

Unified Number-Free Text-to-Motion Generation Via Flow Matching

Supplementary Material

This appendix provides motion representation details (Sec. A), the pyramid continuity derivation (Sec. B), additional training and evaluation specifications (Sec. C), user study details (Sec. D), the hyperparameter ablation study (Sec. E), the group prompt synthesis (Sec. F) and additional HumanML3D results (Sec. G).

Video. In the supplementary video, we show more cases for text-to-interaction generation. We recommend viewing the supplementary video to observe the dynamic motion quality.

A. Motion Representation Details

A.1. Canonical Representation

Canonical Representation (CR) is highly expressive and compatible with neural network architectures, and commonly adopted in recent single-person text-to-motion methods [4, 12]. The motion state x_c for a single person at frame i is defined as:

$$x_c^i = [r^a, r^x, r^z, r^y, \mathbf{j}_l^p, \mathbf{j}_l^v, \mathbf{j}^r, \mathbf{c}^f] \quad (1)$$

where $r^a \in \mathbb{R}$ is the root angular velocity along the Y-axis, $(r^x, r^z) \in \mathbb{R}^2$ are root linear velocities on the XZ-plane, and $r^y \in \mathbb{R}$ is root height. Local joint positions, velocities, and rotations are given by $\mathbf{j}_l^p \in \mathbb{R}^{3(N_j-1)}$, $\mathbf{j}_l^v \in \mathbb{R}^{3N_j}$, and $\mathbf{j}^r \in \mathbb{R}^{6(N_j-1)}$, with $N_j = 22$ denoting the number of joints. Binary foot-ground contact features $\mathbf{c}^f \in \mathbb{R}^4$ are derived by thresholding heel and toe joint velocities [3]. The resulting input dimension is 263 per person.

A.2. Non-Canonical Representation

Non-Canonical Representation (NCR) is suitable for multi-person scenarios [8, 10] because it retains explicit global spatial information by partially canonicalizing joint states to the root frame. The motion state x_{nc} for an interaction sequence at frame i is represented by:

$$x_{nc}^i = [\mathbf{j}_g^p, \mathbf{j}_g^v, \mathbf{j}^r, \mathbf{c}^f], \quad (2)$$

where the motion state x_{nc}^i is defined by global joint positions $\mathbf{j}_g^p \in \mathbb{R}^{3N_j}$ and velocities $\mathbf{j}_g^v \in \mathbb{R}^{3N_j}$ (in the world frame), the 6D representation of local rotations $\mathbf{j}^r \in \mathbb{R}^{6(N_j-1)}$ (in the root frame), and binary foot-ground contact features $\mathbf{c}^f \in \mathbb{R}^4$. $N_j = 22$ denotes the number of joints. The input dimension is 262 per person, totaling 524 for a two-person interaction.

A.3. Unified Representation

While HumanML3D (263) and InterHuman (262) representations are similar in dimensionality for a single person, they differ fundamentally in root frame handling. In CR, global trajectories are implicitly encoded in canonical root features $[r^a, r^x, r^z, r^y]$. Recovering world-frame information requires integrating noisy local velocities, leading to accumulated drift. This error results in unbounded exponential deviation over long sequences [13].

To enable robust multi-person generation with CR-based datasets (e.g., HumanML3D), we convert data to NCR during training to incorporate global spatial information. Conversely, generated NCR outputs are reverted to CR features during inference to maintain compatibility with standard HumanML3D evaluation protocols.

Canonical-to-Non-Canonical Conversion. We extract the global root state $(\mathbf{q}_t, \mathbf{p}_t)$ from the canonical features $[r^a, r^x, r^z, r^y]$ during preprocessing. The global Y-axis rotation angle θ_t is obtained by accumulating the angular velocity r^a . Concurrently, the global root position \mathbf{p}_t is derived by transforming the local linear velocities onto the world frame using the current orientation, followed by temporal integration. For any time step $t \in \{1, \dots, N\}$, the recovery is defined as:

$$\begin{aligned} \theta_t &= \sum_{k=0}^{t-1} r_k^a, \\ \mathbf{q}_t &= [\cos(\theta_t/2), 0, \sin(\theta_t/2), 0]^\top, \\ \mathbf{p}_t &= \sum_{k=1}^t \text{Rot}(\mathbf{q}_k, [r_{k-1}^x, 0, r_{k-1}^z]^\top) + [0, r_t^y, 0]^\top, \end{aligned} \quad (3)$$

where $\text{Rot}(\mathbf{q}, \mathbf{v}) \triangleq \mathbf{q} \hat{\mathbf{v}} \mathbf{q}^*$ denotes the spatial rotation of vector \mathbf{v} by quaternion \mathbf{q} via conjugation. The final output is the global root state $S_t^{\text{root}} = (\mathbf{q}_t, \mathbf{p}_t)$, where \mathbf{q}_t explicitly defines the root rotation and \mathbf{p}_t represents the global root position.

