Towards Storytelling Animations: Joint Synthesis of Human and Camera Motions

Supplementary Material

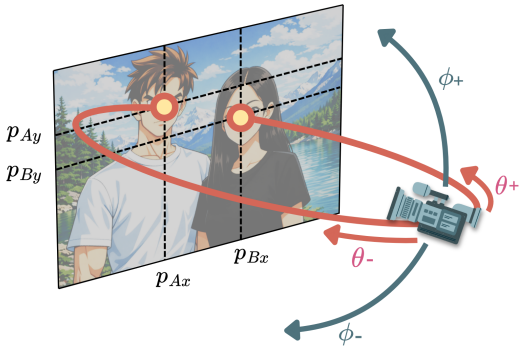

Figure 1. Toric representation for a two-shot. The camera pose is defined by two primary degrees of freedom: an azimuthal rotation ϕ (blue arrows) that enables orbiting around the inter-subject baseline AB, and an elevation coordinate θ (red arrows) for vertical positioning. This parameterization inherently ensures that both protagonists are maintained within the screen-space constraints.

1. Training Strategies for Enhancing Motion Diversity

To further increase the diversity of generated motions, we explore alternative training strategies that leverage different external datasets in combination with our proposed motion-camera dataset. Specifically, we design and evaluate three strategies:

Strategy 1. We first pre-train the backbone model using the HumanML3D dataset, which contains a wide range of single-person motion categories. The pretrained model is then fine-tuned on our motion-camera dataset to adapt to two-person interactions and camera trajectories.

Strategy 2. We jointly train the model using Inter-X, a dataset of two-person interactions, together with our motion-camera dataset. During training, when the batch samples contain camera information, all parameters are updated; when the batch only contains human interactions without cameras, we update only the human-related modules while freezing the camera branch.

Strategy 3. Building on Strategy 2, we further incorporate HumanML3D data. When the batch contains only single-person motions, we update only the human branches, keeping the camera-related parameters frozen.

Table 3 reports the quantitative comparison of different training strategies. The baseline model trained solely on our dataset already shows reasonable performance, while the use of additional datasets brings further gains. In terms of motion quality, Strategy 3 achieves the most consistent improvements. Strategy 2 also provides clear benefits by introducing richer interaction dynamics, whereas Strategy 1 underperforms due to the distributional mismatch introduced by single-person pre-training.

For camera quality, Strategy 2 obtains the lowest SeqFID and FrameFID, indicating the most stable trajectories. Strategy 3 shows modest improvements over the baseline but does not reach the level of Strategy 2, as the integration of HumanML3D reduces the effective update frequency of the camera branch. Strategy 1 remains the weakest among all settings. In terms of motion-camera coordination, both Strategy 2 and Strategy 3 improve upon the baseline, with Strategy 3 achieving the lowest CLIP Loss, reflecting stronger semantic alignment between characters and camera motion.

Overall, Strategy 3 is the most effective when the focus is on motion diversity and human–camera coordination, while Strategy 2 is better suited when stable camera trajectories are the primary concern.

2. Camera Data Representation

In this study, we represent camera states using the Toric manifold, a formulation that decouples camera positioning from intrinsic factors like focal length or aspect ratio (see Figure 1). This approach defines the camera's pose relative to the spatial layout of two primary subjects, A and B, ensuring cinematic constraints are maintained. To derive the Toric coordinates, we extract world-space coordinates for the camera x_C^i and the subjects x_A^i, x_B^i, along with their corresponding screen projections p_A^i, p_B^i. We define the following fundamental vectors:

$$v_{AB}^i = x_B^i - x_A^i, \quad v_{AC}^i = x_C^i - x_A^i, \quad v_{BC}^i = x_C^i - x_B^i \tag{1}$$

While v_{AB}^i captures the relative distance and orientation between subjects, v_{AC}^i and v_{BC}^i represent the perspective rays from the viewpoint to each target.Geometric Interpretation of α^i: The opening angle α^i signifies half of the angular spread between the subjects as perceived by the camera. By utilizing screen-space data, we reconstruct the normalized viewing rays:

$$r_A^i = \frac{\left(\frac{p_{Ax}^i}{S_x}, \frac{p_{Ay}^i}{S_y}, 1\right)}{\left|\left(\frac{p_{Ax}^i}{S_x}, \frac{p_{Ay}^i}{S_y}, 1\right)\right|}, \quad r_B^i = \frac{\left(\frac{p_{Bx}^i}{S_x}, \frac{p_{By}^i}{S_y}, 1\right)}{\left|\left(\frac{p_{Bx}^i}{S_x}, \frac{p_{By}^i}{S_y}, 1\right)\right|} \tag{2}$$

Table 1. Quantitative comparison of three training strategies on character motion generation.

