint arXiv:2403.13307*, 2024.

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.

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*, pp. 5152–5161, 2022.

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*, pp. 1900–1910, 2024.

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

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

Yiming Huang, Weilin Wan, Yue Yang, Chris Callison-Burch, Mark Yatskar, and Lingjie Liu. Como: Controllable motion generation through language guided pose code editing. In *European Conference on Computer Vision*, pp. 180–196. Springer, 2024.

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.

Jaewoo Jeong, Daehee Park, and Kuk-Jin Yoon. Multi-agent long-term 3d human pose forecasting via interaction-aware trajectory conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1617–1628, 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.

Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2151–2162, 2023.

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.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Shunlin Lu, Jingbo Wang, Zeyu Lu, Ling-Hao Chen, Wenxun Dai, Junting Dong, Zhiyang Dou, Bo Dai, and Ruimao Zhang. Scamo: Exploring the scaling law in autoregressive motion generation model. *CVPR 2025*, 2024.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pp. 480–497. Springer, 2022.

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.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1941–1951, 2024.

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

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Mikihiro Tanaka and Kent Fujiwara. Role-aware interaction generation from textual description. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15999–16009, 2023.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9601–9611, 2023.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pp. 358–374. Springer, 2022a.

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

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.

Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. Tlcontrol: Trajectory and language control for human motion synthesis. In *ECCV 2024*. 2024.

Yabiao Wang, Shuo Wang, Jiangning Zhang, Ke Fan, Jiafu Wu, Zhengkai Jiang, and Yong Liu. Temporal and interactive modeling for efficient human-human motion generation. *arXiv preprint arXiv:2408.17135*, 2024.

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.

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, 2023a.

Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 364–373, 2023b.

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 *ECCV 2024*. 2024.

SUPPLEMENTARY MATERIALS

A TEXT2INTERACT IMPLEMENTATION DETAILS

A.1 WORD-LEVEL TOKENIZATION

We use the CLIP (Radford et al., 2021) ViT-L/14 encoder for encoding the text. The text is tokenized by the CLIP tokenizer into word-level tokens for short words and sub-word-level tokens for long words, with <SOT> and <EOT> tokens inserted at the start and end of the text. The maximum number of text tokens is 75. If the number of tokens after tokenization is longer than 75, the text tokens are truncated and additional tokens are discarded.

A.2 WORD-LEVEL CONDITIONING

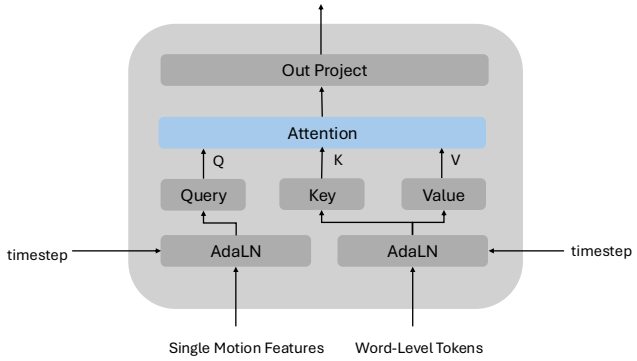

Figure 7: Illustration of the **Word-Level Conditioning** Block

Fig. 7 illustrates the details of the proposed Word-Level Conditioning block. One of the interacting agents' motion features (Single Motion Features) is passed through an Adaptive Layer Norm (AdaLN) for the injection of timestep information. The word-level tokens are passed through a separate AdaLN of the same structure but with different parameters. Then, the normalized and modulated features are passed through the linear layers to yield the query, key, and value tensors, where the query comes from the single motion features and the key and value come from the word-level tokens. Then, an attention output embedding is obtained using query Q, key K, and value V with the attention mechanism (Vaswani et al., 2017):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2}$$

Finally, the output of the attention layer is passed through a linear layer for the reprojection and mixing of the attention head outputs.