Non-Canonical-to-Canonical Conversion. Conversely, extracting the canonical root features $\mathcal{F}_{\text{root}} = \{r^a, r^x, r^z, r^y\}$ from a raw sequence of global joint positions $\mathbf{J} \in \mathbb{R}^{N \times N_j \times 3}$ involves inverse transformation. First, the global root orientation \mathbf{q}_t is estimated via Inverse Kinematics (or directly extracted from \mathbf{j}^r) at each frame, aligning the skeleton to the Z+ direction. Subsequently, the relative velocities and height are computed by projecting the global motion derivatives

onto the local root frame:

$$\begin{aligned} r_t^y &= (\mathbf{p}_t)_y, \\ r_t^a &= \psi_y(\mathbf{q}_{t+1}\mathbf{q}_t^{-1}), \\ [r_t^x, 0, r_t^z]^\top &= \text{Rot}(\mathbf{q}_{t+1}^{-1}, \mathbf{p}_{t+1} - \mathbf{p}_t), \end{aligned} \quad (4)$$

where $(\cdot)_y$ extracts the Y-axis component, $\psi_y(\cdot)$ computes the Y-axis Euler angle from the relative quaternion difference, and $\text{Rot}(\mathbf{q}, \mathbf{v}) \triangleq \mathbf{q}\hat{\mathbf{v}}\mathbf{q}^*$ denotes the spatial rotation of vector \mathbf{v} by quaternion \mathbf{q} via conjugation.¹

B. Pyramid Continuity Derivation

This section provides a detailed derivation of Eq. 11, which ensures continuity at the jump points of the temporal pyramid. The objective is to enforce that the end points of current temporal windows share an identical Gaussian distribution with the start points of subsequent temporal windows.

For the k -th time window $[s_k, e_k]$, we jointly compute the end points $(\hat{z}_{s_k}, \hat{z}_{e_k})$ with noise $\epsilon \sim \mathcal{N}(\mathbf{0}, I)$ and data point z_1 as:

$$\text{Start Point: } \hat{z}_{s_k} = s_k \text{Up}(\text{Down}(z_1, 2^k)) + (1 - s_k)\epsilon, \quad (5)$$

$$\text{End Point: } \hat{z}_{e_k} = e_k \text{Down}(z_1, 2^{k-1}) + (1 - e_k)\epsilon, \quad (6)$$

where $k \in [K, 1]$, and $\text{Up}(\cdot)$ and $\text{Down}(\cdot)$ are standard, non-invertible resampling functions.

During inference, the low-resolution latent motion from stage k is upsampled using the end point \hat{z}_{e_k} , yielding:

$$\text{Up}(\hat{z}_{e_k}) \sim \mathcal{N}(e_k \text{Up}(\text{Down}(z_1, 2^{k-1})), (1 - e_k)^2 \Sigma). \quad (7)$$

Moreover, the start point for the subsequent stage ($k-1$) is:

$$\hat{z}_{s_{k-1}} \sim \mathcal{N}(s_{k-1} \text{Up}(\text{Down}(z_1, 2^{k-1})), (1 - s_{k-1})^2 I), \quad (8)$$

Continuity requires that the distributions at each jump point are identical: $\hat{z}_{s_{k-1}} \stackrel{d}{=} \text{Up}(\hat{z}_{e_k})$. To satisfy it, we apply a linear rescaling and renoising scheme defined as:

$$\hat{z}_{s_{k-1}} = A \text{Up}(\hat{z}_{e_k}) + \alpha n', \quad \text{s.t. } n' \sim \mathcal{N}(\mathbf{0}, \Sigma'), \quad (9)$$

where A is the linear weight, α is the noise weight, and Σ' is a blockwise diagonal covariance matrix (e.g., 4×4 blocks).

