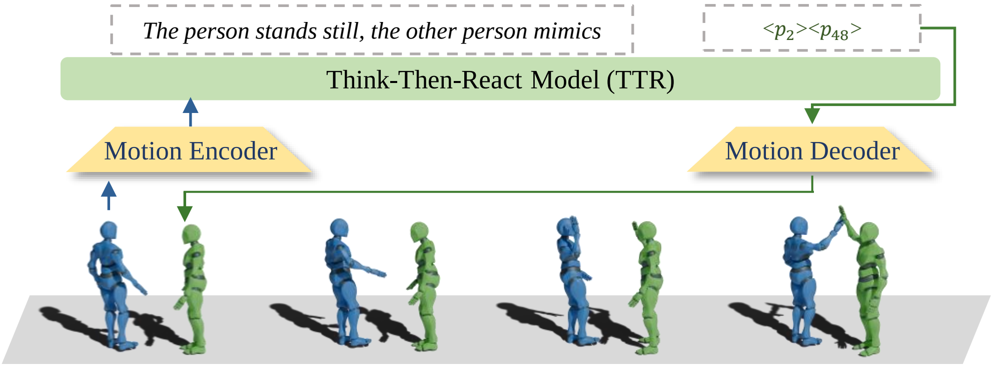
THINKING THEN REACTING: TOWARDS BETTER ACTION-TO-REACTION MOTION GENERATION

Anonymous authors
Paper under double-blind review


The person stands still, the other person mimics
<p_2><p_{48}>
Think-Then-React Model (TTR)
Motion Encoder
Motion Decoder

ABSTRACT

Generating unconstrained interactive human-like reactions in an online manner has significant applications in virtual reality, human-robot interaction, and games. Despite recent advancements in single-person motion generation, it is still challenging to well handle unconstrained action-to-reaction generation for three reasons: 1) the absence of a unified motion representation that encompasses both egocentric pose and absolute space features, 2) the instability during inference due to lack of constraints (e.g., text prompts), and 3) the insufficient utilization of text-motion training data in unconstrained scenario. To address these challenges, we introduce Think-Then-React (TTR), a large language model-based framework designed to generate online and unconstrained human-like reactions. First, we propose a unified space-pose token representation, combining both egocentric pose and absolute space features to enhance action understanding and reaction generation. Second, TTR unifies two processes during inference: a **thinking** process that infers action intentions, and a **reacting** process that predicts precise and semantically appropriate reactions based on the inferred intention. Furthermore, we introduce a multi-task training strategy, that enables us to effectively leverage both motion and text data. Extensive experiments demonstrate that TTR outperforms existing baselines, achieving significant improvements in evaluation metrics, such as reducing FID from 3.988 to 1.942.

1 INTRODUCTION

Predicting reaction motion based on human action has broad applications in areas such as AR/VR, robotics, and intelligent non-player characters in games. Recently, significant advancements have been achieved in the domain of single-person motion generation, conditioned on textual descriptions (Guo et al., 2024; 2022b; Zhang et al., 2023b) and action labels (Xu et al., 2023; Guo et al., 2020). Leveraging well-annotated human 3D motion datasets (Guo et al., 2022a), these models employ various generative frameworks, such as Diffusion Models (Ho et al., 2020; Liang et al., 2024), Variational Autoencoders (VAEs) (Kingma, 2013), and Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), to capture cross-modality distributions. Furthermore, Large Language Models (LLMs) have been applied to human motion generation, demonstrating superior performance (Jiang et al., 2023; Zhang et al., 2024). LLMs are particularly suitable for this task due to their fine-grained text understanding capabilities and the auto-regressive nature of the temporal human motion generation process.

However, generating online and unconstrained human reactions in multi-person scenarios presents a more challenging task due to three primary factors. First, unlike single-person motion representation that solely focuses on egocentric pose, representing human motions in multi-person scenarios necessitates both **egocentric pose** and **absolute space** information. Second, in real-world scenarios, ground-truth textual prompts or action labels are often unavailable. Therefore, models must accurately understand human actions to predict precise reactions that correspond both spatially and temporally. Third, the unconstrained setting also limits effective utilization of text data during training, as text data are not provided during inference.

Several works have focused on the human-human interaction domain. For instance, InterFormer (Chopin et al., 2023) proposes an encoder-decoder architecture for reaction generation, injecting human skeleton priors into attention layers for effective spatial modeling. InterGen (Liang et al., 2024) introduces a mutual attention mechanism within diffusion process for joint action-reaction generation. However, these methods are not directly applicable to real-world applications. They do not consider reaction generation as a online process, where the generated motions should be real-time responses to actions, and they often require auxiliary prompts to condition the generation process. ReGenNet (Xu et al., 2024), which is most similar to our approach, acknowledges the online and unconstrained nature of reaction generation, proposing a transformer decoder-based diffusion model for online reaction generation. It observes that explicitly given the action's intention as a condition, the model can achieve superior performance compared to unconstrained settings, highlighting the necessity of understanding interaction semantics for reaction generation. However, ReGenNet directly models action-to-reaction generation process, without explicitly inferring action intention, leading to subpar performance.

To address these challenges, we propose Think-Then-React (TTR), an LLM-based model designed to predict human reactions in **online** and **unconstrained** settings with the following innovations: First, to address the mismatch between egocentric pose and absolute space representation, we propose a space-pose token representation that effectively unifies motion representations for action and reaction. Specifically, we train a VQ-VAE (Van Den Oord et al., 2017) to encode human pose sequences into discrete tokens, which are readable by LLMs. Before encoding, human poses are spatially normalized to ensure high-quality and fidelity, i.e., the encoded tokens only represent egocentric poses, and are agnostic to absolute space information. To maintain spatial features which are crucial in multi-person interaction scenarios, we propose a space tokenizer that encodes 2D positions and human body orientations in the world frame into discrete space tokens. We then concatenate initial space tokens as prefixes to pose sequences, indicating the initial absolute state before an egocentric motion. Second, to tackle the unconstrained setting, we introduce a novel framework that enables the model to infer the action's intention, which can prompt subsequent reaction generation. Specifically, TTR unifies two processes within one model: a **thinking** process that infers the intent and interaction description from action motion, analogous to the cerebral cortex in the brain, and a **reacting** process that takes both the action motion and inferred description as input, to generate precise and semantically appropriate reactions, akin to the cerebellum. Third, to effectively leverage both motion and text data, we design a multi-task training pipeline consisting of motion-text, space-pose and motion-motion generation tasks. With our proposed training pipeline, TTRis capable to effectively build correlations between text, motion and space modalities.

