PINO: Person-Interaction Noise Optimization for Long-Duration and Customizable Motion Generation of Arbitrary-Sized Groups

Supplementary Material

A. Formulation of Penalty Functions

A.1. Root Position Penalty

To ensure that an individual reaches a specific location at a particular time, we define the root position penalty $\mathcal{L}_{\text{root}}$ as follows:

$$\mathcal{L}_{\text{root}} = \sum_{n \in \mathcal{N}} \max\left(0, \left\|\mathbf{p}^{p}_{\text{root}}(n) - \mathbf{p}_{\text{target}}(n)\right\|^2 - \delta\right), \quad (11)$$

where $\mathbf{p}^{p}_{\text{root}}(n)$ represents the root position of individual p at frame n, and $\mathbf{p}_{\text{target}}(n)$ is the target position at the same frame. The hyperparameter δ acts as a threshold distance, below which no penalty is applied.

This modification relaxes the loss function by penalizing deviations from the target position only when the distance exceeds the threshold δ. This approach prevents overly rigid constraints on positions that are close enough to the target, enabling more natural and flexible motion generation.

A.2. Movement Region Penalty

To restrict movement within a defined area, we introduce the movement region penalty $\mathcal{L}_{\text{region}}$ for a chosen individual p:

$$\mathcal{L}_{\text{region}} = \frac{1}{N}\frac{1}{J}\sum_{n}\sum_{j}\phi\left(\mathbf{p}^{p}_{j}(n)\right), \quad (12)$$

where N is the total number of frames, J is the total number of joints, and $\phi(\mathbf{p})$ penalizes positions outside the desired region. Here, $\mathbf{p}^{p}_{j}(n)$ represents the position of joint j of individual p at frame n.

The penalty function $\phi(\mathbf{p})$ can be flexibly defined according to the specific requirements of the task. For example, to restrict movement within a rectangular cuboid region defined by lower bounds $\mathbf{l} = (l_x, l_y, l_z)$ and upper bounds $\mathbf{u} = (u_x, u_y, u_z)$, $\phi(\mathbf{p})$ can be formulated as:

$$\phi(\mathbf{p}) = \sum_{k \in \{x,y,z\}} \max(0, l_k - p_k) + \max(0, p_k - u_k), \quad (13)$$

where p_k is the k-th coordinate of the position $\mathbf{p}$, and l_k, u_k are the lower and upper bounds of the region along the k-axis, respectively.

A.3. Orientation Penalty

To control facing directions at specific frames while allowing for small deviations, we define the orientation penalty $\mathcal{L}_{\text{orient}}$ with a threshold δ as follows:

$$\mathcal{L}_{\text{orient}} = \sum_{n \in \mathcal{N}} \max\left(0, 1 - \mathbf{d}^{p}(n) \cdot \mathbf{d}_{\text{target}}(n) - \delta\right), \quad (14)$$

where $\mathbf{d}^{p}(n)$ is the normalized facing direction of individual p at frame n, and $\mathbf{d}_{\text{target}}(n)$ is the desired direction. The hyperparameter δ acts as a threshold, below which no penalty is applied.

This modification relaxes the loss function by penalizing deviations from the desired direction only when the cosine similarity between $\mathbf{d}^{p}(n)$ and $\mathbf{d}_{\text{target}}(n)$ falls below $1 - \delta$. By avoiding penalties for minor deviations, this approach enables more natural and flexible motion generation while maintaining alignment with the desired direction.

A.4. Relative Position Penalty

To ensure that the root positions of two individuals remain within a desired distance range, we define the relative position penalty $\mathcal{L}_{\text{relative}}$ as:

$$\mathcal{L}_{\text{relative}} = \sum_{n \in \mathcal{N}} \Big[\max\left(0, d_{\text{min}} - \left\|\mathbf{p}^{1}_{\text{root}}(n) - \mathbf{p}^{2}_{\text{root}}(n)\right\|\right) + \max\left(0, \left\|\mathbf{p}^{1}_{\text{root}}(n) - \mathbf{p}^{2}_{\text{root}}(n)\right\| - d_{\text{max}}\right)\Big], \quad (15)$$

where $\mathbf{p}^{1}_{\text{root}}(n)$ and $\mathbf{p}^{2}_{\text{root}}(n)$ represent the root positions of individuals 1 and 2 at frame n, and d_{min} and d_{max} are the minimum and maximum allowable distances, respectively.

This penalty constrains the distance between the root positions of two individuals to remain within the desired range $[d_{\text{min}}, d_{\text{max}}]$. By adjusting this range, the strictness of the constraint can be flexibly controlled to allow for tighter or more relaxed interactions. The use of $\max$ ensures that penalties are applied only when the distance violates these bounds, enabling natural and adaptable motion generation without imposing overly rigid constraints.