B.1. Matching the Means

We set linear coefficient $A = \frac{s_{k-1}}{e_k}$ to match the distribution means:

$$\hat{z}_{s_{k-1}} = \frac{s_{k-1}}{e_k} \text{Up}(\hat{z}_{e_k}) + \alpha n', \quad \text{s.t. } n' \sim \mathcal{N}(\mathbf{0}, \Sigma'), \quad (10)$$

¹Extracting velocities via forward differences in Eq. 4 inherently reduces the sequence length by one. To maintain dimensional consistency during inference, we pad the missing final frame by replicating the velocities from the $(N-1)_{th}$ frame.

B.2. Matching the Variances

Based on $A = \frac{s_{k-1}}{e_k}$, matching the covariance matrices yields the following constraint:

$$\frac{s_{k-1}^2}{e_k^2} (1 - e_k)^2 \Sigma + \alpha^2 \Sigma' = (1 - s_{k-1})^2 I. \quad (11)$$

Here, we apply nearest-neighbor upsampling to analyze the covariance matrix Σ . In this case, Σ has a blockwise structure with non-zero elements only in the 4×4 blocks along the diagonal (corresponding to those upsampled from the same pixel). Consequently, the corrective noise covariance Σ' must also adopt a similar blockwise structure:

$$\Sigma_{\text{block}} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \Rightarrow \Sigma'_{\text{block}} = \begin{pmatrix} 1 & \gamma & \gamma & \gamma \\ \gamma & 1 & \gamma & \gamma \\ \gamma & \gamma & 1 & \gamma \\ \gamma & \gamma & \gamma & 1 \end{pmatrix}, \quad (12)$$

where $\gamma \in [-1/3, 0]$ is a decorrelation coefficient. The lower bound $\gamma = -1/3$ ensures Σ'_{block} is positive semidefinite.

We further rewrite Eqs. (11) and (12) by considering the equality of their diagonal elements:

$$\frac{s_{k-1}^2}{e_k^2} (1 - e_k)^2 + \alpha^2 = (1 - s_{k-1})^2, \quad (13)$$

and non-diagonal elements:

$$\frac{s_{k-1}^2}{e_k^2} (1 - e_k)^2 + \alpha^2 \gamma = 0. \quad (14)$$

Taking into account the timestep constraints $0 < s_{k-1}$, $e_k < 1$, they can be solved directly:

$$e_k = \frac{s_{k-1}\sqrt{1-\gamma}}{(1-s_{k-1})\sqrt{-\gamma} + s_{k-1}\sqrt{1-\gamma}}, \quad \alpha = \frac{1-s_{k-1}}{\sqrt{1-\gamma}}. \quad (15)$$

Intuitively, it is desirable to maximally preserve the signal at each jump point, which corresponds to minimizing the noise weight α . According to Eq. (15), this is equivalent to minimizing γ . Substituting its minimum value $\gamma = -1/3$ into Eq. (15) yields:

$$\alpha = \frac{\sqrt{3}(1-s_{k-1})}{2}, \quad e_k = \frac{2s_{k-1}}{1+s_{k-1}}. \quad (16)$$

It is worth noting that $e_k > s_{k-1}$, indicating that the timestep is rolled back a bit when adding the corrective noise at each jump point. This yields the final renoising rule (substituting these values into Eq. 9):

$$\hat{z}_{s_{k-1}} = \frac{1+s_{k-1}}{2} \text{Up}(\hat{z}_{e_k}) + \frac{\sqrt{3}(1-s_{k-1})}{2} \mathbf{n}'. \quad (17)$$

C. Training & Evaluation Details

C.1. Evaluation Metrics

Frechet Inception Distance (FID) The FID [5] measures the distribution distance between the generated and real interaction features.

$$\text{FID} = \|\mu_{\text{gt}} - \mu_{\text{pred}}\|^2 - \text{Tr}(\Sigma_{\text{gt}} + \Sigma_{\text{pred}} - 2(\Sigma_{\text{gt}}\Sigma_{\text{pred}})^{\frac{1}{2}})$$

where μ_{gt} and μ_{pred} are the mean ground-truth and generated interaction features, and Σ represents the covariance matrix.

