

# in2IN: Leveraging individual Information to Generate Human Interactions

## Supplementary Material

In this chapter, we will provide additional information and examples about the main paper's contents that have not been included due to space reasons. This additional information will help the reproducibility and further understanding of the contributions that we have previously presented. In Sec. A we will explain how we have generated the individual descriptions for the InterHuman dataset. In Sec. B we will explain the implementation details and results on the individual prior that we have used in DualMDM. Finally, in Sec. C we will present additional examples from the qualitative evaluation of the in2IN and DualMDM contributions.

### A. Extended InterHuman dataset

The InterHuman dataset contains a significant amount of annotated human-human interactions. However, the textual descriptions of the interactions are not focused on the specific individual motions performed by the integrants of the interaction. As our in2IN proposal needs these individual descriptions, we have generated them using LLMs. From the original interaction descriptions we generate the individual ones using the following prompt:

Having the description of an interaction, extract descriptions for the motions of each individual.

–

**Interaction Description:** In an intense boxing match, one person attacks the opponent with a straight punch, and then the opponent falls over.

**Individual Motion 1:** One person is moving and then throws a punch.

**Individual Motion 2:** One person falls over and stays on the ground.

–

**Interaction Description:** <interaction motion description>

The LLM used for this task is gpt3.5\_turbo from OpenAI with a p\_value of 1 and a temperature of 1.5. Using LLMs to generate these individual descriptions automatically has some risks such as hallucinations or the no correspondence between the individual description and the individual. However, as manually annotating this dataset is not feasible and the interactions on it are not very complex, we have decided to use this approach as a proof of concept. In future work, more complex techniques for generating individual descriptions might be tested.

### B. Individual Motion Prior

For our proposal in Sec. 3.2 we need an individual motion prior. As mentioned in Sec. 2.1 there are many existing approaches for single-human motion generation. However, none of them are trained with the motion representation that we use in our interaction model. For this reason, we have proposed a single-human baseline based on our in2IN architecture. The differences with the proposed architecture in Sec. 3.1 is that we have removed the cross-attention modules and we have only retained the individual conditioning.

While this individual motion prior can be theoretically interchangeable with other ones, the prior selected must have been trained with the same motion representation as the interaction model and using the same training and sampling scheduler. We have trained this individual prior with the HumanML3D dataset converted to the InterHuman format. It is important to do that, because the HumanML3D dataset contains relative joint positions and velocities, while the InterHuman dataset has these values in the world frame to properly represent the global positions of the different individuals on the interaction. The rest of the implementation details are the same as the ones described in Sec. 4.3 for the in2IN model. The only difference is that we have trained the individual prior with just the L2 loss.

#### B.1. Results Individual Generation

The results of the evaluation of the individual prior can be observed in Tab. A. While our model does not beat the best models presented in this table, it obtains decent results to be used as a motion prior. While better architectures could have been used, this goes out of the scope of this paper as the only objective was to obtain a decent motion prior able to use the InterHuman motion representation.

### C. Additional Qualitative Evaluation

In addition to the examples shown in Sec. 4.5, we include additional cases that have been used on the qualitative evaluation which illustrate the observations presented in the main paper. In Fig. A and Fig. C we can observe how the interactions generated by our in2IN architecture outperform the ones generated by InterGen. Additionally, in Fig. B and Fig. D we further corroborate the improvements of the exponential schedulers for DualMDM in comparison to the others. It can also be observed the differences for the same  $\lambda$  with different examples. As stated in the main paper, future lines of work could try to propose better blending strategies without scheduler parameters.

Method	R Precision (top 3) $\uparrow$	FID $\downarrow$	MM Dist $\downarrow$	Diversity $\rightarrow$	Multimodality $\uparrow$
Real	$0.797 \pm .002$	$0.002 \pm .000$	$2.974 \pm .008$	$9.503 \pm .065$	-
JL2P	$0.486 \pm .002$	$11.02 \pm .046$	$5.296 \pm .008$	$7.676 \pm .058$	-
Text2Gesture	$0.345 \pm .002$	$7.664 \pm .030$	$6.030 \pm .008$	$6.409 \pm .071$	-
T2M	<b><math>0.740 \pm .003</math></b>	$1.067 \pm .002$	<b><math>3.340 \pm .008</math></b>	$9.188 \pm .002$	$2.090 \pm .083$
MDM	$0.707 \pm .004$	<b><math>0.489 \pm .025</math></b>	$3.631 \pm .023$	<b><math>9.449 \pm .066</math></b>	<b><math>2.873 \pm .111</math></b>
Individual Prior (Ours)	$0.6172 \pm .005$	$5.0631 \pm .150$	$4.2910 \pm .026$	$7.8289 \pm .082$	$0.4354 \pm .024$

Table A. Quantitative evaluation of our individual motion prior in comparison with other models.  $\pm$  indicates the 95% confidence interval. We highlight the **best** results



Figure A. **Interaction Description:** both they lift their right legs to kick one another. The X-axis represents time.



Figure B. **Interaction Description:** Two people greet each other by shaking hands. **Individual Description #1:** One person reaches out their hand to meet the other person's hand, shaking it in a vertical motion. **Individual Description #2:** An individual jumps. The X-axis represents time.

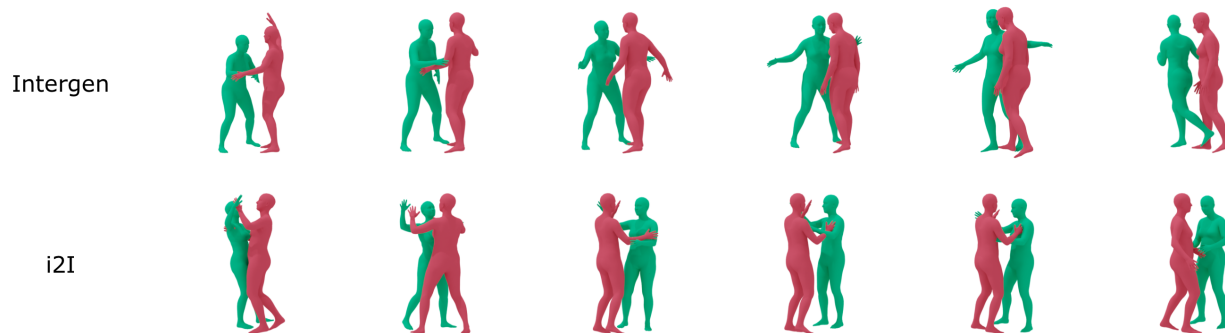


Figure C. **Interaction Description:** the two individuals are dancing ballroom together. The X-axis represents time.



Figure D. **Interaction Description:** Two people salute to each other. **Individual Description #1:** An individual bows forward. **Individual Description #2:** An individual raises their right arm and waves it. The X-axis represents time.