B. Experimental Details

B.1. Optimization Details

In all experiments, we follow a consistent set of configurations to ensure fair and reproducible results. We base our optimization framework on ProgMoGen [20], originally designed for single-person motion optimization. Specifically, we use the Adam optimizer due to its robustness and efficiency in handling motion optimization tasks. The number

Task	Overlap	Root Position	Acceleration
Two-person (Table 2)	✓		
Multi-person (Table 3, 4, C)	✓		
Two-person Extension (Table 6, 7)	✓		✓
Specifying the position (Table B)	✓	✓	

Table A. Penalties applied in each task.

of optimization steps is set to 100 by default; however, we employ an early stopping criterion that halts optimization once the total loss falls below 10^{-6}. This efficiency improvement allows simpler tasks, such as overlap avoidance, to converge more quickly.

For penalty terms, we use fixed hyperparameters throughout all experiments. The overlap penalty threshold, δ_{overlap}, is set to 30 cm to ensure sufficient spatial separation. For orientation constraints, we set the threshold $\delta_{\text{orientation}}$ to 0.2. These settings were chosen to balance optimization precision and computational efficiency. The penalty terms introduced in each experiment is shown in Table A.

B.2. Evaluation Metrics

For the *Overlap* metric, we define an overlap occurrence as any frame in the motion sequence where the root positions of any pair of individuals are closer than 25 cm. We compute the proportion of generated interactions exhibiting overlaps across all samples, ensuring that even minor instances of unnatural proximity are accounted for. This metric prioritizes generating motions with clear spatial separation between individuals.

B.3. Tasks

B.3.1. Multi-Person Interaction Generation

The multi-person interaction generation task evaluates the framework's ability to generate plausible interactions between three or more individuals. Initially, two-person interactions are generated based on the text prompt *"the other person approaches one by walking."* From these interactions, we create eight distinct scenarios. For each scenario, we generate the interaction of a third person 12 times using the same text prompt. During the third person's optimization, we only apply the overlap penalty to guide motion generation, avoiding spatial collisions. This process results in a total of 96 generated interactions, which are quantitatively evaluated using the overlap metric to measure the effectiveness of the penalty as shown in Table 3.

To assess the semantic correctness of the generated interactions, we decompose multi-person interactions into two-person interactions and evaluate them separately. Specifically, we measure FID, overlap, foot skating, and maximum acceleration for each generated pair, where additional individuals are incrementally generated using the first person as a pivot. Since datasets involving more than three individuals are scarce, we use the same dataset as our two-person interaction evaluation. Specifically, we randomly select 300 samples from the InterGen test set to generate motions and compare them with the ground truth data, applying both quantitative metrics and qualitative analysis. This analysis allows us to quantify how well our method preserves interaction realism while avoiding spatial artifacts. The results are presented in Table 4, demonstrating that our approach effectively reduces overlap while maintaining natural motion characteristics.

Figure 3 illustrates three examples of multi-person interaction generation. In the first example, a two-person dance interaction is generated using the prompt "The two people are dancing together". In the second example, a three-person dance interaction is generated. The initial two-person motion is created using the prompt "Two persons danced at the party". A third person's motion is then generated using the person-to-person approach, optimized with the same prompt and guided by an overlap penalty to avoid spatial collisions. In the third example, a multi-person conversation scene is generated. First, the motions of two people are created with the prompt "They are talking and using hand gestures." This two-person interaction is optimized with both an overlap penalty and a root-position penalty to keep each character in its intended area and prevent collisions. One of these two characters is then designated as a hub, and the motions of three additional people are generated sequentially via the person-to-person approach using the same prompt. Each new character is likewise optimized with the overlap penalty and root-position penalty, resulting in a coherent five-person conversation where everyone stands in their assigned locations, gestures naturally, and avoids intersecting with one another.

B.3.2. Motion Extension

For the motion extension task in Table 6, we first generate two-person interactions lasting 8 seconds based on the text prompt *"Two people are dancing together."* From each sequence, the final 50 frames are fixed, and an additional 190 frames are generated as a continuation. To ensure smooth transitions and natural motions, we apply two penalties during optimization. The first penalty, the overlap penalty, prevents collisions between individuals during the extended motion. The second penalty, the acceleration penalty, is applied over the first 25 frames of the extended sequence starting from the transition point to suppress abrupt motions. The weights of all penalties and the settings of all hyperparameters are consistent with those used in other experiments. As elements such as abrupt sliding are difficult to observe solely from still images, qualitative results for this task are included in the supplementary video.

For further evaluation, we compare our motion extension results to ground truth sequences by cropping the generated motions to match the ground truth length. This allows for a fair comparison using quantitative metrics, including FID,

Method	C.Err↓	Overlap↓	PenVol.↓	Foot Skate↓	Max Acc.↓	FID↓	Div.→	R-Prec.↑ (Top3)
InterGen	4.029	0.119	3112.72	0.124	**0.034**	13.278	7.793	0.674
InterGen (inpainting)	**0.0**	0.130	1546.06	0.120	0.191	**12.903**	**7.899**	**0.705**
PINO-InterGen	0.071	**0.050**	**678.32**	**0.119**	0.039	14.575	7.788	0.587

Table B. Evaluation of two-person interaction generation by specifying the positions at the initial and final frames.