Multimodal Distance (MM-Dist) This metric calculates the average Euclidean distance between each text feature and the generated interaction feature.

$$\text{MM-Dist} = \frac{1}{N} \sum_{i=1}^N \|f_{t,i} - f_{m,i}\|$$

where $f_{t,i}$ and $f_{m,i}$ are the features of the i th text-interaction pair.

Diversity All generated interactions are randomly sampled to calculate the average Euclidean distance between two subsets.

$$\text{Diversity} = \frac{1}{X_d} \sum_{i=1}^{X_d} \|x_i - x'_i\|$$

Multimodality (MModality) This metric assesses the variability given multiple text descriptions by calculating the average pairwise Euclidean distance between motion features.

$$\text{MModality} = \frac{1}{J_m \times X_m} \sum_{j=1}^{J_m} \sum_{i=1}^{X_m} \|x_{j,i} - x'_{j,i}\|$$

where $x_{j,i}$ and $x'_{j,i}$ are the features of the j th pair of the i th text description.

C.2. Geometric Loss Function

To enforce physical plausibility and mitigate artifacts, we adapt the geometric loss from [12]. Specifically, we utilize Bone Length (BL) and Foot Contact (FC) regularization:

$$\mathcal{L}_{BL} = \|B(\hat{x}_a) - B(x_a)\|_2^2 + \|B(\hat{x}_b) - B(x_b)\|_2^2, \quad (18)$$

$$\mathcal{L}_{FC} = \frac{1}{N-1} \sum_{i=1}^{N-1} \left\| \left(FK(\hat{x}_{foot}^{i+1}) - FK(\hat{x}_{foot}^i) \right) \cdot f_i \right\|_2^2 \quad (19)$$

$$\mathcal{L}_{\text{geometric}} = \lambda_{BL} \mathcal{L}_{BL} + \lambda_{FC} \mathcal{L}_{FC} \quad (20)$$

where B represents the bone lengths in a predefined human body kinematic tree derived from the global joint positions, and FK denotes the forward kinematic function converting joint rotations into joint positions. Bone length loss \mathcal{L}_{BL} constrains the global joint positions of each person to

satisfy skeleton consistency, which implicitly encodes the kinematic structure. $f_i \in \{0, 1\}^J$ is the binary foot contact mask for each frame i , indicating whether they touch the ground; it mitigates the foot-sliding effect by nullifying velocities when touching the ground.

C.3. Implementation Details

For VAE pretraining, motion transformer encoders and decoders all consist of 11 layers and 8 heads with skip connections by default. The P-Flow and S-Flow transformers share this architecture but employ 13 layers. We employ a frozen CLIP-ViT-L/14 model as the text encoder, yielding text embedding $c \in R^{768}$, and adopt the classifier-free guidance [6] where the 10% random CLIP embeddings are set to zero during training and the guidance coefficient is set to 2.5 during sampling. The hyperparameters used in ILD are: $\lambda_{BL} = 10$, $\lambda_{FC} = 30$, $\lambda_{\text{Geometric}} = 1e-5$, $\lambda_{\text{VAE-Recon}} = 1$, $\lambda_{KL} = 1$, $\lambda_{\text{S-Flow-Recon}} = 0.25$. Specifically, for the P-Flow training, we apply two-stage temporal windows, where $e_2 = 0.5$, $e_1 = 1$, $s_2 = 0$, $s_1 \approx 0.289$.

C.4. Baseline settings.

We compare with various text-to-motion methods in two-person interactive scenarios, including single-person methods VAE-based TEMOS [9] and T2M [3], diffusion-based MDM [12], and the two-person diffusion-based methods ComMDM [11], InterGen [8], in2IN [10], TIM [14], and InterMask [7] based on masked transformer. To ensure fair comparison, the above single-person methods are trained with the same InterHuman training set and test set. To extend single-person motion synthesis models to handle two-person interaction, the networks' input and output dimensions are modified to accommodate the non-canonical representation of the InterHuman dataset. Specifically, we report the results of TIM with a Transformer backbone.

