

MixerMDM: Learnable Composition of Human Motion Diffusion Models

Supplementary Material

This supplementary material aims to enhance the reproducibility and understanding of the previously presented contributions. In Sec. A, we detail the datasets used and describe the modifications made to integrate them into the MixerMDM pipeline. In Sec. B, we outline the implementation specifics of the state-of-the-art models used for comparison, as well as the evaluators employed in the proposed evaluation pipeline. In Sec. C, we complement the quantitative evaluation with additional experiments and ablations. Lastly, in Sec. D, we include additional visual examples illustrating the MixerMDM capabilities.

A. Datasets

A.1. InterHuman

InterHuman [27] is one of the most extensive annotated datasets for human-human interactions, containing 7,779 interactions labeled with textual descriptions. Each individual’s motion within an interaction is represented as a set of poses $x_i = [j^p, j^v, j^r, c^f]$, where x^i denotes the i -th motion timestep. This representation includes joint positions and velocities $j^p, j^v \in \mathbb{R}^{3N_j}$ in the world frame, a 6D representation of local rotations $j^r \in \mathbb{R}^{6N_j}$ in the root frame, and binary foot-ground contact features $c^f \in \mathbb{R}^4$. The number of joints in the InterHuman dataset is $N_j = 22$. Each interaction in the dataset is paired with three textual descriptions summarizing the overall interaction. Additionally, the in2IN [39] framework introduced more detailed textual descriptions, generated by Large Language Models, for the motions performed by each individual in the interaction. We utilized these detailed descriptions to condition the generation of individual models employed in the mixing process.

A.2. HumanML3D

The HumanML3D [15] dataset contains 14,616 individual motions annotated with textual descriptions. Each motion is represented as a set of poses $x_i = [\dot{r}^a, \dot{r}^x, \dot{r}^z, r^y, j^p, j^v, j^r, c^f]$, where x^i denotes the i -th motion timestep. In this format, $\dot{r}^a \in \mathbb{R}$ is the root angular velocity in the Y-axis, $\dot{r}^x, \dot{r}^z \in \mathbb{R}$ are the root linear velocities on the XZ-plane, $r^y \in \mathbb{R}$ is the root height, $j^p, j^v \in \mathbb{R}^{3N_j}$ and $j^r \in \mathbb{R}^{6N_j}$ are the joint positions, velocities, and rotations in the root frame, and $c^f \in \mathbb{R}^4$ are binary foot-ground contact features. The number of joints in the HumanML3D dataset is $N_j = 22$. Since this representation differs from the format used in InterHuman and is not optimal for capturing the relative positions of interactants, we have converted it to the InterHuman format. This conversion involves processing the raw SMPL

motions from HumanML3D to extract the global joint positions and velocities as well as the relative rotations as it is done in the InterHuman pre-processing.

B. Further Implementation Details

B.1. Motion Transformations

Motion transformations are applied to maintain the pre-trained models within their learned distribution. The *centering* function translates the motion to the origin of coordinates in the XZ plane, simultaneously orienting the trajectory initially in the Z+ direction. The *alignment* is a global transformation applied to a motion (x^a) with respect to another (x^b). Firstly, x^a is translated to the initial position of x^b . Secondly, x^a is rotated to match the orientation of the vector of the initial and end positions of x^b . The same transformation is applied to the whole motion, thus not introducing foot sliding, and standardizes global positioning and orientation to the individual models.

B.2. State-of-the-art Implementations

The methods employed in Sec. 4 were implemented using their respective official codebases. We leveraged the original checkpoints of InterGen [27] and the interaction and individual versions of in2IN [39], as they were trained with the same motion representation we use. For MDM [48], originally trained on HumanML3D, we kept the architecture as is, and adapted the output shape of the denoiser to match the size of our motion representation.

B.3. Evaluators

The evaluation metrics for human motion generation require a feature extractor that produces aligned latent representations of both the generated motions and the corresponding conditions (text, in this case). The feature extractor architecture is based on the one used in the InterHuman dataset: a *MotionEncoder* and a *TextEncoder*. The *MotionEncoder* is a Transformer encoder with 8 layers of 8 heads each, which transforms the motion into a 2048-dimensional latent vector. This vector is then compressed to a dimension of 512 using a Multi-Layer Perceptron. The *TextEncoder* is a frozen CLIP-ViT/14 model [38], supplemented with a Transformer encoder with 8 layers of 8 heads each to adapt the CLIP latent space to better match the dataset distribution. We trained a feature extractor for each dataset employed. These models were trained for 500 epochs with a batch size of 64, using the AdamW optimizer [30] with β parameters set to (0.9, 0.999), a weight decay of 10^{-5} , and a learning rate of 10^{-4} .

