# **InterMask**: 3D Human Interaction Generation via Collaborative Masked Modeling

**ICLR 2025**

Muhammad Gohar Javed[1], Chuan Guo[2], Li Cheng[1], Xingyu Li[1]

[1]University of Alberta, [2]Snap Inc.

UNIVERSITY OF ALBERTA

ICLR

# Motivation

## Importance of Human Interactions in AI

**What are human interactions?**

Coordinated **exchanges of movements** between two individuals, captured as 3D skeletal motions.

**What do they represent in human intelligence and AI?**

A reflection of **cognition** and **behavior**, essential for AI to understand and replicate.

**Why model them using Generative AI?**

To enable **realistic**, **contextually appropriate**, and **diverse** behavior generation.

# Motivation

Applications in animation, games, VR

**Animation**

**Games**

**Virtual Reality**


## Challenge

Creating human interactions manually is **time-consuming**, **labor-intensive**, and often **lacks diversity** or true **realism**.

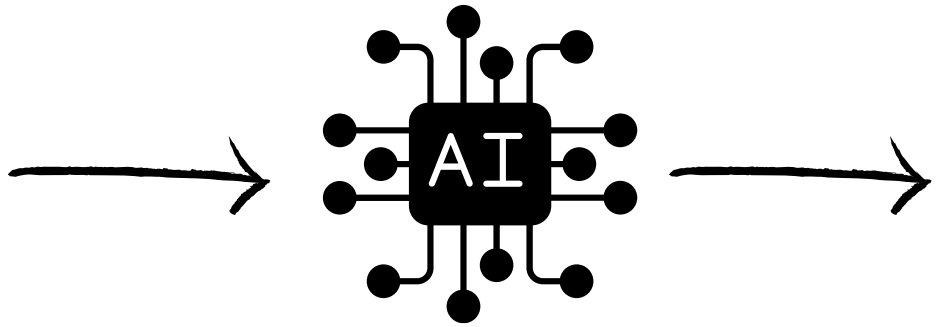
## Generative AI as a solution

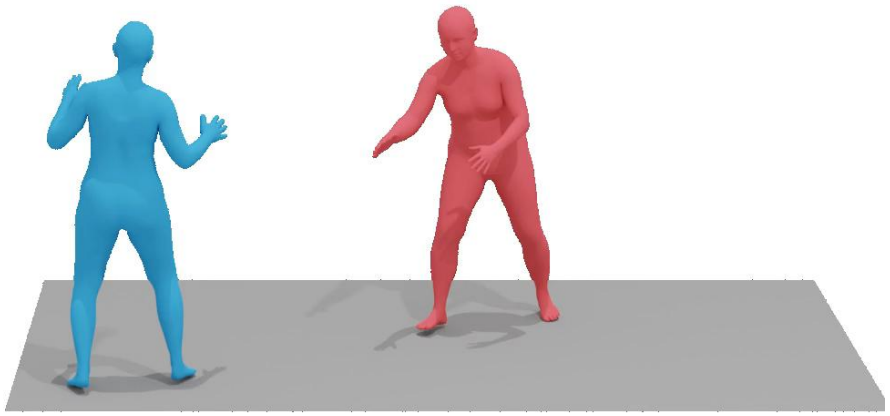
A Generative AI model can produce **realistic**, **human-like**, **diverse**, and **controllable** interactions, with minimal human effort.

# Goal: Human Interaction Generation

Enable interaction generation for smart digital character control in virtual environments

One person **strikes the other** with a sword and the other **dodges**



# Challenges

## Individual Motion

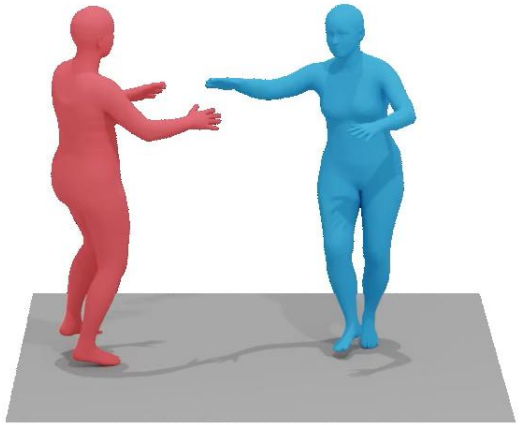
- Pose and Motion Quality
- Temporal Consistency

## Two-person Interaction

- Pose and Motion Quality
- Temporal Consistency
- Interaction Quality
- Spatial Consistency

