InterControl: Zero-shot Human Interaction Generation by Controlling Every Joint

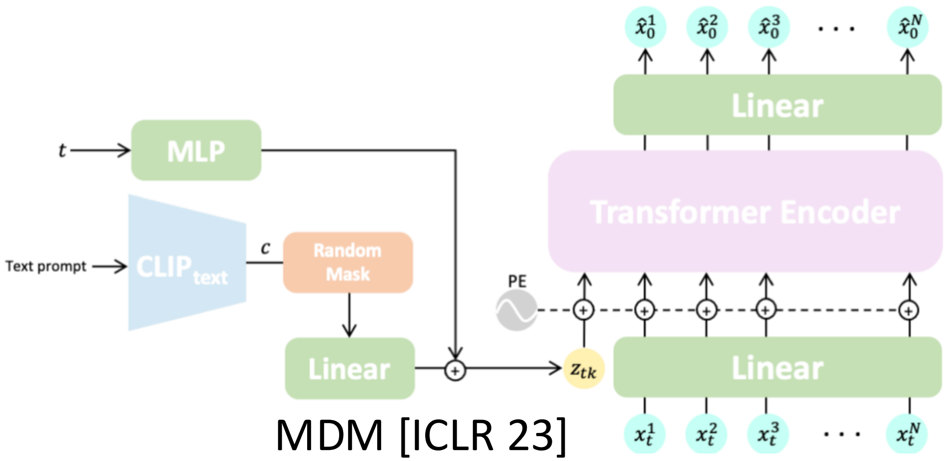
NeurIPS 2024

Zhenzhi Wang, Jingbo Wang, Yixuan Li, Dahua Lin, Bo Dai

CUHK, Shanghai AI Lab, HKU

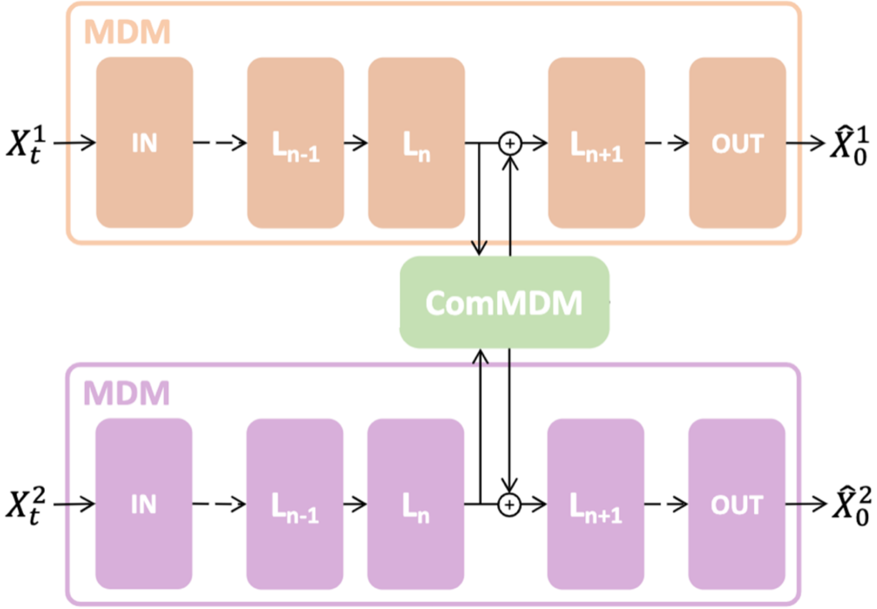
Code: https://github.com/zhenzhiwang/intercontrol

Human Motion Generation

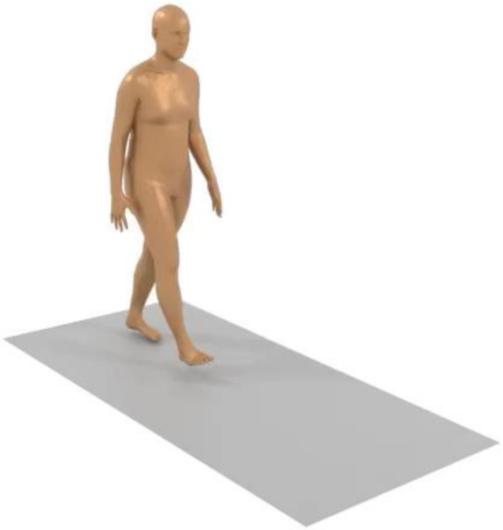


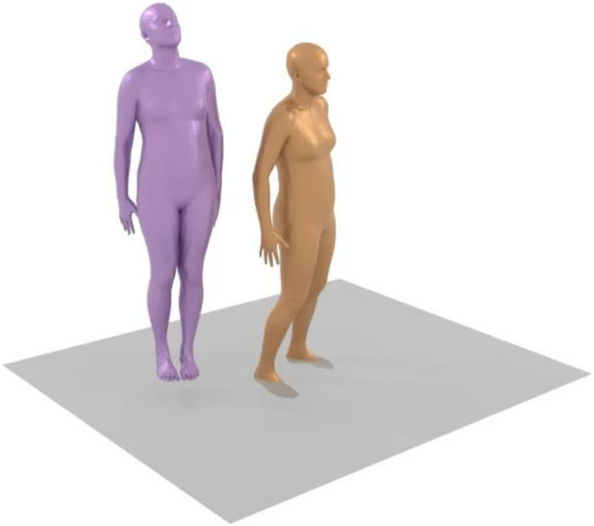
$\hat{x}_0^1$ $\hat{x}_0^2$ $\hat{x}_0^3$ . . . $\hat{x}_0^N$
Linear
t
MLP
Transformer Encoder
Text prompt
CLIP$_{text}$
c
Random Mask
PE
Linear
z_{tk}
Linear
x_t^1 x_t^2 x_t^3 . . . x_t^N

MDM [ICLR 23]