In summary, our main contributions are as follows:

- We introduce a unified space-pose representation that effectively converts both absolute space and egocentric pose features into discrete tokens, facilitating LLM training.
- We propose a novel framework, Think-Then-React, that unifies two processes of inferring action intention and predicting reaction within one model, ensuring reaction generation quality in unconstrained scenario.
- Through extensive experiments, we demonstrate that our approach surpasses existing baselines by substantial margins, achieving an FID improvement from 3.988 to **1.942**, along with other ranking metrics.

2 RELATED WORK

2.1 HUMAN MOTION GENERATION

The field of Human Motion Generation focuses on creating realistic and diverse 3D human motion from various input modalities, including text (Zhang et al., 2023b; Guo et al., 2022b; Jiang et al., 2023; Guo et al., 2024; Liang et al., 2024), action labels (Guo et al., 2020; Xu et al., 2023; Petrovich et al., 2021), and human motion (Chopin et al., 2023; Liang et al., 2024; Xu et al., 2024). Most research has concentrated on text-conditioned single-person motion generation (text-to-motion) tasks. In this area, several works have utilized generative models commonly used in the vision domain, such as GANs, VAEs, and Diffusion Models, to generate human motion sequences. Another prominent approach (Zhang et al., 2023b; Guo et al., 2022b; Jiang et al., 2023; Guo et al., 2024) employs VQ-VAE to encode human motion sequences into one-hot tokens, which are then processed by auto-regressive models. This method converts the high-dimensional generation task into a next-token prediction task, effectively leveraging pre-trained large language models for more accurate text prompt understanding and diverse motion generation.

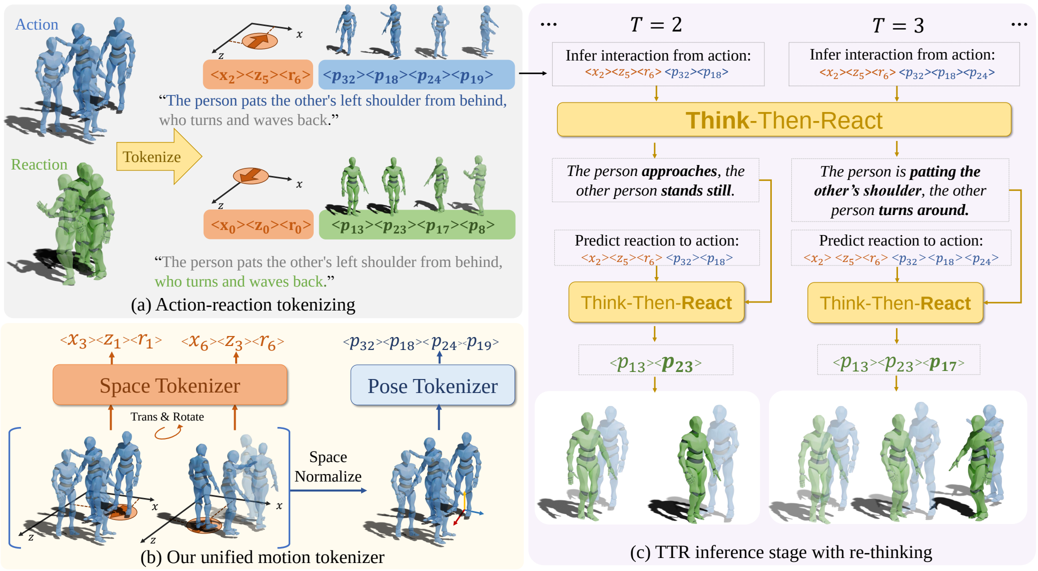
Recently, there has been growing interest in generating human motion in multi-person scenarios. InterGen (Liang et al., 2024) introduces a dual-person interaction dataset with detailed textual descriptions and a diffusion-based model for jointly generating multi-person interactions conditioned on text input. InterFormer (Chopin et al., 2023) utilizes temporal and spatial attention with human skeleton priors to generate human motion sequences reacting to input action sequences. The latest work, ReGenNet (Xu et al., 2024) employs a diffusion model to generate human reactions based on human actions in a unconstrained and online manner, and points out that given action's intention as a condition, the model can achieve superior performance compared to unconstrained settings. However, it directly predicts reaction motion without analyzing semantics of action motion. Our work unifies two processes: a thinking process that infers action semantics, and a reacting process that predicts reaction motion based on action motion and the thinking results, ensuring to generate reaction with appropriate semantics.

2.2 HUMAN MOTION CAPTIONING

Understanding human motion is a crucial challenge in the fields of computer vision, natural language processing, and multi-modality. Some works leverage large multi-modal models pre-trained on internet-scale text-video data to understand videos (Zhang et al., 2023a; Cheng et al., 2024). In the human motion domain, 3D human motion sequences contain more condensed human motion information, making the understanding of human motion more effective and efficient. TM2T (Guo et al., 2022b) utilizes transformers for reciprocal generation of 3D human motion tokens and texts. More recently, MotionGPT (Jiang et al., 2023) proposes leveraging a pre-trained language model and a fine-grained fine-tuning mechanism for both text and human motion token generation. Contrary to these works, our research focuses on understanding multi-person human motion to enhance reaction generation.