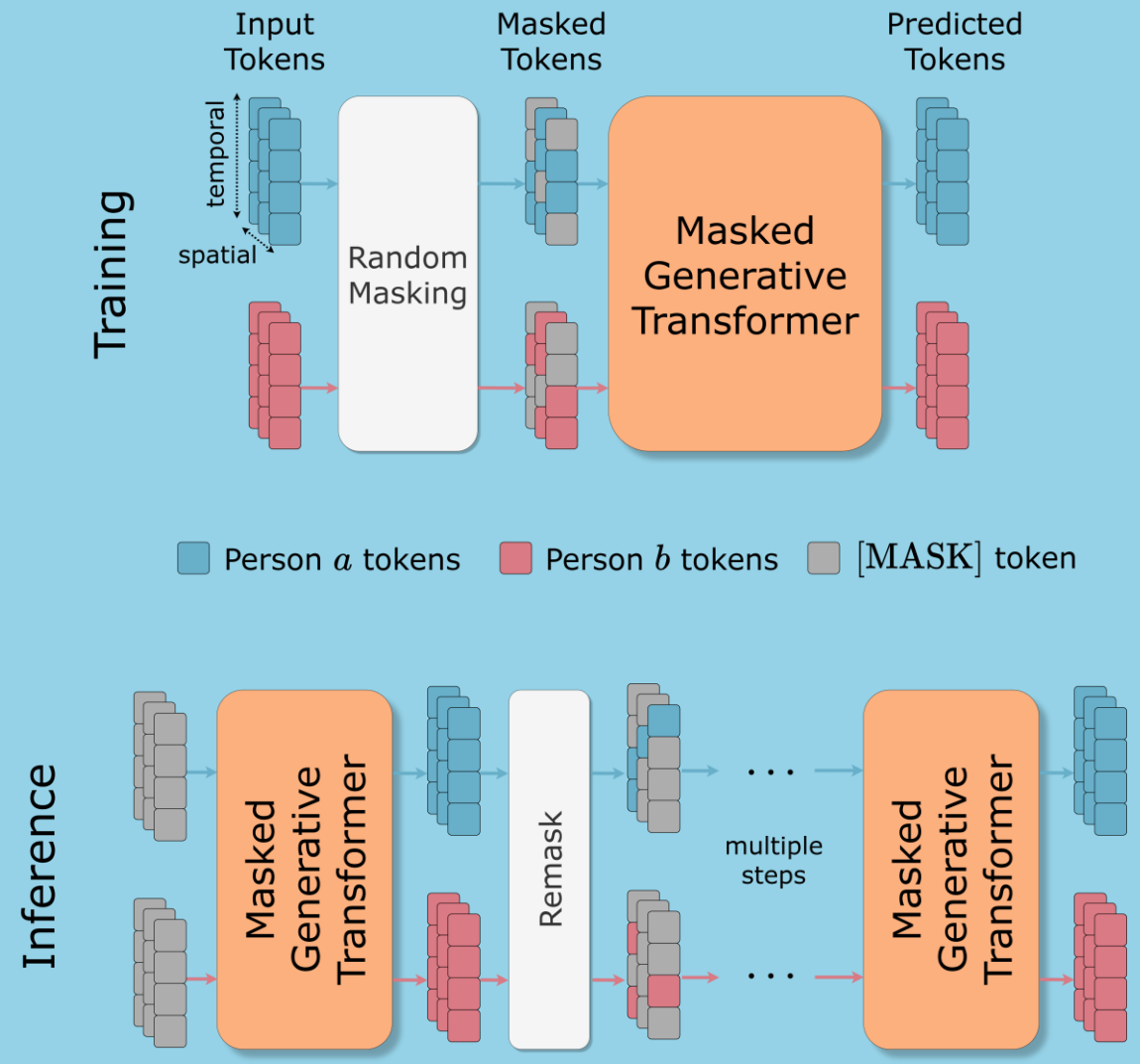
A person **walks in a clockwise circle**