Method	Type	R-Precision \uparrow		FID \downarrow		MM Dist \downarrow		Diversity \rightarrow		MModality \uparrow	
		Interaction	Individual	Interaction	Individual	Interaction	Individual	Interaction	Individual	Interaction	Individual
Ground Truth	-	.701 \pm .01	.563 \pm .00	.273 \pm .01	1.04 \pm .14	3.76 \pm .01	3.44 \pm .00	7.95 \pm .06	16.3 \pm .05	-	-
Diff.Blending [41]	-	.577 \pm .00	.137 \pm .02	33.8 \pm .29	360 \pm 16	3.89 \pm .00	5.18 \pm .01	6.14 \pm .14	11.9 \pm .22	.779 \pm .12	2.74 \pm .01
DualMDM [39]	-	.574 \pm .00	.134 \pm .01	22.9 \pm .19	330 \pm .02	3.85 \pm .01	5.13 \pm .01	7.04 \pm .17	12.5 \pm .28	.935 \pm .12	2.74 \pm .09
MixerMDM (ours)	G	.521 \pm .00	.228 \pm .01	44.5 \pm .99	199 \pm 19	3.92 \pm .00	4.70 \pm .14	6.57 \pm .19	14.0 \pm .53	1.08 \pm .19	2.99 \pm .12
	T	.672\pm.02	.150 \pm .02	21.2\pm.70	245 \pm 5.1	3.85\pm.00	5.05 \pm .00	7.57\pm.07	13.6 \pm .16	1.14 \pm .25	3.04 \pm .21
	S	.391 \pm .01	.257 \pm .01	52.4 \pm 1.8	192 \pm 5.9	3.94 \pm .00	4.69 \pm .03	6.40 \pm .20	14.4 \pm .09	1.11 \pm .02	2.96 \pm .07
	ST	.406 \pm .01	.286\pm.01	47.6 \pm .88	142\pm.75	3.93 \pm .01	4.60\pm.02	6.57 \pm .09	15.1\pm.18	1.23\pm.02	3.26\pm.07

Table A. **Quantitative evaluation with conventional metrics.** Ours $\{G,T,S,TS\}$ uses $\mathcal{M}^a=\mathcal{M}^b=\text{in2IN}$. Mean of 10 evaluations, \pm shows the 95% confidence interval. Best in **bold**.

C. Further Quantitative Examples

In this section, we complement the quantitative study from Sec. 4.4 with additional experiments and ablations. In addition to the proposed metrics, we have evaluated MixerMDM with conventional metrics. Tab. A shows that our method surpasses previous methods on the proposed and conventional metrics.

C.1. Motion Transformations

The *centering* and *alignment* transformation (Sec. 3.1, Sec. B.1) help to maintain the mixed motion within the distribution of the pre-trained models. Tab. B shows the effect of not using the alignment transformation. While producing a performance drop in the interaction evaluation, results still outperform previous methods.

Method	Top-3 R-Prec. \uparrow	FID \downarrow	MM Dist \downarrow	Diversity \rightarrow	MModality \uparrow
	Interaction	Interaction	Interaction	Interaction	Interaction
G	.391 \pm .01	45.7 \pm .31	3.92 \pm .01	6.51 \pm .08	1.15\pm.02
T	.578\pm.00	20.2\pm.04	3.84\pm.00	7.73\pm.09	.990 \pm .01
S	.375 \pm .01	50.6 \pm 1.0	3.93 \pm .01	6.43 \pm .01	1.01 \pm .15
ST	.380 \pm .02	41.1 \pm .00	3.93 \pm .01	6.62 \pm .05	1.01 \pm .01

Table B. **MixerMDM without alignment.** Compare with Tab. A.

C.2. Usability

While using more prompts allows higher controllability, it can hinder usability with tedious descriptions in cases where such controllability is not a priority. Using an LLM (gpt4o-mini) at inference allows using just the interaction prompt and inferring the individual ones. Tab. C shows that this strategy does not affect the interaction motion quality and text-alignment.