D. User Study

We conduct a user study for the number-free text-to-motion task, specific to unseen crowd scenarios with more than 2 agents. We asked 15 users to choose between UMF and state-of-the-art work FreeMotion [2] in a side-by-side view, with both samples generated from the same text prompt. We repeated this process with 21 unseen group prompts (see Fig. 8) per model and 10 repetitions per prompt. Fig. 5 shows that UMF was preferred over the compared models in the majority of cases. This user study was designed to measure the overall rating. After the user gives the rating, a second question asks about the factors that influenced their decisions, such as Text Match (Did the video correctly follow the text description?), Physical Realism (Did the motion look physically possible and natural?) and Interaction Quality (Did the people interact in an appropriate way,

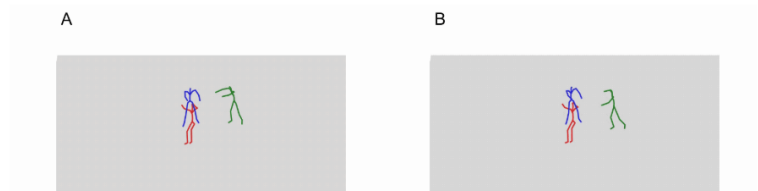
This is Example

Example Question (not counted in total)

Text Description:

Three performers spin around each other in circles.

Video Comparison



Step 1: The Overall Rating (Your Intuitions)

Single Choice: Select one option that best describes your preference

Which animation looks more human-like? *

☐ A 🍌🍌 ☐ A 🍌 ☐ Equal ☐ B 🍌 ☐ B 🍌🍌

Step 2: The Contributing Factors (Your Reasons)

Multiple Choice: Select all factors that influenced your decision. You can skip this part if you chose "Equal" in Step 1.

Which factors contributed most to your response? (Please tick all that apply) *

☐ Text Match: One motion was a better match for the text prompt (e.g. correctly showed 'clapping' while the other didn't).

☐ Motion Realism: The motion itself looked more realistic and natural (e.g. smoother movement, better balance, no 'ice-skating' or limbs passing through bodies).

☐ Interaction Quality: The interaction between people was more appropriate (e.g. taking into account others, not colliding with others).

Other

Next Question

Figure 5. An example question for our number-free text-to-motion user study, using the Streamlit platform.

namely, coordinating their movements by taking into account others). A sample question from this study is presented in Fig.5.

E. Hyperparameter Ablation Study

E.1. Classifier-free Guidance Scale

We conduct an ablation study of the classifier-free guidance scale to balance motion fidelity and text consistency, as shown in Fig. 6. We observe that as the scale increases from 1.0, the R-Top3 improves rapidly and stabilizes around a scale of 2.5 to 4.0. Meanwhile, the FID metric reaches its minimum at a scale of 2.5 and begins to degrade as the scale increases further. Therefore, we adopt a guidance scale of 2.5 for our evaluations, which achieves the optimal trade-off between FID and R-Precision.

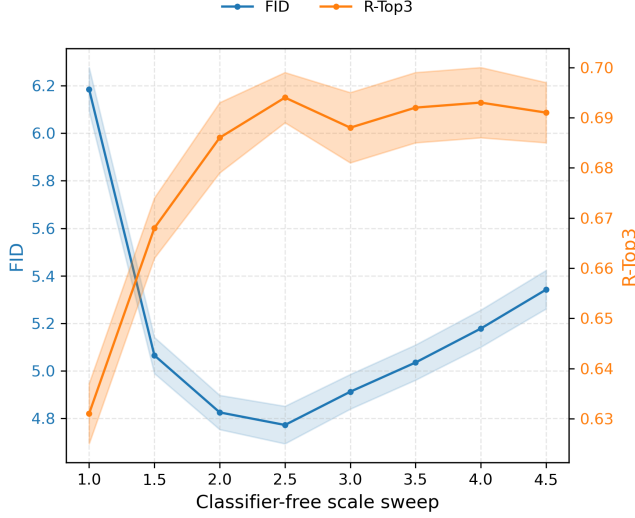


Figure 6. Guidance-scale sweep for the InterHuman dataset. FID (lower is better) and R-precision Top3 (higher is better) metrics as a function of the scales, highlight an optimal trade-off around $s = 2.5$.