Two people are **spinning around in clockwise direction**



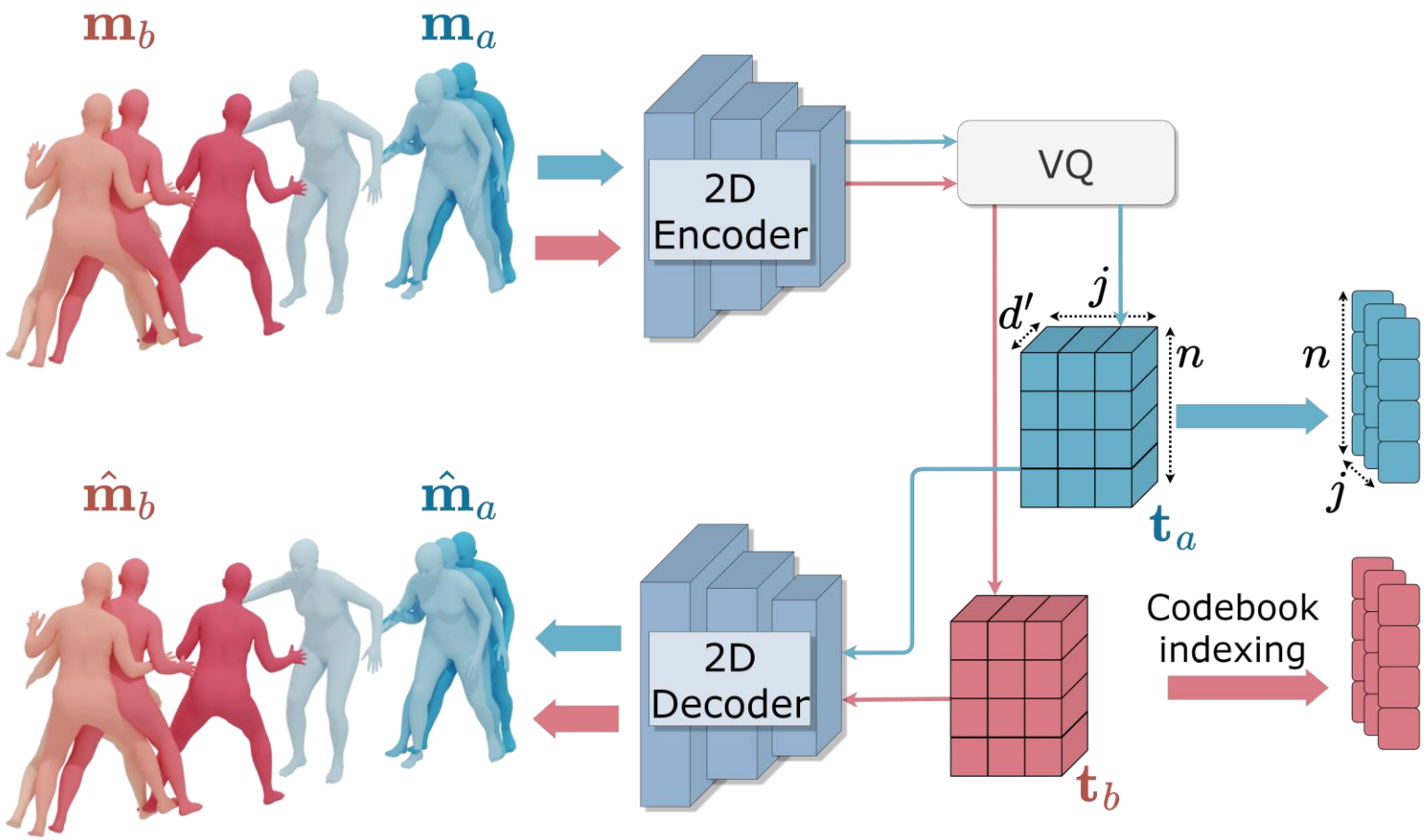
# Overview

- **A Generative Masked Modelling Objective**
  - Input tokens are randomly masked and predicted during training.
  - Fully masked tokens are progressively generated during inference.
- **2D Discrete Motion Token Map**
  - A 2d token map which preserves the spatial and temporal dimension.
- **Collaborative Transformer with Spatio-Temporal Attention**
  - Every token in both motions attends to all the tokens in the sequence. Spatial and temporal dimensions in the interaction data are explicitly modeled.