2.3 MULTI-MODAL LARGE LANGUAGE MODELS

Large Language Models (LLMs), especially multi-modal LLMs, have demonstrated impressive performance across various tasks (Zhang et al., 2023a; Zeng et al., 2024; Jin et al., 2023), benefiting from extensive training data and model size. A key component in integrating data from different modalities is the Vector Quantization (VQ) mechanism, which enables the conversion of representations from continuous space (e.g., image, audio, human motion) to one-hot tokens while preserving high fidelity, allowing for joint training with texts. Our work leverages the pre-trained language model Flan-T5 (Chung et al., 2024; Raffel et al., 2020), an encoder-decoder transformer model, for simultaneously understanding human action and generating reactions.


Action
Reaction
Tokenize
x
z
<x2><z5><r6>
<p32><p18><p24><p19>
"The person pats the other's left shoulder from behind, who turns and waves back."
<x0><z0><r0>
<p13><p23><p17><p8>
"The person pats the other's left shoulder from behind, who turns and waves back."
<x3><z1><r1>
<x6><z3><r6>
<p32><p18><p24><p19>
Space Tokenizer
Pose Tokenizer
Trans & Rotate
Space Normalize
T = 2
T = 3
Infer interaction from action: <x2><z5><r6> <p32><p18>
Infer interaction from action: <x2><z5><r6> <p32><p18><p24>
Think-Then-React
The person approaches, the other person stands still.
The person is patting the other's shoulder, the other person turns around.
Predict reaction to action: <x2><z5><r6> <p32><p18>
Predict reaction to action: <x2> <z5><r6> <p32><p18><p24>
Think-Then-React
Think-Then-React
<p13><p23>
<p13><p23><p17>

(a) Action-reaction tokenizing

(b) Our unified motion tokenizer

(c) TTR inference stage with re-thinking

Figure 1: (a) We propose a unified tokenizing process that encodes human action and reaction while maintaining absolute space feature and egocentric motion feature. (b) To obtain space tokens of a motion, we first extract its initial space state, i.e., 2D position and body orientation. Then we normalize the body center at the origin while facing positive z axis for effectively encoding the following pose sequences. (c) During inference, our method TTR first infers action's intent and semantics. Then TTR could predict corresponding reaction based on both the input action and inferred intent.

3 METHOD

3.1 OVERVIEW

For the task of action-to-reaction, given a human action $\mathrm{a} = \{a_i\}_{i=1}^{N_f}$ over N_f frames, our aim is to generate a corresponding reaction $\mathrm{b} = \{b_i\}_{i=1}^{N_f}$ without any input prompts. Most previous works leverage input prompts but they are often inaccessible in unconstrained interaction setting. As the example in Figure 1 shows, when a robot/avatar meets a human, it can only observe the human behaviors, try to understand her/his intents, and think what the robot/avatar is expected to react. There is no prompt available to tell what they are going to do. Furthermore, to better align with real-world scenarios, the generation process during inference is online manner with time interval l, i.e., when generating b_k once every l time steps with $\{a_i\}_{i=1}^{k-l}$ as input, and l is much smaller than N_f. Such a problem is challenging due to the complexity of representing two-person motion and their spatial relationship, the absence of prompts during inference.

To address the above problem, we propose a unified framework, Think-Then-React (TTR), for both action understanding and reaction generation. As illustrated in Figure 1, we first train a unified tokenizer to convert both egocentric poses and and absolute spatial location and direction information into tokens in Section 3.2. Then we propose a unified Large Language Model (LLM) based model that are pre-trained on three categories of motion and language related tasks, such as describing a motion, and then fine-tune the model with instructions of predicting a reaction from a given action in Section 3.3.

To avoid confusion to previous works, we denote **pose** by human body pose or movement in a small time fraction like "taking one step forward", and a **motion** consists of an initial space state with following poses, like "a person walks forward starting from the center point".

3.2 UNIFIED MOTION REPRESENTATION

To represent one or two persons (denoted by p1 and p2) in an absolute coordinate system, we normalize their centers at the origin while facing positive z axis. Then for each frame, we extract the 3D skeletons' joint position, velocity and rotation as normalized (or egocentric) pose feature. Before normalizing, we save the persons' absolute 2D positions and orientations, denoted as x, z and r, to maintain absolute space feature. Based on pose and space features of p1 and p2, we propose a unified tokenizing pipeline to convert them into LLM-readable tokens.

3.2.1 EGOCENTRIC POSE TOKENIZER

Our aim is to convert continuous pose features into discrete pose tokens like "$<p_{128}><p_{42}>$...". To achieve this, we adopt VQ-VAE (Van Den Oord et al., 2017), similar to Jiang et al. (2023), as the egocentric pose tokenizer. The pose tokenizer consists of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. $\mathcal{E}$ first encodes continuous pose features, i.e., the 22 joints' position, rotation and velocity vector $\mathrm{m} = \{p_i\}_{i=1}^{M}$ into N discrete pose tokens, where L denotes the number of frames or original motion. Specifically, $\mathcal{E}_p$ and $\mathcal{D}_p$ are 1D convolution networks with down-sample and up-sample blocks.

We first obtain the latent representation of a motion $\hat{\mathrm{c}} = \mathcal{E}(\mathrm{m})$. Then, we set up a learnable codebook $C = \{c_i\}_{i=1}^{K} \subset \mathbb{R}^d$ with K entries in size d. A quantization operation $Q(\cdot)$ is applied on the encoded motion latent features by replacing each row vector $\hat{c_i} \in \hat{\mathrm{c}}$ with its nearest codebook entry c_k. The process is formulated as:

$$\mathrm{c}_q = Q(\hat{\mathrm{c}}) := (\underset{c_k \in C}{\arg\min}\, ||\hat{\mathrm{c}_i} - \mathrm{c_k}||) \in \mathbb{R}^{N \times d} \tag{1}$$

Then, we obtain the reconstructed pose feature $\hat{\mathrm{m}}$ through the decoder $\hat{\mathrm{m}} = \mathcal{D}(\mathrm{c}_q)$. The overall process of the VQ-VAE can be formulated as:

$$\hat{\mathrm{m}} = \mathcal{D}(Q(\mathcal{E}(\mathrm{m}))). \tag{2}$$

This is trained via a reconstruction loss with codebook commitment loss. Noting that the $argmin$ operation is non-differentiable, we simply copy the gradients from $\mathcal{D}_m$ to $\mathcal{E}_m$ as the estimated gradient. Furthermore, for smoother reconstructed motion and a stable training process, we add an extra velocity regularization in the reconstruction loss and employ exponential moving average (EMA) with codebook reset techniques, following Zhang et al. (2023b). More details about this section are provided in the appendix.