Method	Top-3 R-Prec. \uparrow	FID \downarrow	MM Dist \downarrow	Diversity \rightarrow	MModality \uparrow
	Interaction	Interaction	Interaction	Interaction	Interaction
G	.451 \pm .00	46.7 \pm .18	3.92 \pm .01	6.61 \pm .12	1.03 \pm .18
T	.651\pm.02	21.2\pm.77	3.83\pm.01	7.69\pm.17	.998 \pm .00
S	.412 \pm .03	49.3 \pm 1.2	3.93 \pm .01	6.40 \pm .09	1.11\pm.00
ST	.341 \pm .00	49.1 \pm 1.4	3.94 \pm .00	6.39 \pm .04	1.03 \pm .05

Table C. **MixerMDM LLM aided.** Compare with Tab. A.

D. Further Qualitative Examples

In this section, we complement the qualitative study from Sec. 4.4 with additional examples. Fig. B show the superior individual controllability of MixerMDM when compared to previous approaches. With MixerMDM, we can achieve detailed control of the individual motions while still preserving the dynamics of the interaction. This is achieved thanks to the adversarial training that promotes preserving the main interaction dynamics as well as the individual ones. Fig. A shows another example of consistency on this dual control. We refer the reader to the attached video for a more detailed visualization of all the examples that we showed and discussed in this section.

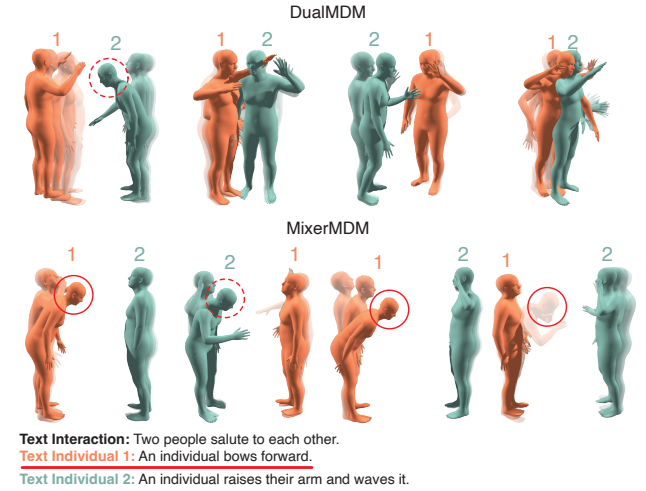


Figure A. **Consistency.** When an individual variation (underlined) is performed in one of the interactions, MixerMDM achieves a greater level of consistency generating the mixed motion.