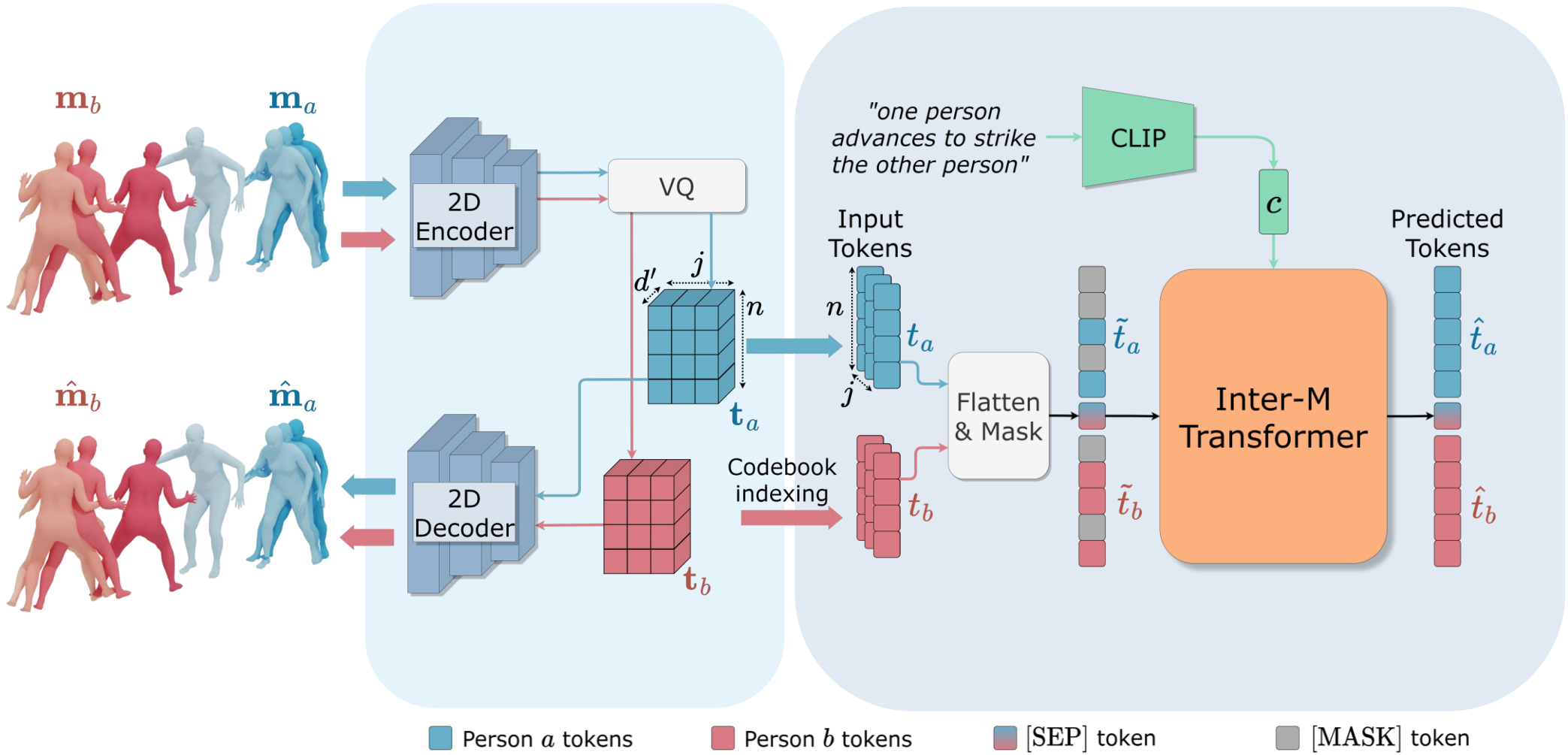


# 2D Motion Token Map

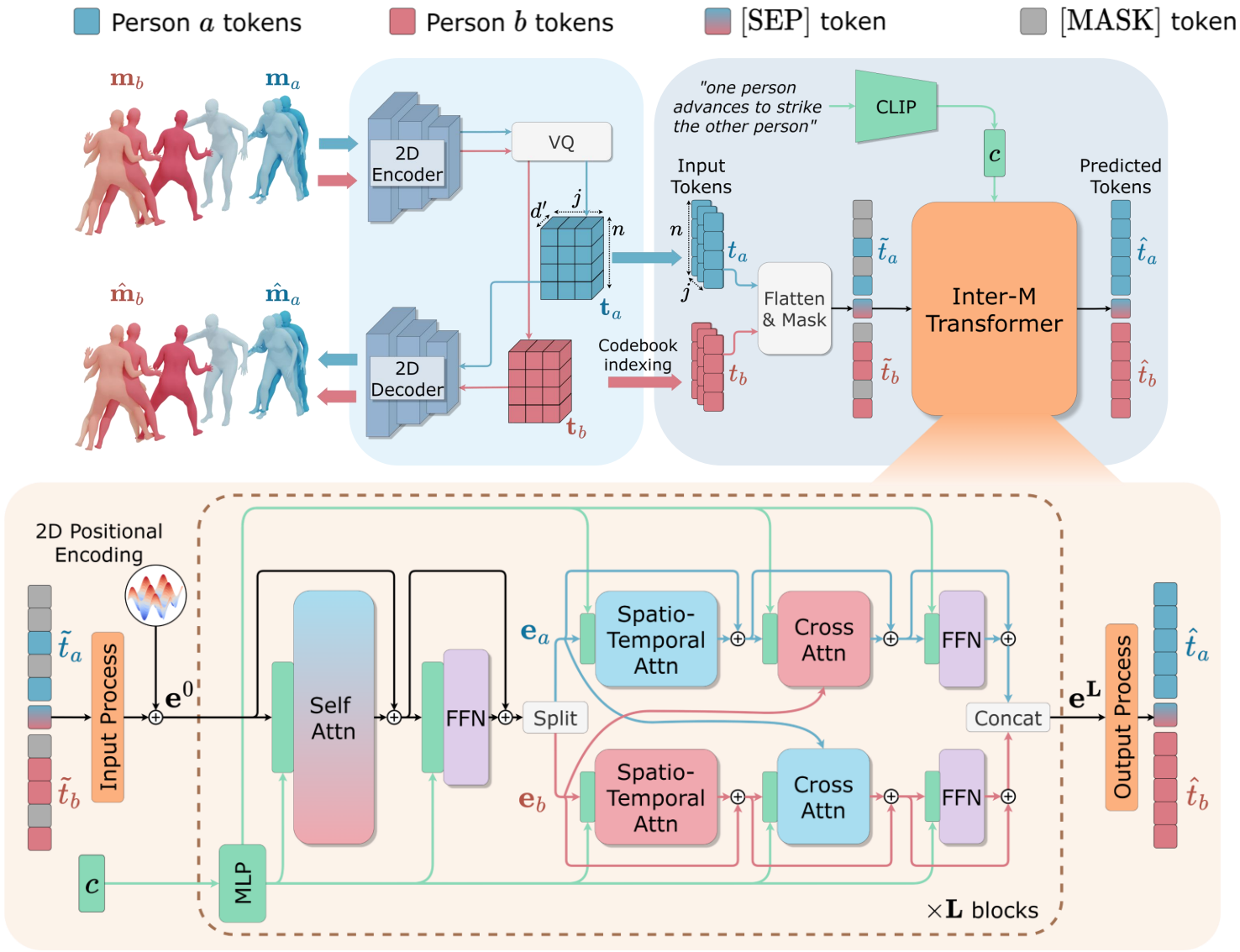
2D motion Vector Quantized Variational AutoEncoder (VQ-VAE)