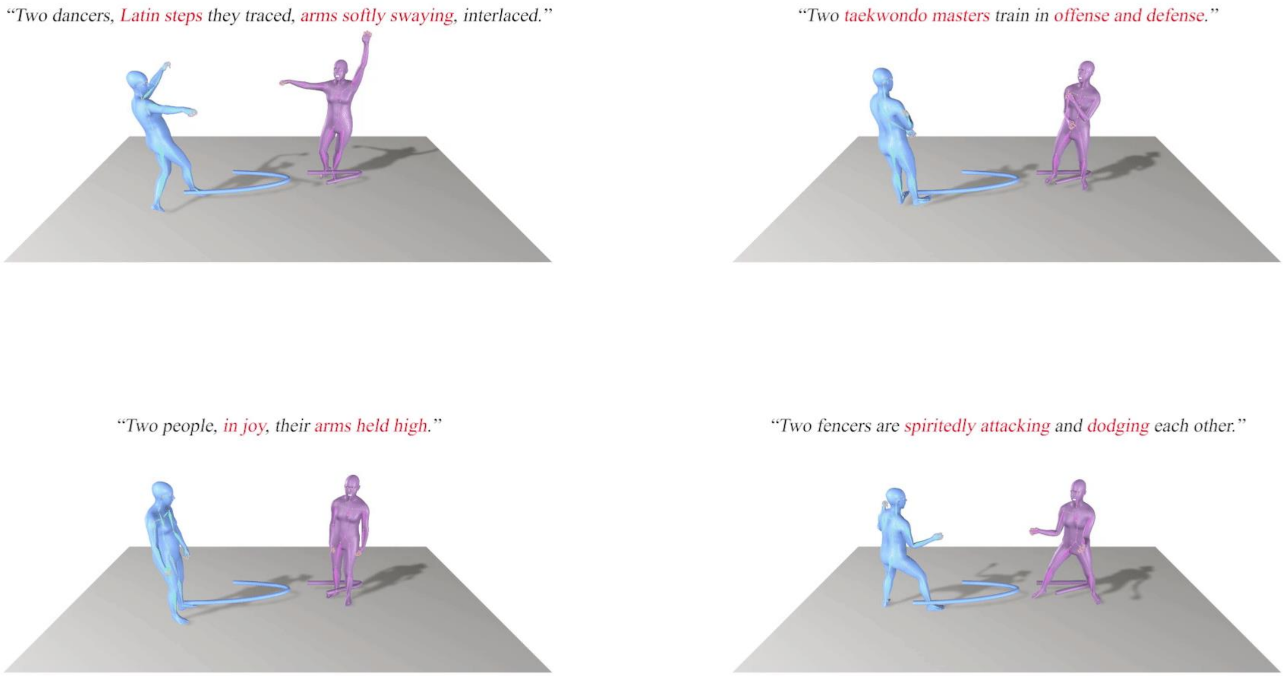


PriorMDM [ICLR 24]





Multi-Person Interaction



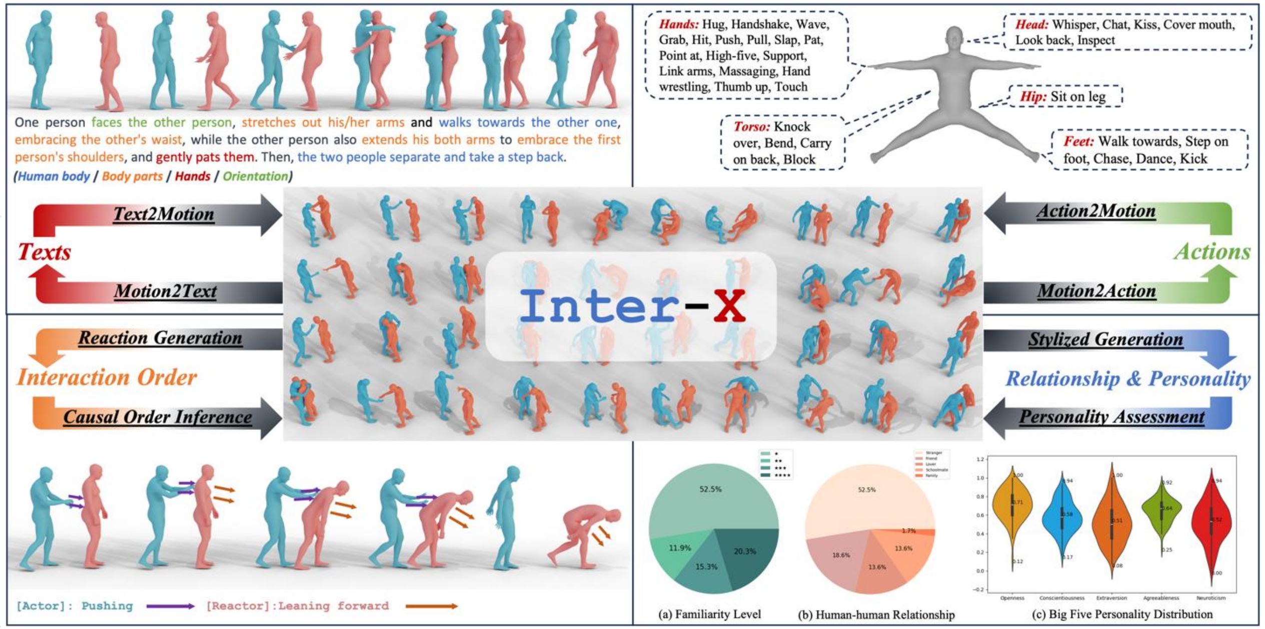
“Two dancers, Latin steps they traced, arms softly swaying, interlaced.”

“Two taekwondo masters train in offense and defense.”

“Two people, in joy, their arms held high.”

“Two fencers are spiritedly attacking and dodging each other.”

InterHuman [IJCV24]



One person faces the other person, stretches out his/her arms and walks towards the other one, embracing the other's waist, while the other person also extends his both arms to embrace the first person's shoulders, and gently pats them. Then, the two people separate and take a step back.