A.3 MOTION-MOTION INTERACTION

A.3.1 SELF-ATTENTION

The Self-Attention module in the Motion-Motion Interaction block is responsible for the processing of one of the interacting agents' motion features. Fig. 8 is an illustration of this. The motion features are first modulated by an Adaptive Layer Norm (AdaLN) block, which injects timestep information by scaling the features with mean and variance determined by the timestep. Then, projection layers calculate the query, key, and value tensors separately, which are used to calculate the attention output using the attention mechanism (Vaswani et al., 2017). Finally, the attention output is projected with a linear output projection layer to give the final output.

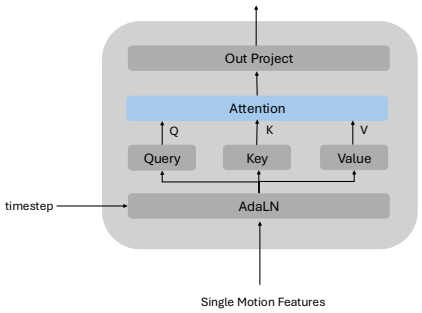

Figure 8: Illustration of the **Self-Attention** module in the Motion-Motion Interaction Block.

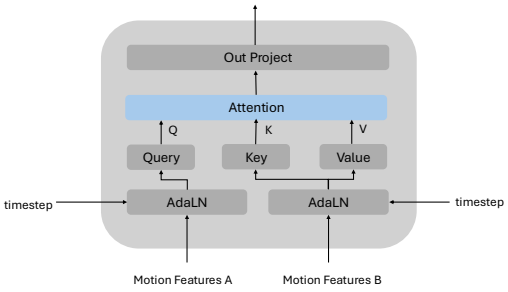

Figure 9: Illustration of the **Cross-Attention** module in the Motion-Motion Interaction Block.

A.3.2 Cross-Attention

The Cross-Attention module in the Motion-Motion Interaction block is responsible for modeling the inter-agent interaction. Fig. 9 provides an illustration. In the Cross-Attention module, the motion features of agent A and agent B (vice versa) are used to calculate features using separate AdaLN blocks for the timestep information. Then, the motion features of agent A are passed through the query projection layer to give the query features. The motion features of agent B are passed through the key and value projection layer to calculate the key and value features. The query, key, and value features undergo an attention mechanism to obtain the attention output feature, which is subsequently projected by an linear output layer to form the final output.

B InterCompose Implementation Details

B.1 Two Person Prompt Synthesis

We use a prompt template for synthesizing diverse and plausible two-person interaction descriptions from an LLM (Liu et al., 2024) based on coarse-grained themes and fine-grained tags, along with real examples from the InterHuman (Liang et al., 2024) dataset for styling reference. The complete template is provided below in Fig. 10.

B.2 Single Person Prompt Synthesis

After obtaining two-person descriptions, we use a separate prompt template to synthesize pairs of single-person descriptions that are self-contained, coherent, and consistent with the given two-person descriptions. The LLM infers the corresponding single-person motion according to the provided two-person interaction information, while reasoning the plausible single-person motion if the two-person prompt does not provide complete information. The complete prompt template is given in Fig. 11. Examples of two-person prompts and corresponding single-person prompts are given in Fig. 12.

B.3 Single Person Motion Generation

We use MoMask (Guo et al., 2024) to generate the single-person motion conditions for the subsequent reaction generation. In addition, we trained a length estimator on the InterHuman (Liang et al., 2024) text-motion pairs to estimate the correct length of the corresponding motion given a two-person motion prompt, and used the predicted length by the estimator to guide the single-person motion generation.

For each two-person prompt, we use the LLM (Liu et al., 2024) to provide two prompts for the two interactants and generate two single-person motions, one with each prompt.

```
You write compact, vivid descriptions of **two-person interactions**.

Each output sentence MUST:
• mention exactly two unnamed people ("one person. . .  the other person. . . "),
• focus on body / arms / legs (ignore faces / fingers / appearance),
• be <=25 words,
• clearly match the given *Theme* and *Tags*,
• be entirely different from the examples.


Theme: {theme}
Tags : {tags}
Reference examples (k):
{example1}
{example2}
. . .
{examplek}

Now craft {m} brand-new descriptions.  Return **only** a JSON array of
strings.
```