Method	FID ↓	Diversity ↑	InterFID ↓	Coverage ↑	Density ↑
Strategy 1	$2.783^{\pm.098}$	$0.432^{\pm.013}$	$0.948^{\pm.128}$	$0.040^{\pm.006}$	$0.255^{\pm.047}$
Strategy 2	$1.520^{\pm.066}$	$0.652^{\pm.021}$	$0.886^{\pm.050}$	$0.428^{\pm.039}$	$0.317^{\pm.023}$
Strategy 3	$\mathbf{1.061}^{\pm.034}$	$\mathbf{0.890}^{\pm.039}$	$\mathbf{0.651}^{\pm.025}$	$\mathbf{0.868}^{\pm.058}$	$\mathbf{0.968}^{\pm.098}$
No Strategy	$1.985^{\pm.080}$	$0.575^{\pm.025}$	$0.910^{\pm.095}$	$0.310^{\pm.032}$	$0.280^{\pm.046}$

Table 2. Quantitative comparison of three training strategies on camera motion generation.

Method	SeqFID ↓	FrameFID ↓	Diversity ↑	Coverage ↑	Density ↑
Strategy 1	$0.463^{\pm.150}$	$0.161^{\pm.058}$	$0.489^{\pm.045}$	$0.549^{\pm.118}$	$1.247^{\pm.080}$
Strategy 2	$\mathbf{0.165}^{\pm.004}$	$\mathbf{0.120}^{\pm.002}$	$0.701^{\pm.032}$	$\mathbf{0.760}^{\pm.020}$	$\mathbf{2.410}^{\pm.070}$
Strategy 3	$0.174^{\pm.003}$	$0.127^{\pm.001}$	$\mathbf{0.697}^{\pm.035}$	$0.751^{\pm.016}$	$2.357^{\pm.065}$
No Strategy	$0.420^{\pm.015}$	$0.185^{\pm.006}$	$0.510^{\pm.039}$	$0.640^{\pm.025}$	$1.950^{\pm.088}$

Table 3. Quantitative comparison of three training strategies on character-camera motion coordination.

Method	CLIP Loss ↓
Strategy 1	$3.508^{\pm.339}$
Strategy 2	$2.750^{\pm.235}$
Strategy 3	$\mathbf{2.422}^{\pm.064}$
No Strategy	$2.950^{\pm.183}$

where the scaling factors S_x and S_y are determined by the horizontal field-of-view fov_x and aspect ratio ρ:

$$S_x = \frac{1}{\tan(\frac{\text{fov}_x}{2})}, \qquad S_y = S_x \cdot \rho \tag{3}$$

The resulting opening angle is calculated as:

$$\alpha^i = \arccos\big(r_A^{i\top} r_B^i\big), \qquad \alpha^i \in (0, \pi) \tag{4}$$

For computational robustness, we constrain α^i to the interval $[0.001,\ \pi - 0.001]$.Toric Coordinates (θ^i, ϕ^i): The radial coordinate θ^i is derived from an auxiliary angle β^i, which measures the separation between the subject-to-subject baseline and the camera-to-subject viewing ray:

$$\beta^i = \arccos\left(\frac{v_{AB}^{i\top} v_{AC}^i}{|v_{AB}^i|\,|v_{AC}^i|}\right) \tag{5}$$

The final radial coordinate is then defined and clamped for stability:

$$\theta^i = 2\beta^i, \qquad 0 < \theta^i < 2(\pi - \alpha^i) \tag{6}$$

Furthermore, the longitudinal coordinate ϕ^i describes the camera's azimuthal rotation around the baseline. Given the plane normal $n^i = v_{AC}^i \times v_{AB}^i$ and an orthogonal reference vector z^i, ϕ^i is expressed as:

$$\phi^i = \tfrac{\pi}{2} - \angle_{\pm}(z^i, n^i;\ v_{AB}^i) \tag{7}$$

Feature Representation: For a temporal sequence of N frames, the camera state is aggregated into a learning descriptor:

$$c^{1:N} = {p_A^i, p_B^i, \theta^i, \phi^i}_{i=1}^N \in \mathbb{R}^{6N} \tag{8}$$

This representation effectively combines screen-space composition (p_A^i, p_B^i) with Toric spatial placement (θ^i, ϕ^i), while α^i is omitted due to its redundancy.