3.2.2 ABSOLUTE SPACE TOKENIZER

For better generalization capability, all motions, including actions and reactions, are normalized to the original point and same direction before being tokenized. Therefore, absolute space information, i.e., the human body 2D position and orientation of each person, is omitted. To extend egocentric pose tokens with absolute space information, we propose converting the position and rotation of a person's center point into LLM-readable tokens.

As shown in Figure 1, before normalizing the human motion, we first extract the center point's features, i.e., the position x and z in the absolute coordinate system where the up-down direction is represented by the y-axis by default, and orientation r. We then compute the range of x, z, and r across the dataset to get the maximum and minimum values. These ranges are uniformly divided into N_b bins, converting each continuous value to discrete tokens. For example, $x = 0.55$ will be represented as token "$<x_{15}>$" when all the x positions are in $[-1, 1]$ and divided into $N_b = 20$ bins.

Finally, we use a unified coding system to represent action, reaction, and their relative information. Specifically, at each time step t, we apply absolute space tokenizer to encode x, z, and r of the center point at the beginning into egocentric pose tokens, and apply pose tokenizer to encode a series of normalized motions before next time step $t+1$ into pose tokens. Such tokens enable training a model that can understand and generate motion and language simultaneously effectively and efficiently in the subsequent phase.

3.3 UNIFIED LLM BASED MOTION UNDERSTANDING AND GENERATION

3.3.1 PRE-TRAINING

To adapt a large language model into a motion-language model (we combine "space" and "pose" to call it "motion"), we first pre-train the model with multiple tasks in diverse formats. The pre-training tasks can be categorized into three main types:

(1) Motion - Text. To enable the model to understand and generate human motion tokens, we combine the action and reaction token sequences to construct prompts, which are then fed into the model to generate corresponding textual descriptions, and vice versa. For example, the input sequence could be "Describe the interaction. Action: $<x_0><z_1><r_2><p_2><p_7>$..., Reaction: $<x_7><z_7><r_8><p_1><p_9>$...", and the target response is: "*One waves the right hand, and the other one waves back*". However, reaction motions are not given during the inference phase. Therefore, the reaction motion is randomly dropped during the training phase to enable the model to infer the interaction from the action motion solely. We also pre-train our model by the instructions on predicting action and reaction token sequences from an interaction description prompt in text.

(2) Pose - Space. Spatial information is represented by orthogonal one-hot tokens, but it may be helpful to infuse auxiliary spatial information into the model. Specifically, we design two tasks: i) Egocentric pose to absolute space: Given space token and subsequent pose tokens of t time step, we train the model to predict the space tokens of $t+1$ time step. For example, given input space token $<z_{12}>$ and a pose token $<p_{56}>$, which represents "stepping foward", the target output should be $<z_{13}>$, denoting spatial transition. ii) Absolute space to egocentric pose: Similarly, given space tokens of t and $t+1$ time step, the model is trained to predict pose tokens between them.

(3) Motion - Motion. To capture more fine-grained action-reaction relationship, , we take the first half of action and the second half of reaction sequence of tokens as input, with their corresponding initial space tokens, and pre-train the model to complete the remaining motion clips, i.e., the second half of action and the first half of reaction, and vice versa. For example, given a sequence consisting ten time steps $t_{1:10}$, we fed space tokens at t_1 and pose tokens in $t_{1:5}$ of action, along with space tokens at t_6 and pose tokens in $t_{6:10}$ of reaction, then supervise the model to predict action poses in $t_{6:10}$ and reaction poses in $t_{1:5}$.

During pre-training, we jointly train all the tasks in a non-causal manner for better efficiency. Owning to our unified motion and language architecture and space-pose token representation, single person motion and text data can be seamlessly integrated into the training process. We adopt HumanML3D (Guo et al., 2022a), a large scale single person motion-text dataset to facilitate pre-training. To avoid overfitting, we prepare 20 prompt templates for each task and randomly mask out 15% of motion tokens during training. In addition, we adopt random clipping of motions as augmentation. We also find that text generation tasks converge much faster than motion generation tasks. Therefore, we adopt a weighted sampling strategy for pre-training tasks, using the validation losses for each task as the weights to sample training tasks for the next training epoch.

3.3.2 FINE-TUNING

After pre-training, the motion-language model is well-structured with knowledge of motion, space, and text. To make the model applicable to online action-to-reaction generation, we fine-tune it in a causal manner, focusing on two tasks: thinking and reacting.

The **thinking** task involves understanding action motion, e.g., "the person is waving hand", and inferring its possible interaction, e.g., " two persons wave goodbye to each other", or reaction, "the other person waves back". At each training iteration, we randomly choose the first quarter, half, or the entire action sequence as input to predict the entire interaction caption. However, the entire action motion is not given in the early stage of inference, thus the inferred description based on action motion clips may not be accurate, thus we adopt periodical **re-thinking** in the inference phase for each N_r action tokens given, to dynamically adjust the prompt for reaction generation. We define N_r as re-thinking interval.

For the **reacting** task, we aim to supervise the model to generate reaction motions conditioned on the generated descriptions during the thinking process. However, in the early stages of fine-tuning, the inferred interaction descriptions are not accurate enough to guide the reaction generation process.

Thus, we adopt a teacher forcing approach. In the early stages, the model takes the ground-truth text prompt t as a condition to generate the entire reaction sequence. Meanwhile, we monitor the validation loss and text generation metrics. When the metrics tend to converge, we begin to sample predicted interaction captions $\hat{\mathrm{t}}$ by the model and use them as reaction generation conditions. This process ensures alignment between training and inference, as ground-truth prompts are inaccessible during inference.