(Human body / Body parts / Hands / Orientation)

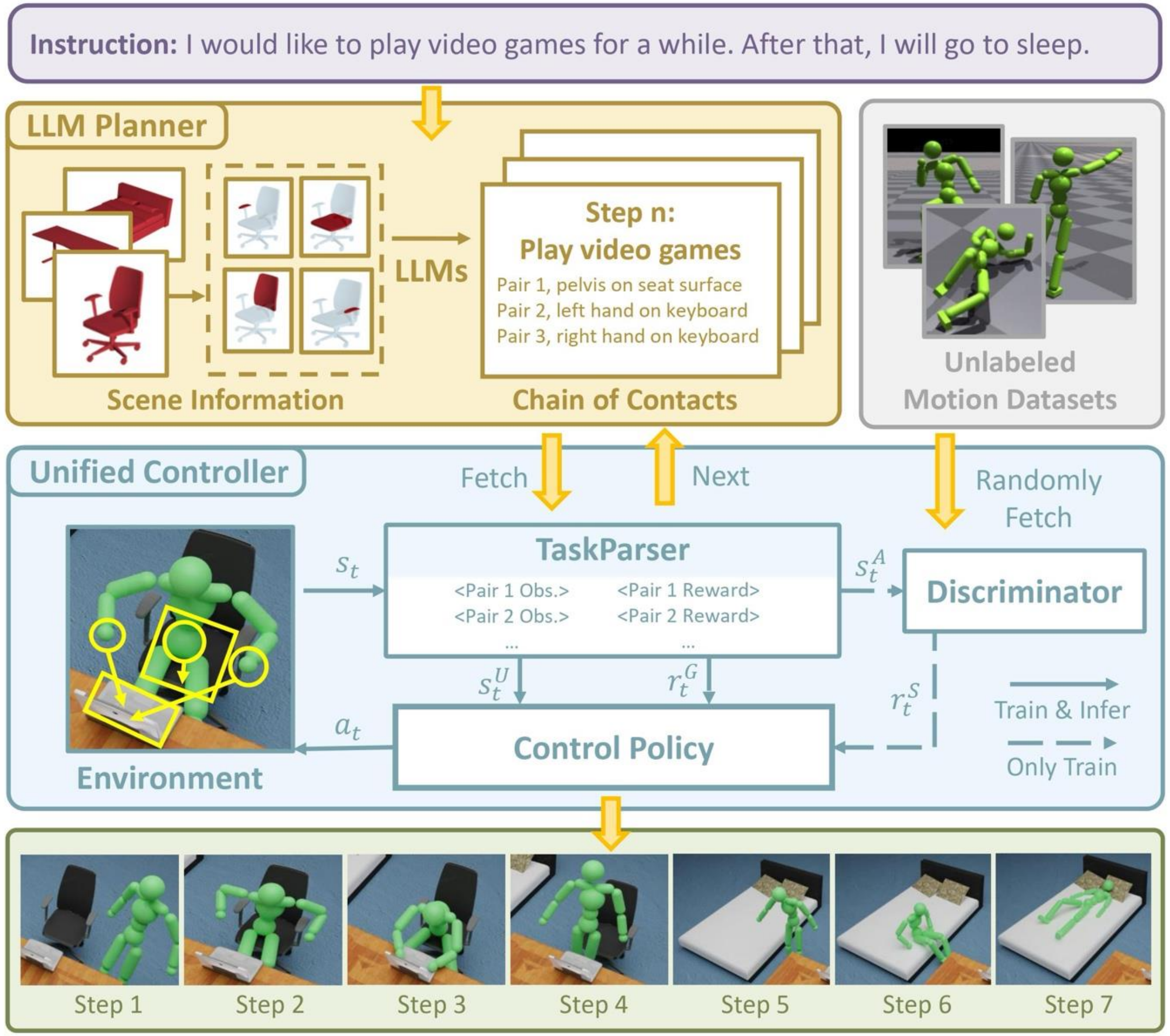
(a) Familiarity Level (b) Human-human Relationship (c) Big Five Personality Distribution

Inter-X [CVPR 24]

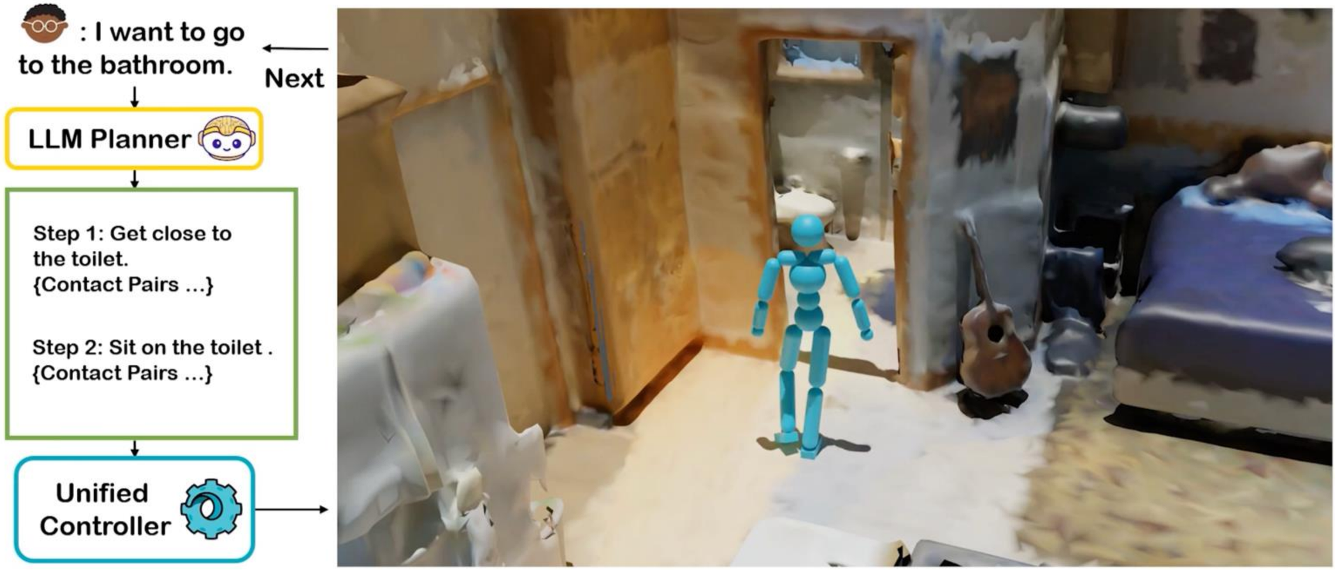
Our work was done in Nov, 2023. Thus, back to that time, could we generate multi-person interactions only by single-person motion data?