Figure 10: Prompt template for generating two-person interaction descriptions.

```
Given the following description of a two-person interaction:

{two-person text}

Independently describe the motion of each person involved, using only
information implied by the full interaction. Do not mention or refer to the
other person in either description.  Focus only on body, arms, and legs —
ignore facial expressions, fingers, or appearance.

Use "the person" to refer to each.  Assume shared context (e.g., dancing,
greeting, arguing), but isolate each description.

Output JSON in this exact format:
  {{"1": {{"person1": "{description1}", "person2": "{description2}"}}}}

Each description must be one sentence, <=15 words, specific, and motion-focused
with relevant context.
```

Figure 11: Prompt template for generating single-person interaction descriptions.

B.4 Two Person Motion Composition

We trained a reaction generation network that uses a given sequence of joints as condition $\mathbf{x}_{cond} \in \mathbb{R}^{T\times 22\times 3}$ to generate the reaction $\mathbf{x} \in \mathbb{R}^{T\times 262}$, consisting of the full interaction in the complete InterGen (Liang et al., 2024) joint representation.

The network is trained on the InterHuman (Liang et al., 2024) training dataset with one person's joints not noised and all other terms noised to simulate the reaction generation tasks. At test time, the condition joints are provided at each time step of denoising and after the final denoising step.

Example 1.
two-person text: One person leans back, arms outstretched, while the other steps forward, pressing their chest lightly against the first's, hands resting on their hips.
single-person text A: The person leans back with arms outstretched.
single-person text B: The person steps forward, chest pressed lightly, hands on hips.

Example 2.
two-person text: One person claps twice, and the other responds by jumping in place, their legs kicking out wildly with excitement.
single-person text A: The person raises both arms and brings hands together sharply twice.
single-person text B: The person leaps upward, legs swinging outward vigorously.

Example 3.
two-person text: One person lunges with a punch, the other person blocks with crossed arms and counters with a swift kick to the thigh.
single-person text A: The person steps forward, extending one arm sharply in a punching motion.
single-person text B: The person raises both arms to cross in front, then swings one leg outward quickly.

Example 4.
two-person text: one person steps forward aggressively, arms raised, while the other person backs away, hands outstretched to resist the advancing confrontation.
single-person text A: The person steps forward aggressively with arms raised.
single-person text B: The person backs away with hands outstretched to resist.

Example 5.
two-person text: One person stumbles backward from alcohol, and the other person swiftly wraps an arm around their waist to steady them.
single-person text A: The person stumbles backward, legs unsteady from alcohol.
single-person text B: The person moves an arm quickly to wrap around a waist.

Example 6.
two-person text: One person shoves the other's shoulder, causing them to stagger, then crosses their arms in defiance as the other retreats.
single-person text A: The person extends their arm sharply, then pulls it back and crosses both arms tightly.
single-person text B: The person stumbles backward from a sudden force, then turns away while stepping back.

Figure 12: Examples of generated and two-person interaction descriptions with corresponding single-person descriptions.

B.5 Motion Filtering

Before filtering, we first use the trained neural motion evaluator to project all generated motions to the 512-dimensional motion latent space, to compare with the latents of a 500-sample held-out motion dataset. The motion filtering step consists of a k-nearest neighbors filtering with a maximum 20 nearest neighbors for each sample in the held-out set. We also calculate the distances between each of the nearest neighbors with the held-out motion sample to make sure the distance is in the