4 EXPERIMENT

We evaluate our proposed method with strong baselines and further analyze contributions of different components, and the impact of key parameters.

4.1 EXPERIMENT SETUP

Dataset. We evaluate all the methods on Inter-X dataset, which consists about 9K training samples and 1,708 test samples. Each sample is a action-reaction sequence and 3 corresponding textual description. As supplementation, we mix our pre-training data with single person motion-text dataset HumanML3D (Guo et al., 2022a), that consists more than 23K annotated motion sequences, and we uniformly sample frames for both datasets to 30 FPS.

Evaluation Metrics. Following single-person motion generation (Zhang et al., 2023b), we adopt the following metrics to quantitatively evaluate the generated motion: R-Precision measures the ranking of Euclidean distances between motion and text features. Accuracy (Acc.) assesses how likely a generated motion could be successfully recognized as its interaction label, like “high-five”. Frechet Inception Distance (Heusel et al., 2017) (FID) evaluates the similarity in feature space between predicted and ground-truth motion. Multimodal Distance (MMDist.) calculates the average Euclidean distance between generated motion and the corresponding text description. Diversity (Div.) measures the feature diversity within generated motions. All the metrics reported are calculated with batch size set to 32, and accumulated across the test dataset, and we test each method for 20 times with different seeds to calculate the final results at 95% confidence interval.

Evaluation Model. All the metrics mentioned above requires an encoder to extract motion features. For single person text-to-motion generation tasks, a motion-text matching model are commonly trained as human motion feature extractor. A simple way to transfer this method to interaction domain is to directly train an interaction-to-text matching model $\mathcal{M}(\mathrm{a}, \hat{\mathrm{b}}, \mathrm{t})$, where action sequence a and predicted reaction sequence $\hat{\mathrm{b}}$ together is regarded as a generated interaction sequence, or a reaction-to-text match model $\mathcal{M}(\hat{\mathrm{b}}, \mathrm{t})$. However, the former one may focus too much on the ground-truth action input, leading insufficient discriminative power of $\hat{\mathrm{b}}$’s quality.

To address the issue, we simply uniformly mask off a large portion of a, obtaining down-sampled action motion sequence a', that serves as a semantic hint for the matching process while not introducing too much emphasis on input action sequence. The final evaluation model consists of an masked interaction encoder $\mathcal{E}_i(\mathrm{a}', \hat{\mathrm{r}})$ and an interaction text encoder $\mathcal{E}_l(\mathrm{t})$. We use contrastive loss following CLIP (Radford et al., 2021), that encourages paired motion and text features to be close geometrically. In addition, we add a classification head after the predicted motion features, to simultaneously predict interaction labels, such as “high-five”.

Baselines. To evaluate the performance of our method TTRon online and unconstrained setting, we compare TTRwith the following baselines: 1) **InterFormer** (Chopin et al., 2023) is a transformer based action-to-reaction generation model that leverages human skeleton as prior knowledge for efficient attention process. 2) **MotionGPT** (Jiang et al., 2023) is a motion-language model that leverages an LLM for motion and text generation, we change its backbone LLM to Flan-T5-base(Chung et al., 2024; Raffel et al., 2020) and extend MotionGPT to handle multi-person scenarios. 3) **InterGen** (Liang et al., 2024) proposes a mutual attention mechanism within diffusion process for human interaction generation, we reproduce and adapt IngerGen to action-to-reaction generation. 4) **ReGenNet** (Xu et al., 2024) is latest state-of-the-art model on action-to-reaction generation. It adopts a transformer decoder based diffusion model, that directly predicts human reaction given action input in unconstrained and online manner as ours.

Table 1: Comparison to state-of-the-art baselines and ablation studies of our method on Inter-X dataset. ↑ or ↓ denotes a higher or lower value is better, and → means that the value closer to real is better. We use ± to represent 95% confidence interval and highlight the best results in **bold**. For ablation methods (in grey), PT, M, P, S, and SP are abbreviations for pre-training, motion, pose, space, and single-person data, respectively.

Methods	R-Precision↑			Acc.↑	FID↓	MMDist↓	Div.→
	Top-1	Top-2	Top-3				
Real	$0.511^{\pm.003}$	$0.682^{\pm.002}$	$0.776^{\pm.002}$	$0.463^{\pm.000}$	$0.000^{\pm.000}$	$5.348^{\pm.002}$	$2.498^{\pm.005}$
InterFormer	$0.172^{\pm.012}$	$0.292^{\pm.013}$	$0.343^{\pm.012}$	$0.171^{\pm.009}$	$10.468^{\pm.021}$	$7.831^{\pm.018}$	$3.505^{\pm.023}$
MotionGPT	$0.238^{\pm.003}$	$0.354^{\pm.004}$	$0.441^{\pm.003}$	$0.186^{\pm.002}$	$5.823^{\pm.048}$	$6.211^{\pm.005}$	$2.615^{\pm.007}$
InterGen	$0.326^{\pm.036}$	$0.423^{\pm.063}$	$0.525^{\pm.053}$	$0.254^{\pm.019}$	$5.506^{\pm.257}$	$6.182^{\pm.038}$	$2.284^{\pm.009}$
ReGenNet	$0.384^{\pm.005}$	$0.483^{\pm.002}$	$0.572^{\pm.003}$	$0.297^{\pm.004}$	$3.988^{\pm.048}$	$5.867^{\pm.009}$	$\mathbf{2.502^{\pm.001}}$
TTR(Ours)	$\mathbf{0.423^{\pm.005}}$	$\mathbf{0.599^{\pm.003}}$	$\mathbf{0.693^{\pm.003}}$	$\mathbf{0.318^{\pm.003}}$	$\mathbf{1.942^{\pm.017}}$	$\mathbf{5.643^{\pm.003}}$	$2.629^{\pm.006}$
w/o Think	$0.367^{\pm.003}$	$0.491^{\pm.027}$	$0.584^{\pm.008}$	$0.230^{\pm.036}$	$3.828^{\pm.016}$	$6.186^{\pm.055}$	$2.609^{\pm.006}$
w/o All PT.	$0.398^{\pm.007}$	$0.531^{\pm.002}$	$0.628^{\pm.003}$	$0.288^{\pm.002}$	$3.467^{\pm.113}$	$5.822^{\pm.003}$	$2.909^{\pm.053}$
w/o M-M PT.	$0.408^{\pm.005}$	$0.563^{\pm.004}$	$0.646^{\pm.005}$	$0.293^{\pm.002}$	$2.874^{\pm.020}$	$5.736^{\pm.003}$	$2.553^{\pm.006}$
w/o P-S PT.	$0.417^{\pm.004}$	$0.582^{\pm.004}$	$0.664^{\pm.004}$	$0.308^{\pm.003}$	$2.685^{\pm.024}$	$5.699^{\pm.004}$	$2.859^{\pm.007}$
w/o M-T PT.	$0.406^{\pm.003}$	$0.557^{\pm.004}$	$0.637^{\pm.004}$	$0.304^{\pm.003}$	$2.580^{\pm.021}$	$5.822^{\pm.003}$	$2.889^{\pm.005}$
w/o SP Data	$0.414^{\pm.004}$	$0.592^{\pm.005}$	$0.685^{\pm.003}$	$0.315^{\pm.004}$	$2.007^{\pm.015}$	$5.667^{\pm.003}$	$2.611^{\pm.005}$