Advantages: the interaction will not be restricted by only two people defined by the dataset.

Definition of Interaction



Instruction: I would like to play video games for a while. After that, I will go to sleep.



: I want to go to the bathroom.

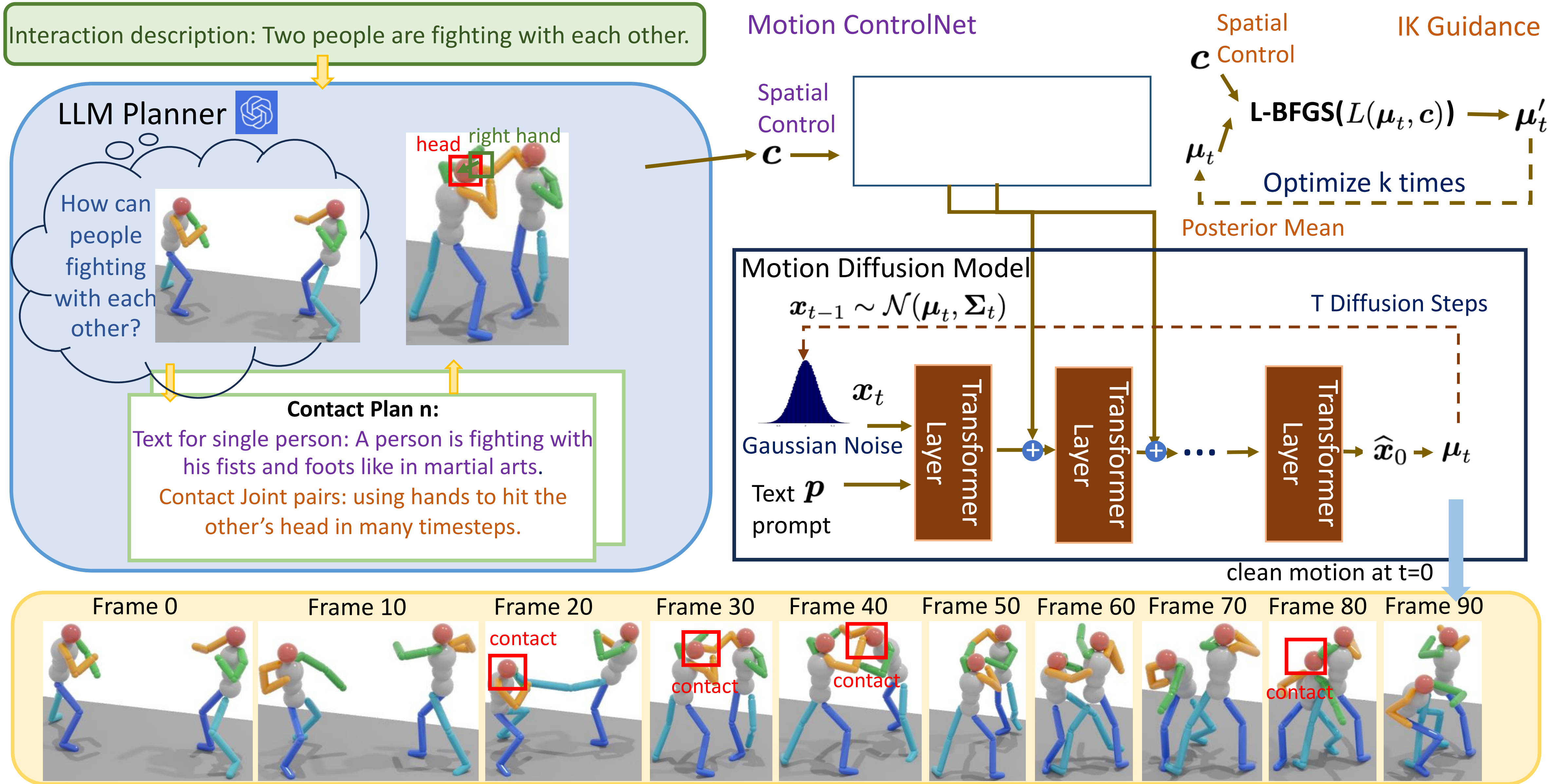
Step 1: Get close to the toilet.
{Contact Pairs ...}

Step 2: Sit on the toilet .
{Contact Pairs ...}

UniHSI [ICLR24]

Yes! We find that interaction could be represented by joint distances. And the semantics of interactions could be understood by Large Language Models.

Spatial Control



Interaction description: Two people are fighting with each other.

LLM Planner

How can people fighting with each other?

head

right hand

Contact Plan n:

Text for single person: A person is fighting with his fists and foots like in martial arts.

Contact Joint pairs: using hands to hit the other's head in many timesteps.