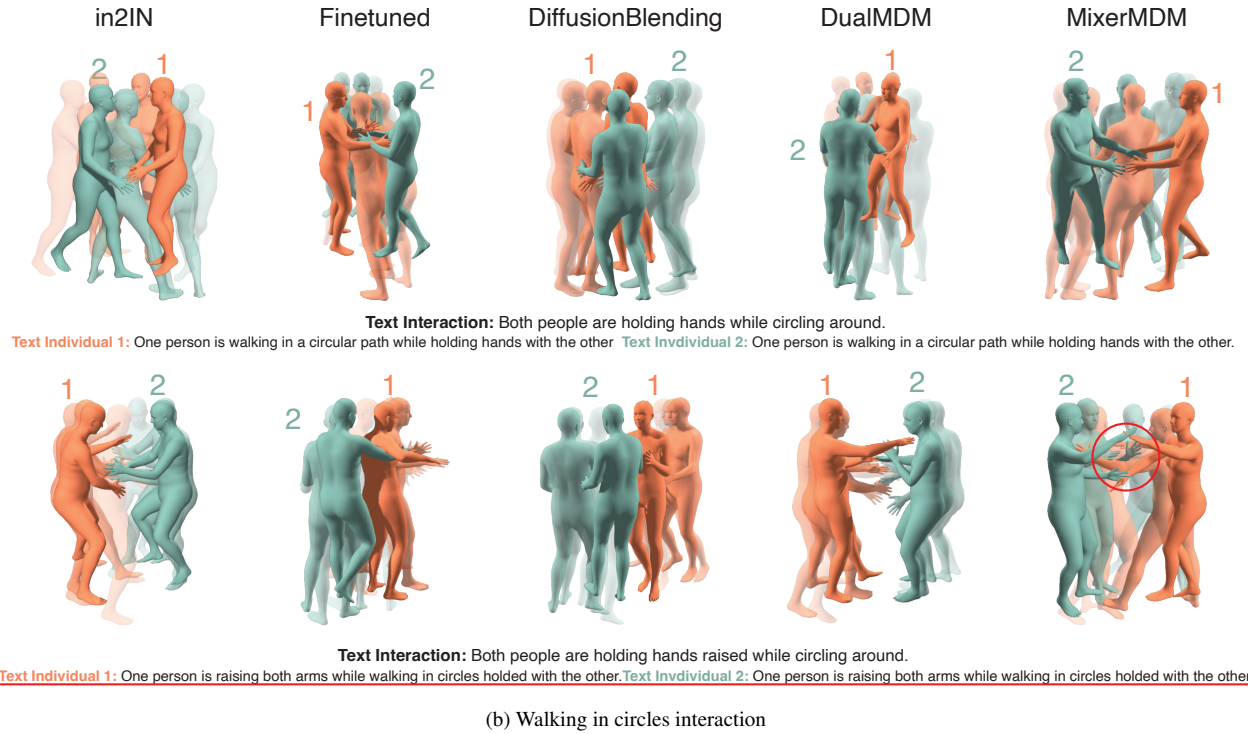
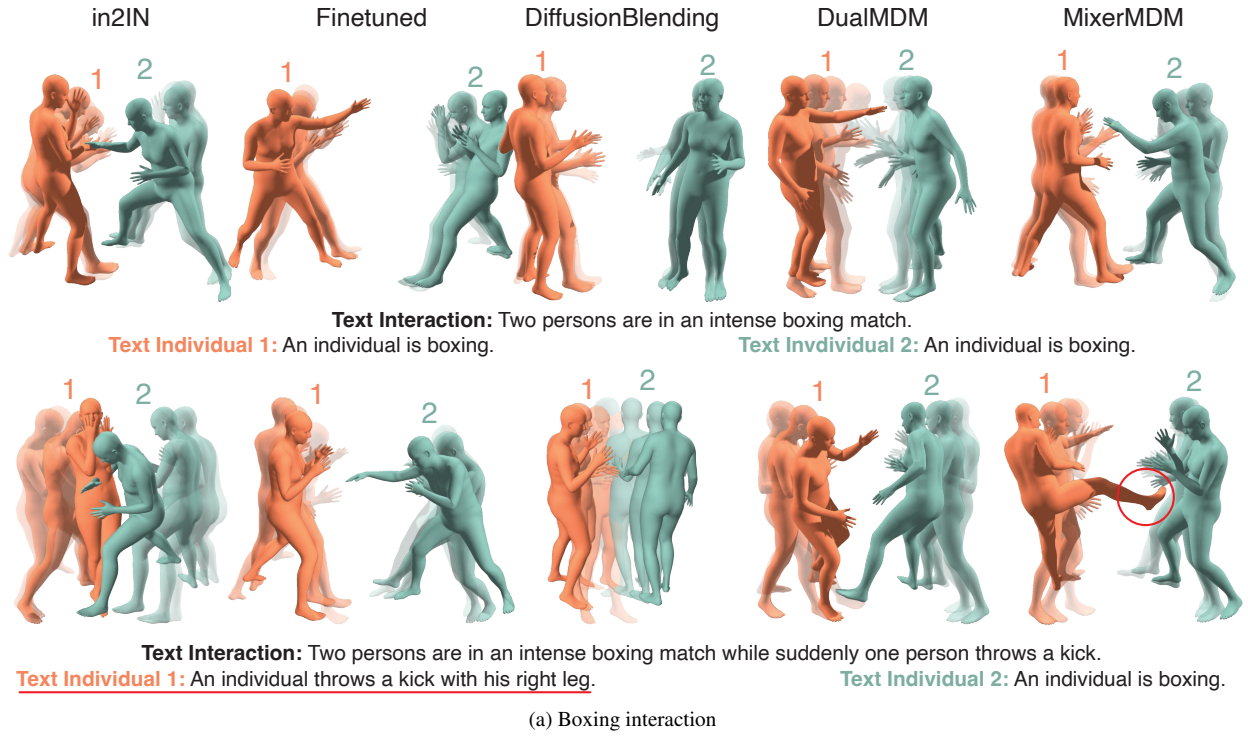


Figure B. **Controllability.** While all methods can properly generate an interaction (top), when a variation in one of the individual conditions is applied (bottom, underlined), MixerMDM generates the most aligned motion to the overall interaction and individual textual descriptions.