Implementation Details. For motion representation, we use the same strategy as Liang et al. (2024), which combines local joint positions, rotations, velocities, and foot-ground contact as the feature of human motion. Regarding the tokenizers, we adopt a temporal down-sample rate $d = 4$ and $N_t = 256$ for motion tokens, each motion token are in $d_c = 512$ in the codebook. We divide all space tokens into $N_b = 100$ bins. The motion VQ-VAE is trained for 150K steps with batch size set to 256 and learning rate fixed at 1e-4 on a single Tesla V100 GPU. For the LLM, we adopt Flan-T5-base (Chung et al., 2024; Raffel et al., 2020) as our base model, with extended vocabulary. We warm up the learning rate for 1,000 steps, peaking at 1e-4 for the pre-training phase, and use the same learning rate for fine-tuning. Both the pre-training and fine-tuning phases are trained on a single machine with 8 Tesla V100 GPUs. The training batch size is set to 32 for the LLM and we monitor the validation loss and reaction generation metrics for early-stopping, resulting about 200K pre-training steps and 40K fine-tuning steps. During inference, we set the reaction time interval l as 4 tokens, i.e., about 0.5 second with the fps of motion being 30. The re-thinking interval is set to N_r as 8 tokens to balance efficiency and generation quality. For the evaluation model, we down-sample the action motion by 30 times, resulting 1 fps action motion as input.

4.2 COMPARISON TO BASELINES

As shown in the upper side of Table 1, our proposed method TTR significantly outperforms baseline methods in terms of ranking, accuracy, FID and multimodal distance, showing superior human reaction generation quality. Compared to MotionGPT, which adopts a similar motion-language architecture, TTR expresses stronger performance, which we attribute to our unified representation of motion via space and pose tokenizers, enabling effective individual pose and inter-person spatial relationship representation. TTR also surpasses the diffusion-based methods, InterGen and ReGenNet, with our think-then-react architecture, improving generated motions by describing the observed action and reasoning what reaction is expected on semantic level. In addition, ReGenNet and MotionGPT get closer diversity to the real than our model. We mainly attribute to that, TTR may conduct multiple re-thinking processes during inference, and the inferred semantics may bring a higher diversity.

4.3 ABLATION STUDY OF KEY COMPONENTS

To evaluate the effectiveness of our proposed key designs, we conduct detailed ablation studies by removing each of them to observe how much drop compared to the full version of our TTR method. The larger drop indicates more contribution. The results are shown in grey lines of Table 1. According to the drops in FID, all designs, including thinking, pre-training tasks and using single person data in pre-training, have positive contributions to the final performance, and thinking contributes the most. Some detailed findings and analyses are as follows.

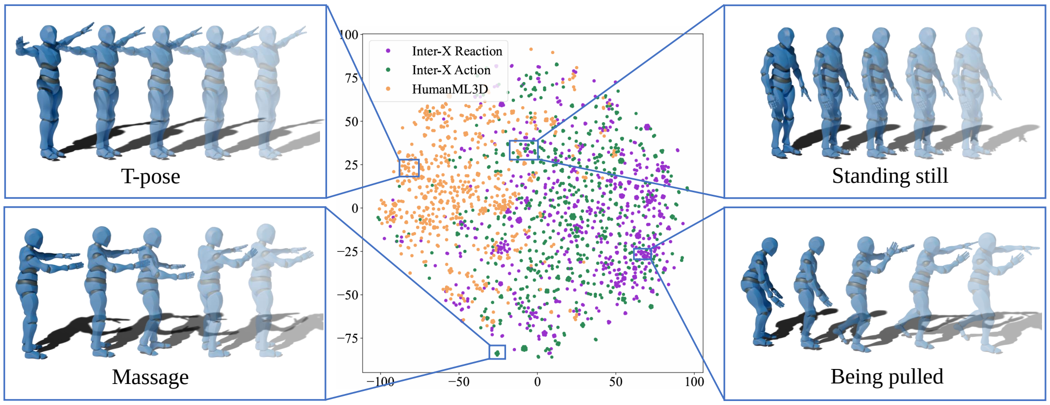


Figure 2: Visualization of a person's motion sequences in Inter-X dataset and HumanML3D dataset.

First, we skip **thinking** stage during inference, and find that the performance drops significantly in FID from 1.9 to 3.8. This supports the necessity of our propose thinking process before reacting. We also notice decreasing diversity of generated samples, as the model relies solely on input action, and cannot explicitly capture and infer action's intent, thus leading to more rigid motion in some cases.