Motion ControlNet

Spatial Control $\boldsymbol{c}$

Spatial Control $\boldsymbol{c}$

IK Guidance

L-BFGS($L(\boldsymbol{\mu}_t, \boldsymbol{c})$) → $\boldsymbol{\mu}'_t$

$\boldsymbol{\mu}_t$

Optimize k times

Posterior Mean

Motion Diffusion Model

$\boldsymbol{x}_{t-1} \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$

T Diffusion Steps

$\boldsymbol{x}_t$

Gaussian Noise

Text $\boldsymbol{p}$ prompt

Transformer Layer

Transformer Layer

Transformer Layer

$\widehat{\boldsymbol{x}}_0$ → $\boldsymbol{\mu}_t$

clean motion at t=0

Frame 0, Frame 10, Frame 20, Frame 30, Frame 40, Frame 50, Frame 60, Frame 70, Frame 80, Frame 90

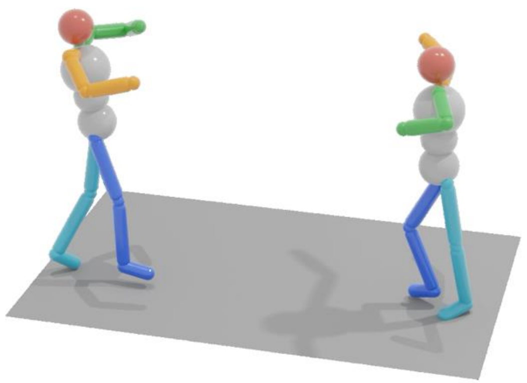
contact

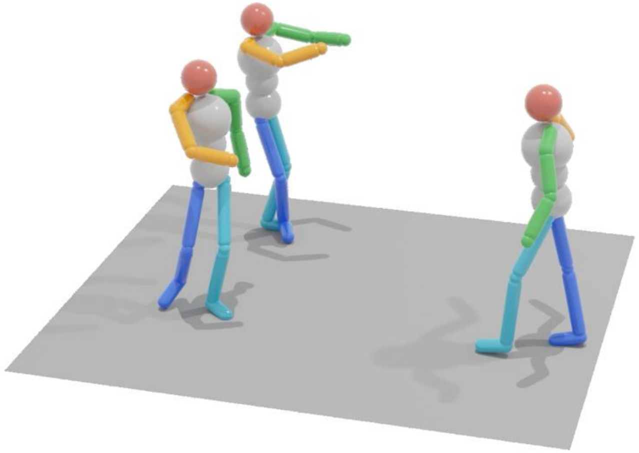
contact

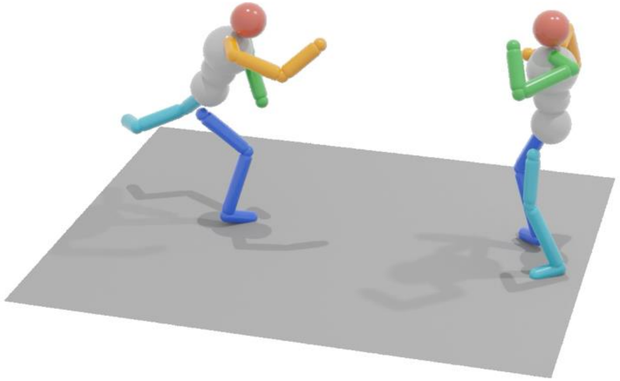
contact

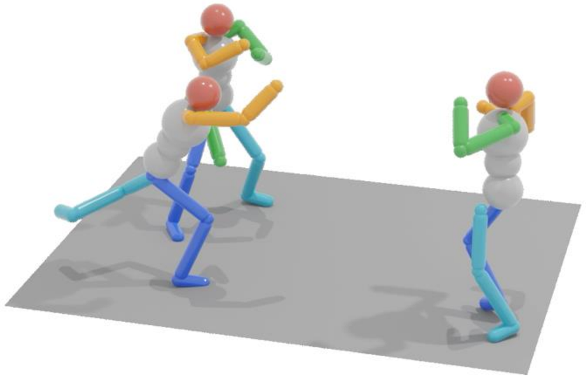
contact

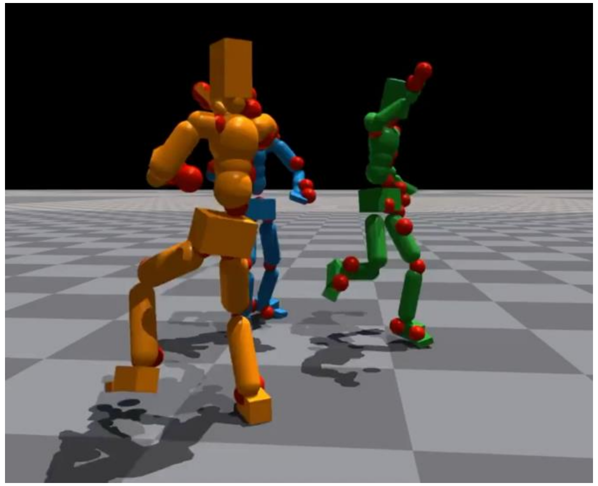
Zero-shot Interaction Generation



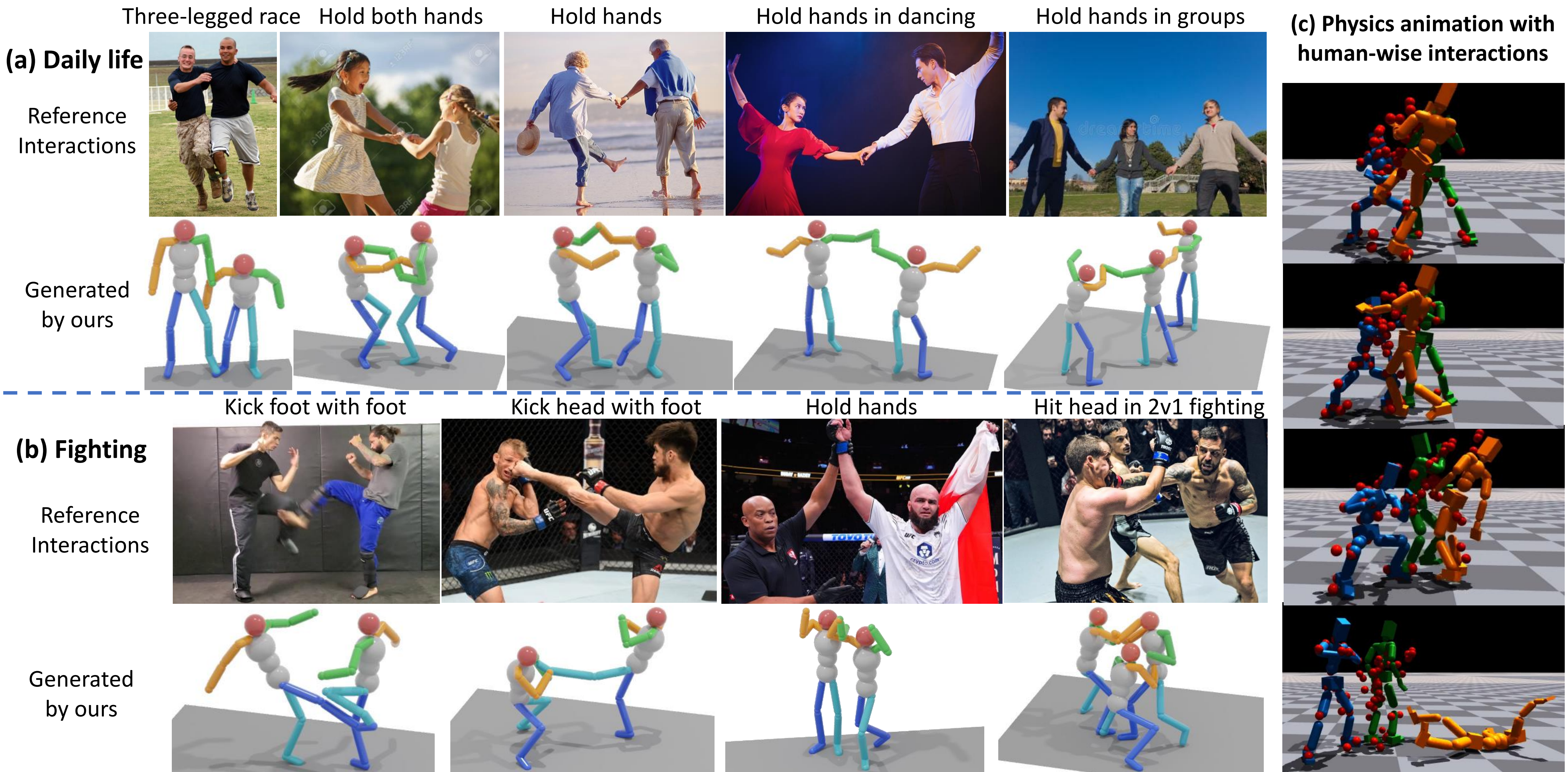








Our interaction are realistic



Quantitative comparisons

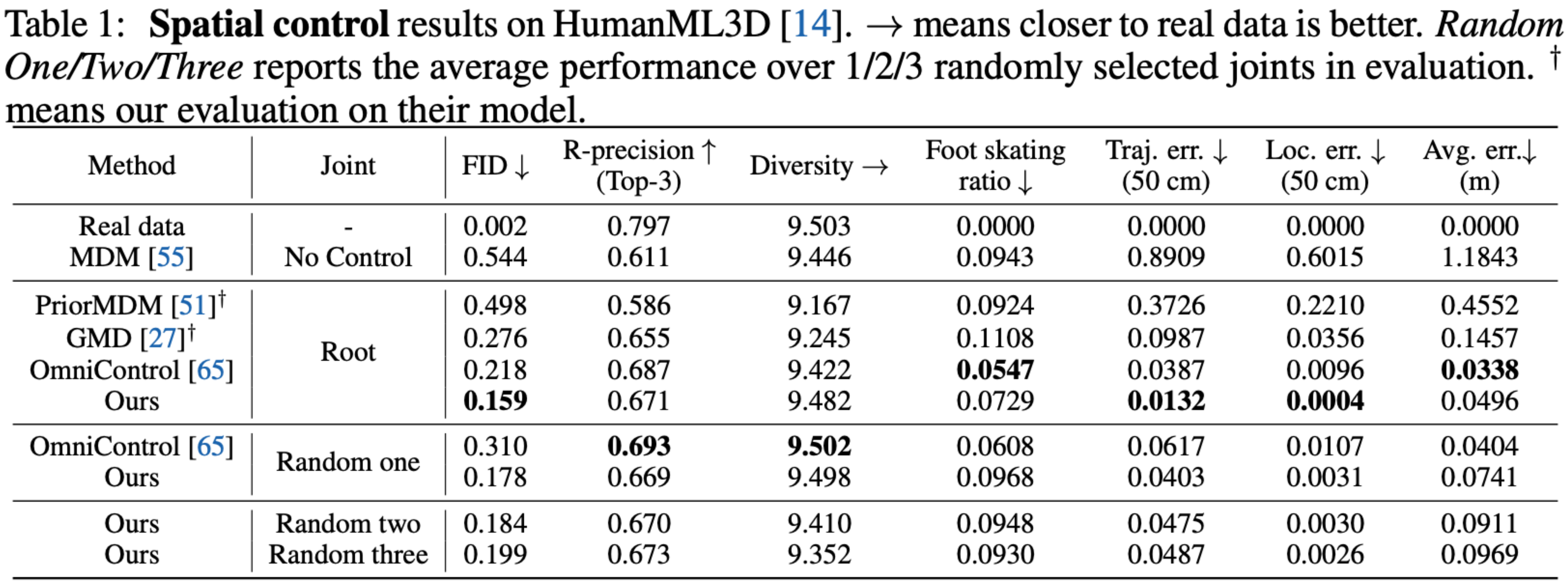


Table 1: **Spatial control** results on HumanML3D [14]. → means closer to real data is better. *Random One/Two/Three* reports the average performance over 1/2/3 randomly selected joints in evaluation. † means our evaluation on their model.