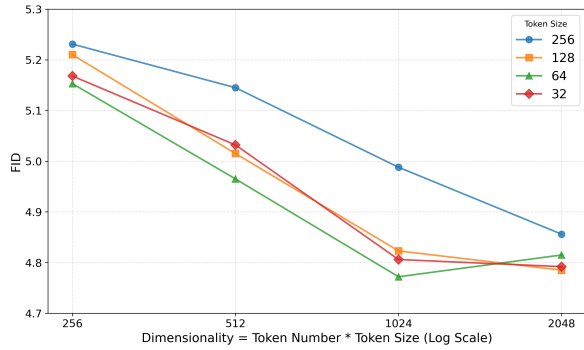


Figure 7. Latent space dimensionality sweep for the InterHuman dataset. FID metric as a function of the dimensionality highlights a performance sweet spot around dimensionality of 1024, corresponding to latent token size of 64 and latent token number of 16.

E.2. Latent Space Dimensionality

Fig. 7 illustrates the influence of token size and dimensionality on the single-agent tokenizer’s generation quality. We evaluate token sizes $\in \{32, 64, 128, 256\}$ across varying dimensionalities (log scale). Results indicate that while increasing dimensionality generally improves FID, the performance gain behaves differently across token sizes. Notably, the configuration with a token size of 64 achieves the best FID when the dimensionality reaches 1024. Larger token sizes (e.g., 256) do not exhibit better performance despite the increased dimensionality. Consequently, we select a token size of 64 and a total dimensionality of 1024 (16 tokens) for our UMF to ensure fidelity.

E.3. Physical Performance

Table 6. Physical performance comparison of FreeMotion and UMF across different agent counts N . We randomly sample 100 unseen group prompts (see Fig. 8) for each agent count and report the mean collision score and foot skating ratio.

Agent Count	Collision Score↓		Foot Skating↓	
	FreeMotion	UMF	FreeMotion	UMF
2	0.6815	0.4762	0.0712	0.0609
3	0.6923	0.4788	0.0854	0.0645
4	0.7156	0.4812	0.1189	0.0692
5	0.7341	0.4845	0.1405	0.0741

Tab. 6 demonstrates that UMF significantly mitigates the error accumulation inherent in autoregressive generation. While FreeMotion exhibits a degrading trend in physical plausibility as the number of agents increases, UMF maintains stability across all agent counts. This validates the effectiveness of S-Flow’s joint probabilistic modeling in preserving global interaction consistency.

E.4. Loss Function Influence

Table 7. Ablation study on the influence of the loss function in the single-agent tokenizer. ‘BL’ and ‘FC’ loss refers to the bone length and foot contact.

Methods	R Precision top 1 ↑	FID ↓	Diversity →
Ground Truth	0.452 ± 0.008	0.273 ± 0.007	7.948 ± 0.064
UMF w/o KL Loss	0.462 ± 0.005	4.895 ± 0.062	8.015 ± 0.028
UMF w/o Reconstruction Loss	0.125 ± 0.009	45.210 ± 1.034	4.103 ± 0.145
UMF w/o BL Loss	0.451 ± 0.004	5.124 ± 0.088	7.962 ± 0.035
UMF w/o FC Loss	0.445 ± 0.005	5.480 ± 0.095	7.910 ± 0.031
UMF w/o Geometric Loss	0.382 ± 0.006	7.655 ± 0.124	7.528 ± 0.048
UMF	0.467 ± 0.004	4.772 ± 0.079	8.039 ± 0.032

Tab. 7 presents a detailed comparison of the loss designs in the single-agent tokenizer. Overall, the reconstruction loss contributes the most to the model performance. Without the reconstruction loss, the model fails to learn valid motion representations, resulting in an FID score of 45.21 and drastically reduced R-Precision. The geometric loss also proves to be significant. Removing it leads to a noticeable decline in all metrics, highlighting the importance of physical constraints. In contrast, the KL loss and individual BL/FC losses have a relatively smaller impact compared to the reconstruction objective, but their removal still degrades performance. Ultimately, the full UMF configuration achieves the best balance across all metrics, validating our design choices.

F. Group Prompt Synthesis

Fig. 8 illustrates the prompt template employed to synthesize group scenarios via LLM [1]. We condition the model on dyadic descriptions from InterHuman, instructing it to

You are an expert in human interaction, social dynamics, and narrative logic.

****Task:****

Analyze the provided **text description** ({{overall_desc}}) which details the interaction between **two (2) people**. Based **only** on the described actions and cues of these two people, you must ****infer and generate plausible actions for the unseen, hypothetical PARTICIPANTS**** in the requested scenario (3-person, 4-person, or 5-person).