Second, to evaluate the effectiveness of **pre-training**, we omit the pre-training stage, and directly train our model TTRfor thinking and reacting tasks. As shown in Table 1, our model's performance deteriorates without a fine-grained pre-training phase from 1.9 to 3.4 in FID. This indicates that pre-training can effectively adapt a language model (Flan-T5-base) into a motion and language model. We further removing three kinds of pre-training tasks: motion-motion (M-M PT.), pose-space (P-S PT.), and motion-text (M-T PT.). The results show that the without any task, the performance obviously gets worse, from 1.9 to 2.5 - 2.8 in FID, indicating their positive contribution to the final performance and complementary values to each other.

Third, to see how much **single-person data** helps reaction generation, we remove single person motion-text data, i.e., the data from HumanML3D dataset, from our training set. The result (w/o SP Data) shows that the model performs worse without training on HumanML3D, which proves that our unified motion encoder and motion-language architecture can leverage both single- and multi-person data, alleviating the insufficiency of training data. However, the benefit from single-person data is not as large as we expect.

4.4 ANALYSIS ON OVERLAPPING BETWEEN SINGLE- AND MULTI-PERSON MOTIONS

To investigate the reason of small contribution from single-person data, we further visualize motion sequences of single-person motion (HumanML3D), two-person action (Inter-X Action) and reaction (Inter-X Reaction) in the same space, as presented in Figure 2. Specifically, each motion sequence is represented by a vector, whose dimensions correspond to motion tokens in our codebook, and then all motion sequence vectors are projected to a two-dimensional plane by t-SNE tool. As shown in Figure 2, the single- and two-person motion sequences have little overlap. When doing case studies, we find that most two-person motion are unique, e.g., massage and being pulled, and will never be used in single-person motion. Similarly, most single-person motions are unique too, e.g., T-pose, and seldomly appear in multi-person interaction. There are only a few overlapped motions, e.g., standing still. In addition, when comparing action and reaction sequences in multi-person interaction, we have some interesting findings. When reactions are close to actions, the motion usually belongs to symmetrical interactions, e.g., pulling or being pulled; whereas, when actions are far from reactions, the motion usually belongs to asymmetrical interaction, e.g., massage.

4.5 IMPACT OF DOWN-SAMPLING PARAMETER IN MATCHING MODEL FOR EVALUATION

As described in Section 4.1, we propose down-sampling action motion sequence to avoid matching models for evaluation pay too much attention to input action rather than output reaction. We conduct

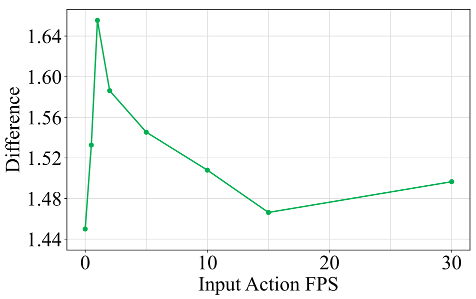


Figure 3: Impact of input action FPS (#Frame per Second) to the summed ranking scores between results of two methods.

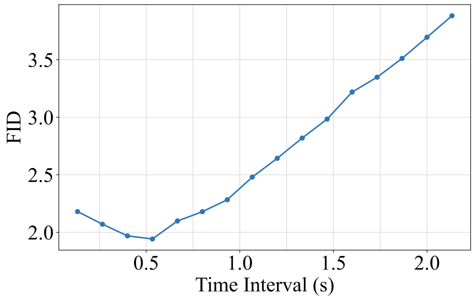


Figure 4: Impact of time interval l that our TTR model thinks and reacts to the generative quality measurement FID.

an experiment to change the down-sampling parameter FPS (#Frames per Seconds) and calculate the difference between the baseline RenGenNet model and a random generated result in terms of FID. As presented in Figure 4, the difference keeps positive as the baseline results are definitely better than the random ones and we prefer larger difference to better distinguish the two methods. Difference is lowest when FPS equals to 0, which meaning we do not use input action motion while only matching generated reaction motion with ground-truth. It goes up to the peak when FPS equals 1 and quickly goes down to low values, even close to the lowest when FPS is about 15. This indicates that it is necessary to concatenate input action with generated reaction to compose a meaningful interaction in evaluation, otherwise the motion-text matching model cannot well recognize the interaction. However, only 1 FPS is enough. With larger FPS, the matching models will be disturbed by input action rather than the generated reaction. Thus, we choose 1 FPS, corresponding to the largest difference, as our final setting.

4.6 IMPACT OF TIME INTERVAL

Our aim is to generate real-time reaction online, and thus time interval is an important parameter to both quality and latency. We change the time interval l from about 0.1 to 2.0 seconds (1 token to 16 tokens) and observe how it impacts generative quality measure FID. As shown in Figure 4, FID falls down first until $l = 0.53$ (4 tokens) and then continues rising up. This indicate that the best time interval is about 0.5 second. When the time interval is too short, our TTR model cannot get enough information to re-think what the input action means and will bring some randomness into predicting appropriate reaction. When the time interval gets too long, our TTR model give slow responses to the input action sequences and generates coarse-grained reaction.

5 CONCLUSION

In this paper, we propose a novel framework Think-Then-React to address the action-to-reaction motion generation problem. First, we propose a unified motion encoder that tokenize a person's starting location at time t and poses between t and $t + 1$ separately. Then we design motion and text related tasks to pre-train a large language model backbone to understand and generate both language and motion. We also fine-tune the model to think what the input action means and what an appropriate reaction is, and then generate reaction motions. Experimental results show that our proposed TTR method outperforms all baselines in all metrics except for diversity. Our proposed thinking phase and all pre-training tasks contribute to the best performance. We find that although our proposed unified motion encoder enable leveraging single-person data in pre-training, it brings limited benefit due to the little overlapped poses between single-person motion and multi-person interaction. In the future, we plan to explore more effective method for single-person and multi-person dataset.