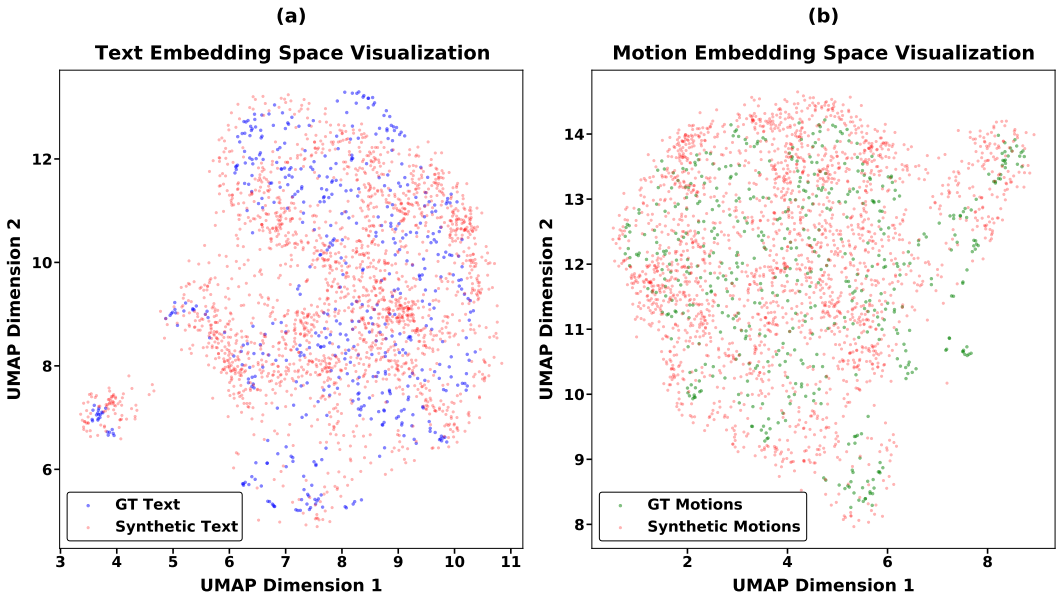

Figure 13: UMAP (McInnes et al., 2018) visualizations of evaluator and CLIP (Radford et al., 2021) embeddings of (a) text and (b) two-person motions from the InterHuman (Liang et al., 2024) held-out subset and filtered synthesized dataset.

predefined annulus $d_{min} < r < d_{max}$. Finally, we use the neural motion evaluator to filter out all the remaining motions with text-motion cosine similarity less than 0.58, an empirically set threshold.

C Motion Embedding Space Visualizations

Fig. 13 demonstrates a dimensionality-reduced visualization of the text and two-person motion embeddings of the InterHuman (Liang et al., 2024) held-out dataset, extracted from CLIP (Radford et al., 2021) and trained motion evaluator. We utilize UMAP (McInnes et al., 2018) (Uniform Manifold Approximation and Projection), a popular dimension reduction technique that preserves the local and global structure of high-dimensional data in a low-dimensional space. As shown in the figure, both the generated text (Fig. 13 (a)) and motion (Fig. 13 (b)) descriptions from our pipeline have good coverage in most high-density areas of the held-out dataset, while covering many underrepresented areas that lacks held-out data samples, highlighting our pipeline's capability in enhancing data diversity.

D User Study Details

We conducted a user study to evaluate our Text2Interact model with and without fine-tuning. Human evaluators are asked to choose between 10 pairs of two-person interaction videos and determine the one out of each pair that is more faithful to the text prompt or more natural as an interaction. Fig. 14 is an illustration of how the user study is conducted.

E Societal Impacts

Our work on Text2Interact introduces a scalable framework for high-fidelity and diverse text-to-two-person interaction generation. While the technology has the potential to significantly benefit domains such as animation, virtual reality, assistive robotics, and embodied AI, it also raises several ethical and societal considerations.

Positive Impacts. The proposed method can facilitate content creation in media, education, and human-computer interaction by automating the generation of complex, realistic human interactions.