****Core Goal:****

Your descriptions of the unseen participants MUST be a **direct and logical consequence** of the actions described in the text. The known actions (e.g., gestures, gaze directions, conversation turns, or physical reactions) should function as **clues** that imply the presence and actions of the unseen people.

****Input Data:****

• **Description of 2-Person Interaction:** `{{overall_desc}}`

****Constraints & Guidance (Strictly follow):****

1. **Logical Interaction Consistency:** The inferred actions of the unseen participant(s) MUST be socially or physically logical to make the **entire** scenario coherent.
2. **Contextual Adherence:** All inferences MUST strictly align with the context provided in {{overall_desc}} (whether it is a casual conversation, a formal meeting, a dance, or a conflict).
3. **Example of the Required Reasoning Process:**
 - **IF the Text Says:** `Person A smiles and extends their right hand forward as if to shake hands, while Person B stands back and watches.`
 - **A VALID 3-Person Inference would be:** `A third person steps forward to grasp Person A's extended hand for a handshake.`
 - **Why:** The **described action** (offering a handshake) implies a **recipient** (the third person) is present.
4. **Focus on Observable Action:** Descriptions must be specific, concrete actions or distinct social cues (e.g., speaking, waving, approaching).
5. **STRICT PROHIBITION:** You MUST NOT invent irrelevant background characters. The **only** added people you may describe are **active participants** whose presence is **directly** implied by the scenario.

****Instructions:****

1. Thoroughly analyze the interaction logic implied by the {{overall_desc}}.
2. Deduce what **other participants** **must** be doing to make the described scene logical.
3. Generate the **exact** number of variations required below (3 variations for each). **Your response MUST start **directly** with the Markdown heading.**

****Required Output Format (Provide ONLY the following without showing your thinking):****

3-Person Scenarios (3 Variations)

(Each describes the plausible actions of the **one unseen third person)**

1. [Description 1...]
2. [Description 2...]
3. [Description 3...]

4-Person Scenarios (3 Variations)

(Each describes the plausible actions of the **two unseen third and fourth persons)**

1. [Description 1...]
2. [Description 2...]
3. [Description 3...]

5-Person Scenario (3 Variations)

(Describes the plausible actions of the **three unseen third, fourth, and fifth persons)**

1. [Description 1...]
2. [Description 2...]
3. [Description 3...]

Figure 8. Prompt template for inferring multi-person interaction scenarios from text descriptions.

infer the actions of unseen participants based on observable social cues and narrative logic. To facilitate robust zero-shot

evaluation, the prompt strictly enforces that added characters appear as direct causal consequences of the original in-

teraction, generating three distinct variations for 3-, 4-, and 5-person configurations.

G. Additional HumanML3D Results.

Table 8. Quantitative VAE results on the HumanML3D dataset.

	FID	RTop3	Diveristy	MPJPE	PAMPJPE	ACCL
UMF-VAE	0.0051	0.735	7.70	11.5	8.4	4.2



Figure 9. Qualitative results on HumanML3D.

References

- [1] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blisstein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 5
- [2] Ke Fan, Junshu Tang, Weijian Cao, Ran Yi, Moran Li, Jingyu Gong, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. Freemotion: A unified framework for number-free text-to-motion synthesis. In *European Conference on Computer Vision*, pages 93–109. Springer, 2025. 3
- [3] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 1, 3
- [4] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 1
- [5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 3
- [6] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3
- [7] Muhammad Gohar Javed, Chuan Guo, Li Cheng, and Xingyu Li. Intermask: 3d human interaction generation via collaborative masked modelling. *arXiv preprint arXiv:2410.10010*, 2024. 3
- [8] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Interger: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, pages 1–21, 2024. 1, 3
- [9] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. *arXiv preprint arXiv:2204.14109*, 2022. 3
- [10] Pablo Ruiz Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and Jose Garcia-Rodriguez. in2in: Leveraging individual information to generate human interactions. *arXiv preprint arXiv:2404.09988*, 2024. 1, 3
- [11] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 3
- [12] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 1, 3
- [13] Timo Von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer graphics forum*, pages 349–360. Wiley Online Library, 2017. 1
- [14] Yabiao Wang, Shuo Wang, Jiangning Zhang, Ke Fan, Jiafu Wu, Zhucun Xue, and Yong Liu. Timotion: Temporal and interactive framework for efficient human-human motion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7169–7178, 2025. 3