REFERENCES

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

Baptiste Chopin, Hao Tang, Naima Otberdout, Mohamed Daoudi, and Nicu Sebe. Interaction transformer for human reaction generation. *IEEE Transactions on Multimedia*, 25:8842–8854, 2023.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2021–2029, 2020.

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, 2022a.

Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pp. 580–597. Springer, 2022b.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

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.

Chuhao Jin, Wenhui Tan, Jiange Yang, Bei Liu, Ruihua Song, Limin Wang, and Jianlong Fu. Alphablock: Embodied finetuning for vision-language reasoning in robot manipulation. *arXiv preprint arXiv:2305.18898*, 2023.

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

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*, pp. 1–21, 2024.

Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10985–10995, 2021.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Liang Xu, Ziyang Song, Dongliang Wang, Jing Su, Zhicheng Fang, Chenjing Ding, Weihao Gan, Yichao Yan, Xin Jin, Xiaokang Yang, et al. Actformer: A gan-based transformer towards general action-conditioned 3d human motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2228–2238, 2023.

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*, pp. 1759–1769, 2024.

Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, Tao Kong, and Ruihua Song. What matters in training a gpt4-style language model with multimodal inputs? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7930–7957, 2024.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023a.

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, 2023b.

Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7368–7376, 2024.

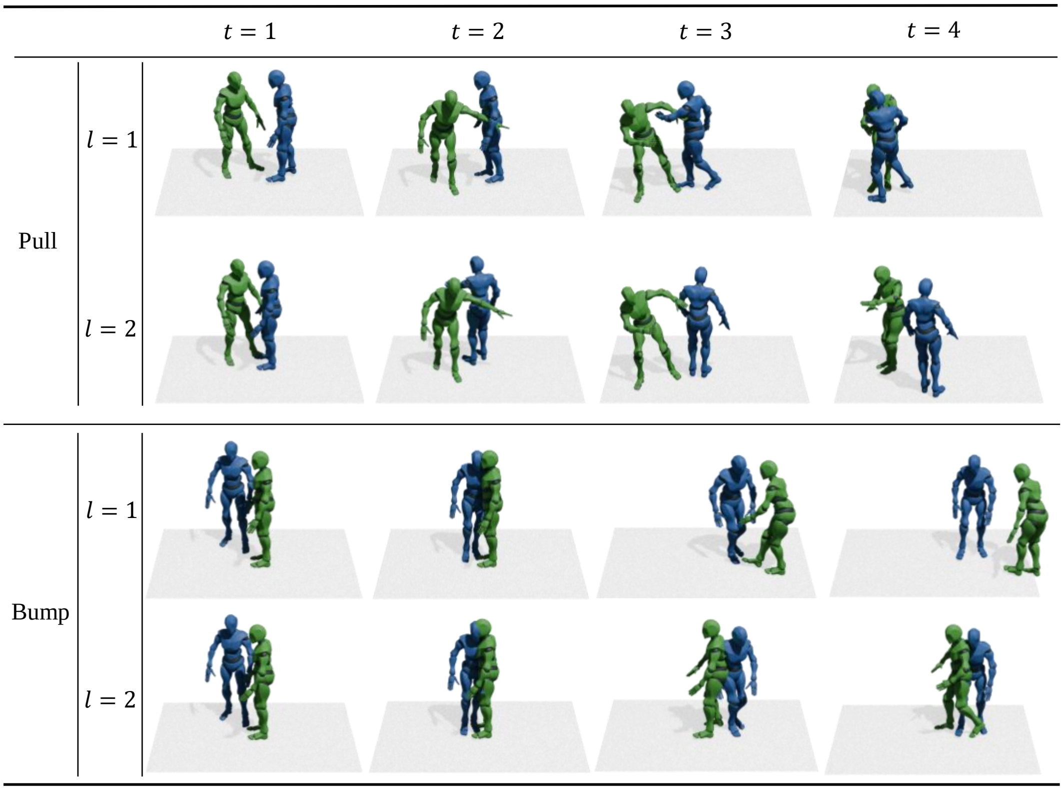


Figure 5: Too large time interval l leads to motion latency and penetration.

A APPENDIX

A.1 POSE TOKENIZER

We adopt a similar architecture to Guo et al. (2022b) as our pose tokenizer. The encoder/decoder consists of two down-sample/up-sample 1D convolution layers and three 1D ResNet blocks He et al. (2016). We set the width of the auto-encoder to 512. To train the model on both Inter-X and HumanML3D datasets for 200,000 steps, with batch size set to 256, and learning rate set to 1e-4. To train the model, we apply L1-loss on both pose feature and velocity reconstruction, and a commitment loss for the embedding process. The weight set to velocity loss is 0.5 and commitment loss is 0.02.

A.2 MATCHING MODEL

For the motion-text matching model, we adopt a similar architecture to InterCLIP (Liang et al., 2024), which consists of an eight-layer motion transformer encoder and an eight-layer text transformer encoder. The hidden size is set to 768 and attention heads is set to 8. We add an learnable token to the motion encoder and extract its feature in the last layer of motion encoder as the pooled motion feature. To conduct motion classification, we add a classification head (an MLP) after the pooled motion feature. We use the text embedding layer from clip-vit-large-patch14 (Radford et al., 2021), which is frozen during training. We train the model for 40 epochs with batch size set to 128. The learning rate is warmed-up to 0.001 in the first 1,000 steps.

A.3 CASE STUDY

We provide case studies on time interval in Figure 5 and re-thinking process in Figure 6.

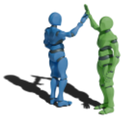
Time	Motion	Thinking results
$t = 1$		*“The first person stands still, the second person stands in front of her/him.”*
$t = 3$		*“Two people standing facing each other. The first person reaches her/his right hand, she/he grasps the other person's hands and they shake twice.”*
$t = 5$		*“The first person raises her/his right hand. The second person extends her/his hands and give she/him a high-five.”*
$t = 7$		*“Two people stand facing each other and raise their right hands. They gave each other a hive-five above their heads.”*

Figure 6: A case of re-thinking process.