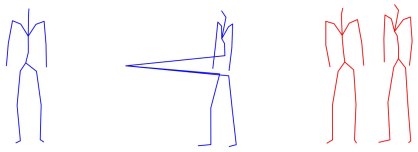

Figure A. Initial frames of interaction sequences generated by the inpainting-based method (left) and our optimization-based method (right) in an interaction generation task where root positions are specified at the initial and final frames. The inpainting-based method fails to coordinate the masked, specified root joint with other generated joints, resulting in unnatural skeleton outputs, whereas our method avoids such failures, maintaining consistency with the positional constraints depicted in the initial frames.

overlap, foot skating, and maximum acceleration. As shown in Table 7, our method consistently yields improved FID and reduced motion artifacts, demonstrating the capability of our approach to generate smoother and more natural extended motions.

C. Additional Analysis

C.1. Comparison with Inpainting-Based Methods

To further analyze the effectiveness of our approach, we compare it against an inpainting-based method using ground truth (GT) motion data. For the inpainting method, we extract the root positions of the first and final frames from the GT motion data. If the GT sequence exceeds 10 seconds, a random 10-second segment is selected for this process. These root positions are used as constraints, with the corresponding frames masked during generation to guide the motion synthesis.

As shown in Table B, the inpainting-based method achieves better scores on semantic metrics like FID. However, it exhibits significantly higher values in quantitative metrics such as Max Acc., indicating issues with motion realism. Specifically, as illustrated in Figure A, the inpainting approach often forces only the root positions of the start and end frames to match the specified constraints, while the relative positions of other joints are misaligned. This results in unnatural motion artifacts. In contrast, our method generates plausible motions that adhere to the positional constraints while maintaining realistic joint configurations throughout the sequence. These results highlight the robustness of our approach in preserving the natural structure of motions while fulfilling the specified constraints.

Method	Pair	FID↓	Overlap↓	Foot Skate↓	Max Acc.↓
in2IN	(1,2)	12.283	0.130	0.148	**0.041**
	(1,3)	13.128	0.642	0.149	**0.044**
	(1,4)	13.171	0.838	**0.148**	**0.046**
	(1,5)	**13.161**	0.920	**0.148**	**0.047**
PINO-in2IN	(1,2)	**12.183**	**0.000**	**0.146**	**0.041**
	(1,3)	**12.811**	**0.020**	0.150	0.045
	(1,4)	**12.967**	**0.058**	0.152	0.049
	(1,5)	13.347	**0.103**	0.153	0.051

Table C. Evaluation of multi-person interaction generation.

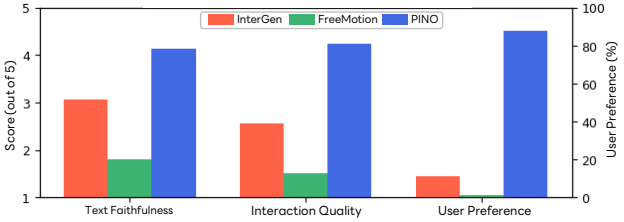

Figure B. User Study.

C.2. Other baseline model

We conducted preliminary comparisons with in2IN [26], a recent framework for multi-person motion generation. As shown in Table C, the evaluation follows a similar setting to that of Table 4 in the main paper, where additional individuals are generated incrementally using the first character as a pivot. We observe that in2IN exhibits a comparable trend to InterGen, with increasing overlap as more individuals are introduced. In contrast, our proposed PINO-in2IN significantly reduces such overlap while maintaining competitive or improved performance across FID, Foot Skate, and Max Acc., demonstrating its effectiveness in producing coherent multi-person interactions.

C.3. User Study

We conducted a user study with 35 participants to evaluate the quality of motions generated by InterGen [19], FreeMotion and ours across five interaction types: "dance (2/3 person.)", "talk" (Fig. 3), and "jump," "rock-paper-scissors" (Fig. 4). For each interaction, participants were shown three animations generated from the same text prompt, presented

in a randomized order to avoid positional bias. The participants were asked to rate each animation on a 5-point Likert scale for Text Faithfulness and Interaction Quality, which includes physical plausibility factors such as foot contact and absence of overlap. After rating each animation, participants were additionally asked to select the motion they preferred overall for each interaction.

In total, each participant evaluated all five interactions, resulting in 175 evaluation instances. To improve clarity, the animations included ground planes and shadows. As shown in Fig. B, our method achieved consistently higher scores in both rating criteria and was most frequently selected in the preference vote.

C.4. Inference speed

PINO is optimization based and slower than feed-forward methods. On an NVIDIA H100 GPU, inference takes approximately 1 minute per person for a 10-second motion (300 frames at 30 FPS) with the overlap penalty and early stopping, and up to 10 minutes with additional penalties. While not real-time, it is effective for high-quality offline content of arbitrary-sized group motion, whose training data is scarce. Accelerating optimization is future work.

C.5. Additional Visualization

In order to provide better understanding of the effects of our proposal, we include a supplementary video containing the generated results. We highly recommend viewing the video, as the results further emphasize the ability of our method in producing realistic multi-person interactions.