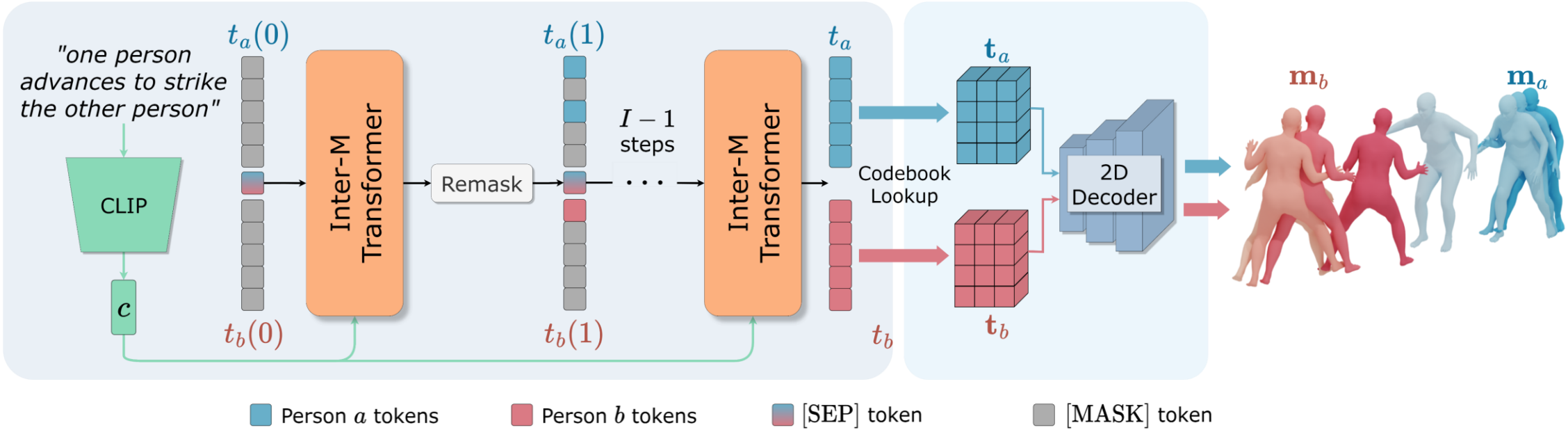
# Interaction Masked Generative Transformer



# Interaction Masked Generative Transformer



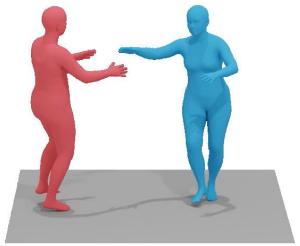
# Interaction Generation Inference



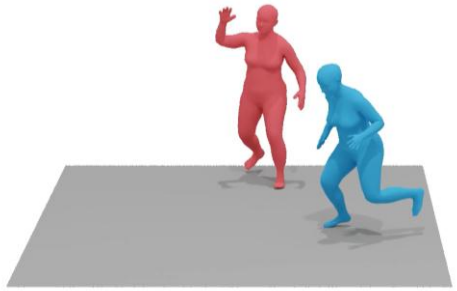
# Interaction Generation Results

## Qualitative

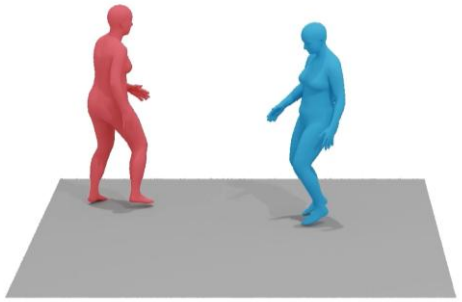
Two people are **spinning around in clockwise direction**



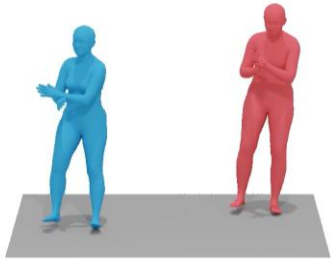
The first **runs to their right** and the other begins to **chase them**



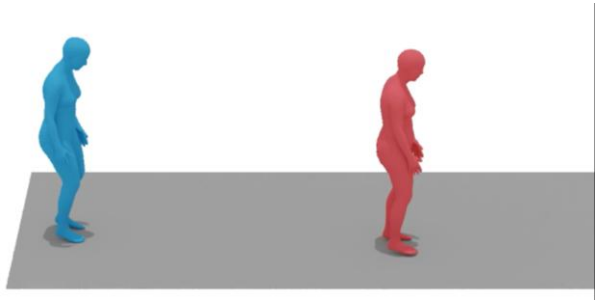
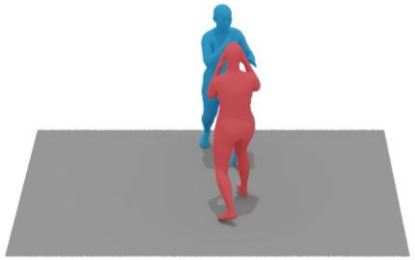
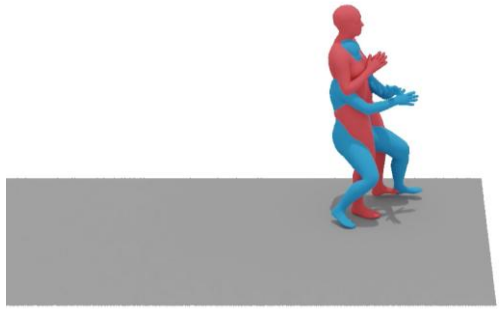
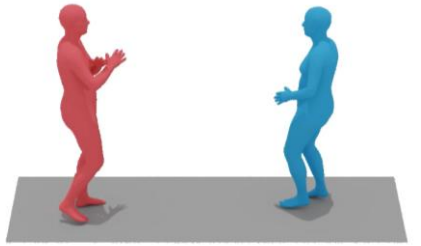
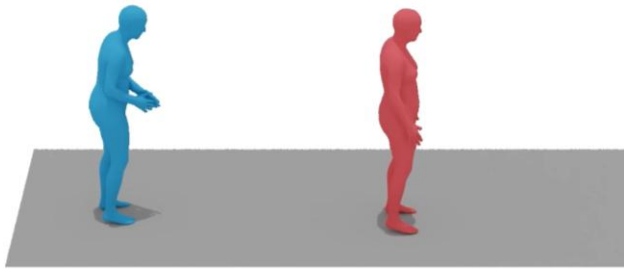
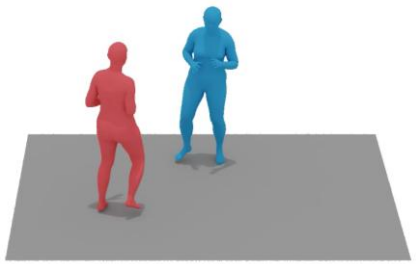
They both **swing their hands four times** and finally **raise their right feet**