Method	Joint	FID ↓	R-precision ↑ (Top-3)	Diversity →	Foot skating ratio ↓	Traj. err. ↓ (50 cm)	Loc. err. ↓ (50 cm)	Avg. err.↓ (m)
Real data	-	0.002	0.797	9.503	0.0000	0.0000	0.0000	0.0000
MDM [55]	No Control	0.544	0.611	9.446	0.0943	0.8909	0.6015	1.1843
PriorMDM [51]†	Root	0.498	0.586	9.167	0.0924	0.3726	0.2210	0.4552
GMD [27]†		0.276	0.655	9.245	0.1108	0.0987	0.0356	0.1457
OmniControl [65]		0.218	0.687	9.422	**0.0547**	0.0387	0.0096	**0.0338**
Ours		**0.159**	0.671	9.482	0.0729	**0.0132**	**0.0004**	0.0496
OmniControl [65]	Random one	0.310	**0.693**	**9.502**	0.0608	0.0617	0.0107	0.0404
Ours		0.178	0.669	9.498	0.0968	0.0403	0.0031	0.0741
Ours	Random two	0.184	0.670	9.410	0.0948	0.0475	0.0030	0.0911
Ours	Random three	0.199	0.673	9.352	0.0930	0.0487	0.0026	0.0969

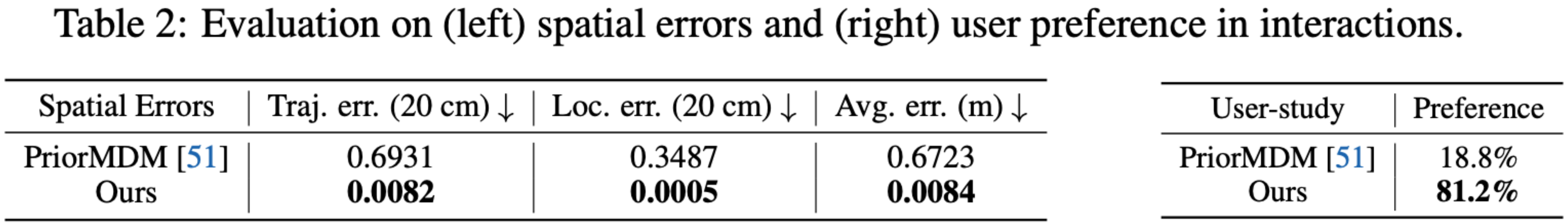


Table 2: Evaluation on (left) spatial errors and (right) user preference in interactions.

Spatial Errors	Traj. err. (20 cm) ↓	Loc. err. (20 cm) ↓	Avg. err. (m) ↓
PriorMDM [51]	0.6931	0.3487	0.6723
Ours	**0.0082**	**0.0005**	**0.0084**

User-study	Preference
PriorMDM [51]	18.8%
Ours	**81.2%**

Quantitative comparisons

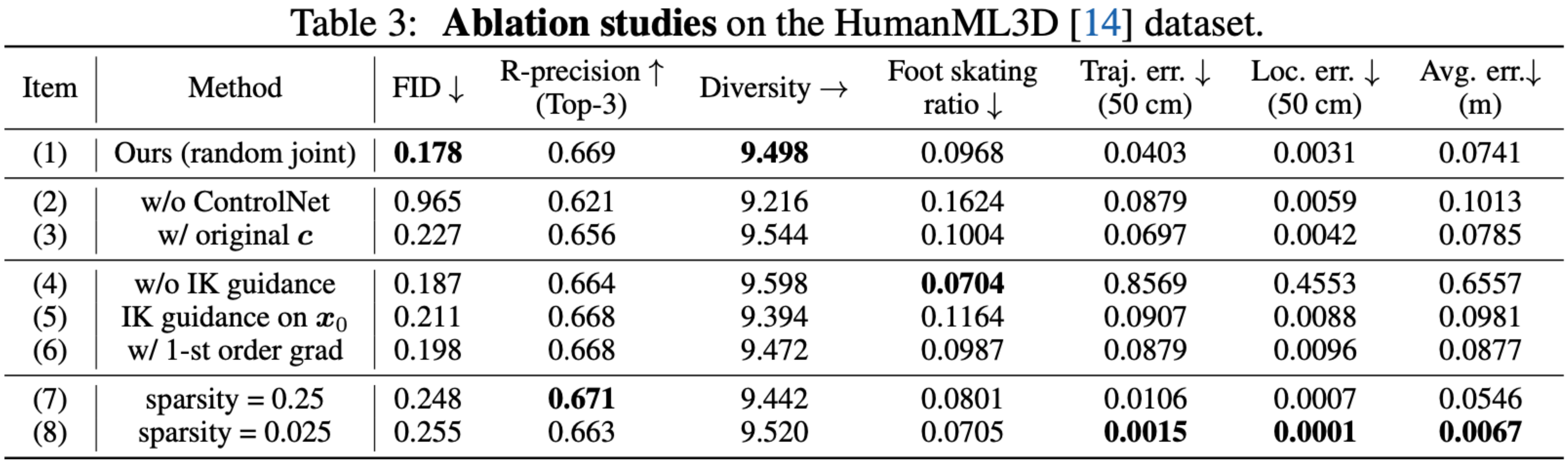


Table 3: **Ablation studies** on the HumanML3D [14] dataset.

Item	Method	FID ↓	R-precision ↑ (Top-3)	Diversity →	Foot skating ratio ↓	Traj. err. ↓ (50 cm)	Loc. err. ↓ (50 cm)	Avg. err.↓ (m)
(1)	Ours (random joint)	**0.178**	0.669	**9.498**	0.0968	0.0403	0.0031	0.0741
(2)	w/o ControlNet	0.965	0.621	9.216	0.1624	0.0879	0.0059	0.1013
(3)	w/ original $\boldsymbol{c}$	0.227	0.656	9.544	0.1004	0.0697	0.0042	0.0785
(4)	w/o IK guidance	0.187	0.664	9.598	**0.0704**	0.8569	0.4553	0.6557
(5)	IK guidance on $\boldsymbol{x}_0$	0.211	0.668	9.394	0.1164	0.0907	0.0088	0.0981
(6)	w/ 1-st order grad	0.198	0.668	9.472	0.0987	0.0879	0.0096	0.0877
(7)	sparsity = 0.25	0.248	**0.671**	9.442	0.0801	0.0106	0.0007	0.0546
(8)	sparsity = 0.025	0.255	0.663	9.520	0.0705	**0.0015**	**0.0001**	**0.0067**