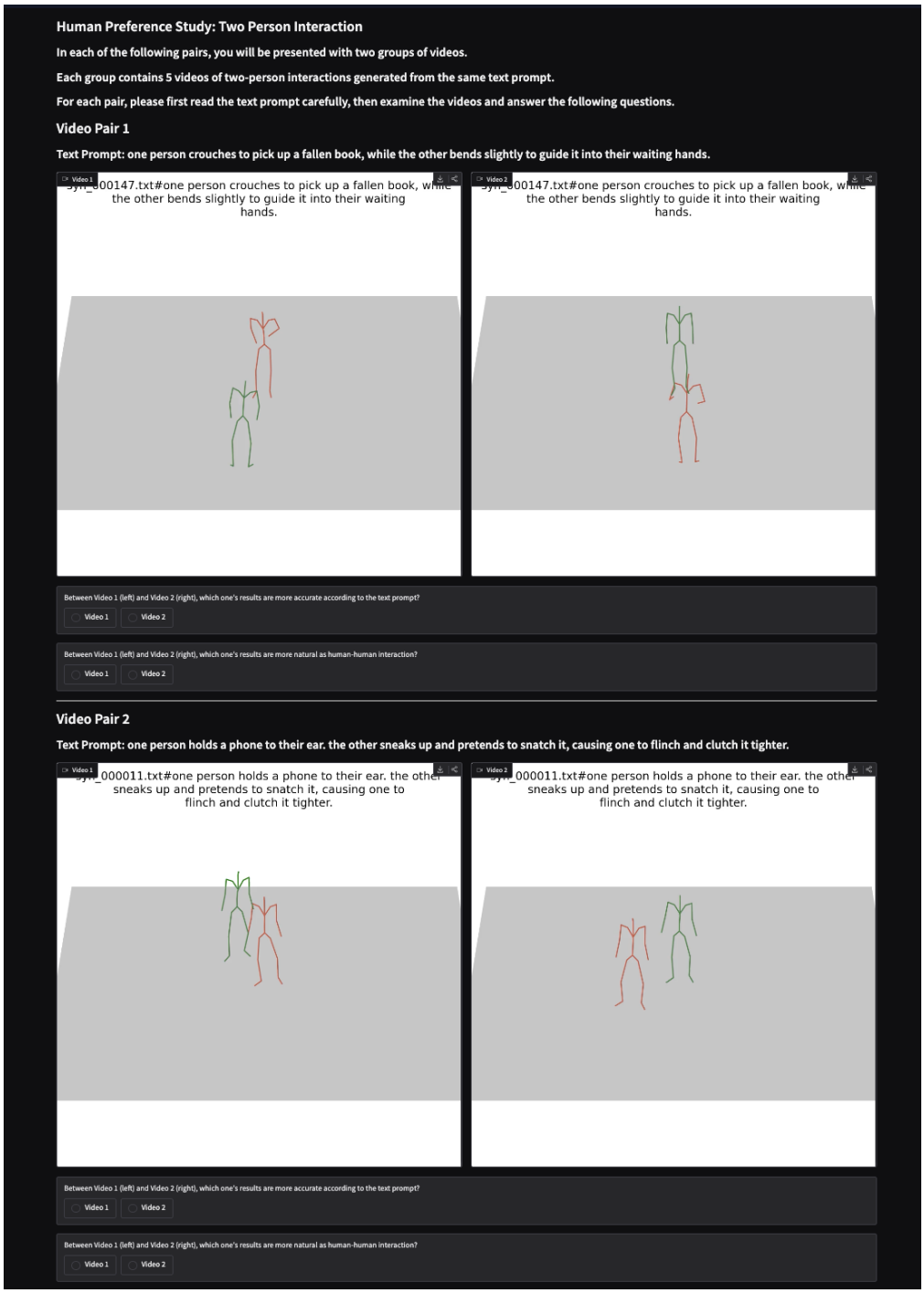

Figure 14: A screenshot of the user study

It may lower the barrier for creating high-quality motion data, particularly for under-resourced languages or motion types, and help simulate social interactions for training embodied agents or improving accessibility tools for individuals with disabilities.

Risks and Limitations. As with many generative models, there is a potential risk of misuse, such as generating deceptive or misleading content (e.g., synthetic surveillance or manipulated footage). Although our method focuses solely on body motion and excludes facial expressions or identity features, generated motion could still be used out of context or embedded in misleading visual narratives. Additionally, there is a risk of dataset bias being amplified if the single-person priors or LLM-generated text reflect culturally specific or stereotyped behaviors. We recommend future users apply careful evaluation and transparency practices when deploying this technology.

Mitigations. Our dataset curation and filtering process emphasizes diversity and alignment with real-world motion distributions to reduce representation biases. Furthermore, our model does not generate personally identifiable information or faces, and we encourage its use only in applications that respect human dignity, consent, and privacy.