First person **lifts right leg to strike**, while other person **responds by raising their right leg**



# Qualitative Comparison

| | One person **sneaks up** on **the other** from behind | The first person **raises the right leg aggressively** towards the **second** | Two people **bow** to each other |
|---|---|---|---|
| **InterGen** *IJCV 2024* |  |  |  |
| **in2IN** *CVPRW 2024* |  |  |  |
| **InterMask** *Ours* |  |  |  |

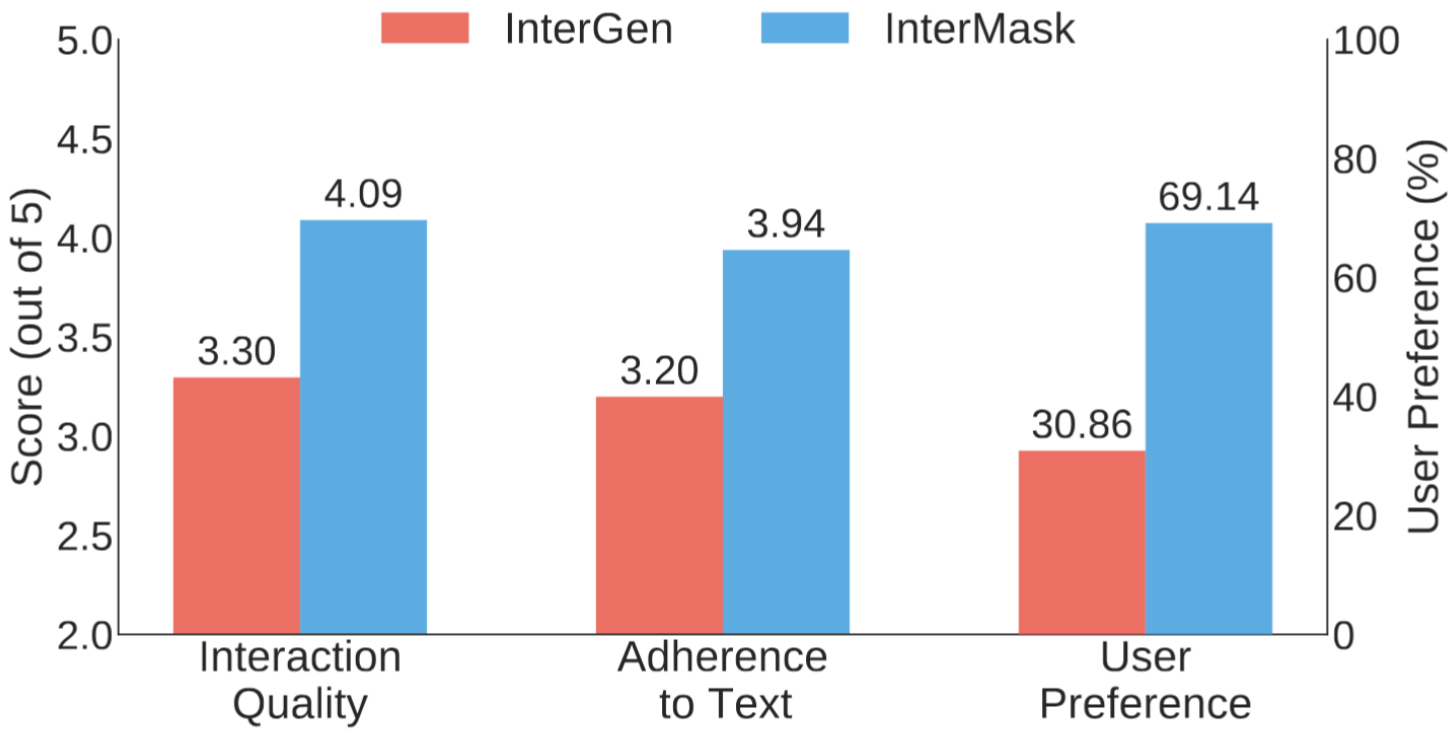
# Quantitative Comparison

| Dataset | Method | R Precision↑ | | | FID↓ | MM Dist↓ | Diversity→ | MModality↑ |
|---|---|---|---|---|---|---|---|---|
| | | Top 1 | Top 2 | Top 3 | | | | |
| Inter Human | 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 | $0.238^{\pm.012}$ | $0.325^{\pm.010}$ | $0.464^{\pm.014}$ | $13.769^{\pm.072}$ | $5.731^{\pm.013}$ | $7.046^{\pm.022}$ | $1.387^{\pm.076}$ |
| | MDM | $0.153^{\pm.012}$ | $0.260^{\pm.009}$ | $0.339^{\pm.012}$ | $9.167^{\pm.056}$ | $7.125^{\pm.018}$ | $7.602^{\pm.045}$ | $\mathbf{2.350}^{\pm.080}$ |
| | ComMDM | $0.223^{\pm.009}$ | $0.334^{\pm.008}$ | $0.466^{\pm.010}$ | $7.069^{\pm.054}$ | $6.212^{\pm.021}$ | $7.244^{\pm.038}$ | $1.822^{\pm.052}$ |
| | InterGen | $0.371^{\pm.010}$ | $0.515^{\pm.012}$ | $0.624^{\pm.010}$ | $5.918^{\pm.079}$ | $5.108^{\pm.014}$ | $7.387^{\pm.029}$ | $2.141^{\pm.063}$ |
| | MoMat-MoGen | $\mathbf{0.449}^{\pm.004}$ | $0.591^{\pm.003}$ | $0.666^{\pm.004}$ | $5.674^{\pm.085}$ | $\mathbf{3.790}^{\pm.001}$ | $8.021^{\pm.35}$ | $1.295^{\pm.023}$ |
| | in2IN | $0.425^{\pm.008}$ | $0.576^{\pm.008}$ | $0.662^{\pm.009}$ | $5.535^{\pm.120}$ | $3.803^{\pm.002}$ | $7.953^{\pm.047}$ | $1.215^{\pm.023}$ |