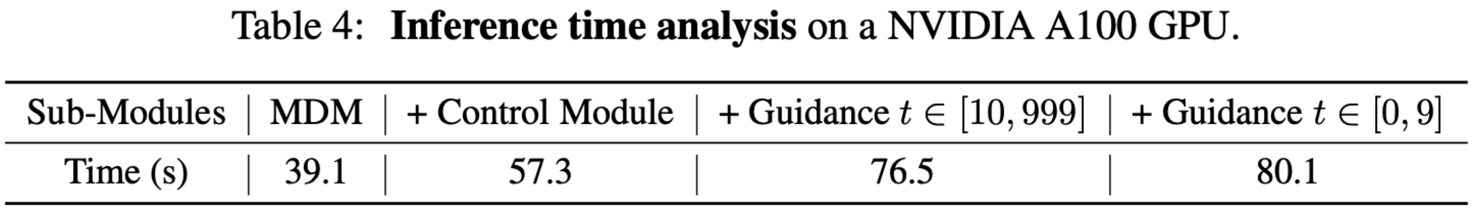
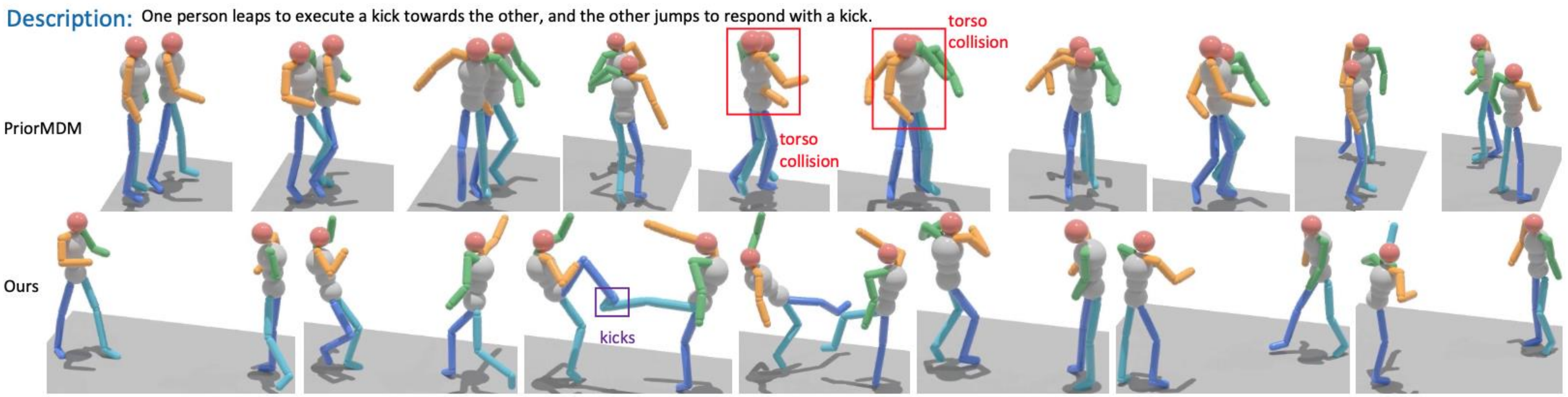


Table 4: **Inference time analysis** on a NVIDIA A100 GPU.

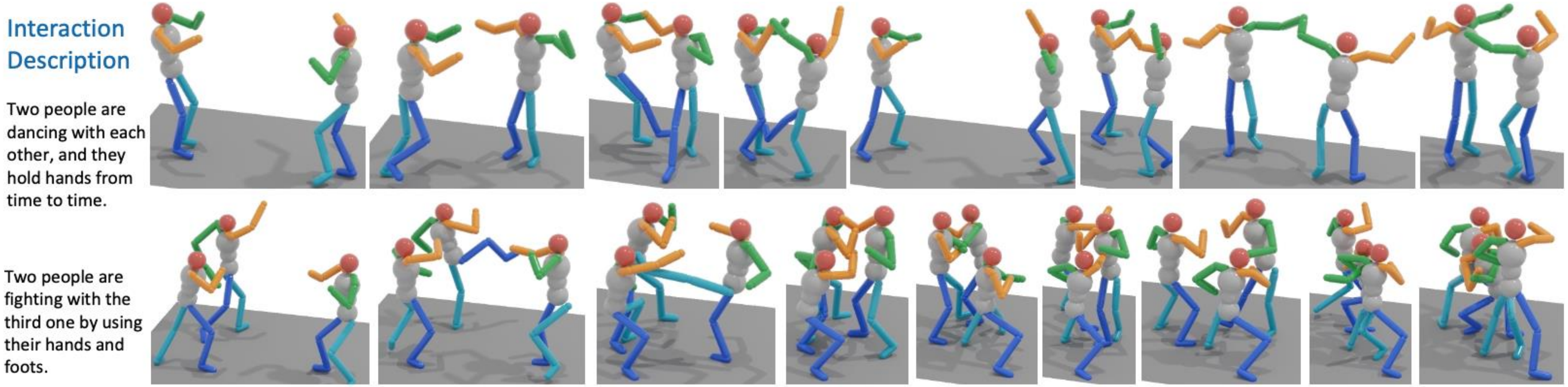
Sub-Modules	MDM	+ Control Module	+ Guidance $t \in [10, 999]$	+ Guidance $t \in [0, 9]$
Time (s)	39.1	57.3	76.5	80.1

Qualitative comparisons



Description: One person leaps to execute a kick towards the other, and the other jumps to respond with a kick.

Figure 3: Comparison with PriorMDM [51] in **user-study** of zero-shot human interaction generation.



Interaction Description

Two people are dancing with each other, and they hold hands from time to time.

Two people are fighting with the third one by using their hands and foots.

Figure 4: **Qualitative results** of zero-shot human interaction generation.

Thank you!