| | **InterMask** | $\mathbf{0.449}^{\pm.004}$ | $\mathbf{0.599}^{\pm.005}$ | $\mathbf{0.683}^{\pm004}$ | $\mathbf{5.154}^{\pm.061}$ | $\mathbf{3.790}^{\pm.002}$ | $\mathbf{7.944}^{\pm.033}$ | $1.737^{\pm.020}$ |
| InterX | Ground Truth | $0.429^{\pm.004}$ | $0.626^{\pm.003}$ | $0.736^{\pm.003}$ | $0.002^{\pm.0002}$ | $3.536^{\pm.013}$ | $9.734^{\pm.078}$ | - |
| | T2M | $0.184^{\pm.010}$ | $0.298^{\pm.006}$ | $0.396^{\pm.005}$ | $5.481^{\pm.382}$ | $9.576^{\pm.006}$ | $2.771^{\pm.151}$ | $2.761^{\pm.042}$ |
| | MDM | $0.203^{\pm.009}$ | $0.329^{\pm.007}$ | $0.426^{\pm.005}$ | $23.701^{\pm.057}$ | $9.548^{\pm.014}$ | $5.856^{\pm.077}$ | $3.490^{\pm.061}$ |
| | ComMDM | $0.090^{\pm.002}$ | $0.165^{\pm.004}$ | $0.236^{\pm.004}$ | $29.266^{\pm.067}$ | $6.870^{\pm.017}$ | $4.734^{\pm.067}$ | $0.771^{\pm.053}$ |
| | InterGen | $0.207^{\pm.004}$ | $0.335^{\pm.005}$ | $0.429^{\pm.005}$ | $5.207^{\pm.216}$ | $9.580^{\pm.011}$ | $7.788^{\pm.208}$ | $\mathbf{3.686}^{\pm.052}$ |
| | **InterMask** | $\mathbf{0.403}^{\pm.005}$ | $\mathbf{0.595}^{\pm.004}$ | $\mathbf{0.705}^{\pm005}$ | $\mathbf{0.399}^{\pm.013}$ | $\mathbf{3.705}^{\pm.017}$ | $\mathbf{9.046}^{\pm.073}$ | $2.261^{\pm.081}$ |

## Insights

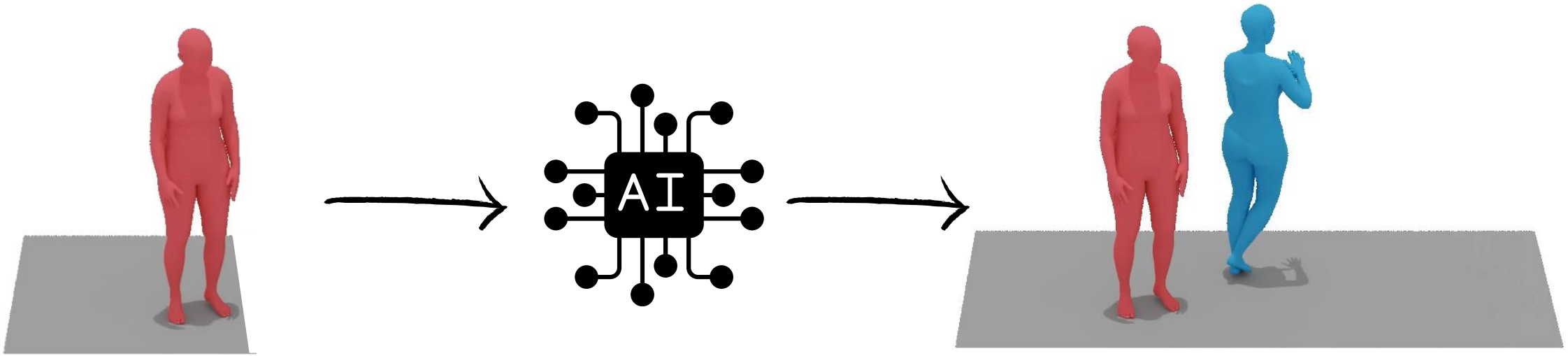
| | |
|---|---|
| Lower FID | Superior Fidelity and Quality |
| Higher R Prec & Lower MMDist | Better Adherence to Text Prompts |
| Optimal Diversity | Wide Range of Interaction Scenarios |
| Optimal Diversity but Lower MModality | Prioritizes Text Adherence over Variety |
| Higher Performance Gain on InterX | Superior Scalability with more Data |

# User Study


InterGen
InterMask
Score (out of 5)
5.0
4.5
4.0
3.5
3.0
2.5
2.0
User Preference (%)
100
80
60
40
20
0
3.30
4.09
3.20
3.94
30.86
69.14
Interaction Quality
Adherence to Text
User Preference

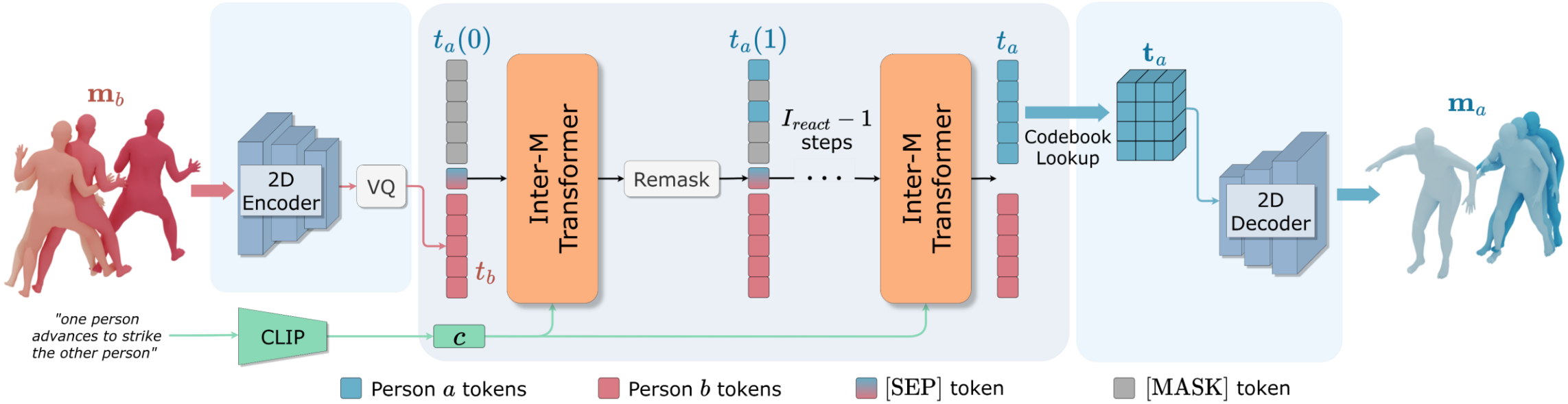
# Application: Human Reaction Generation

Enable reaction generation for smart digital character response in virtual environments



# Reaction Generation Inference

InterMask seamlessly supports Reaction Generation



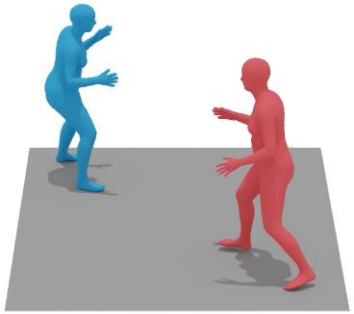
# Reaction Generation Results

## Qualitative

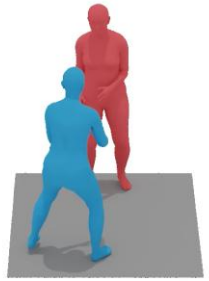
Reference Input Motion

Generated Output Motion

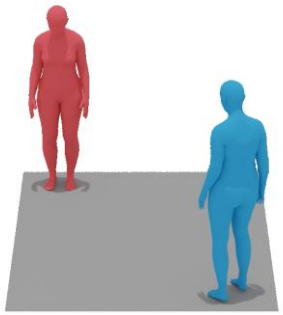
*without text description*



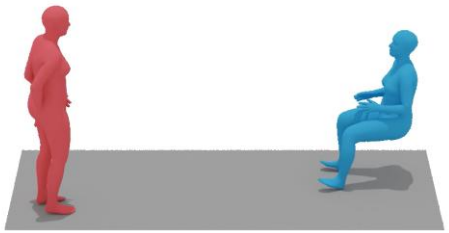
These two **raise their left hands** and extend them towards the left



One person **approaches the other**



One person **takes 4 steps towards the other**, while the other is **sitting on a chair holding a piece of paper**



# Conclusion

- **Goal**:
  - Generative AI framework to model interactions between two digital human characters for applications in animations, VR.
- **Primary Contribution**:
  - A novel masked generative framework for collaborative modeling of two-person interactions.
- **Technical Advancements**:
  - A novel 2D motion token map preserving both spatial and temporal dimensions of
  - Inter-M Transformer, a specialized model with custom attention modules and masking techniques.
- **Results**:
  - State-of-the-art performance with improved interaction quality, spatial coherence and text adherence.
  - Seamless Reaction Generation Support.

# Thank You!