1 Comparison of Sampling Method

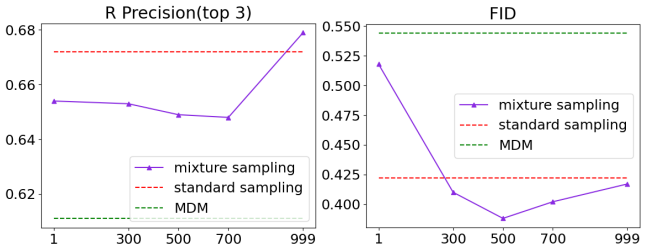

Fig. 1: Results of different sampling strategies are presented in terms of R Precision (top 3) and FID. The x-axis represents the number of steps α for 3D denoising. The red dashed line represents denoising a 3D noise only in the 3D domain, which is the standard sampling method. The green dashed line represents the results of MDM.

In Section 3.4 of the paper, we discuss how our model can denoise and reconstruct 3D movements from both 3D and 2D noise. To investigate if incorporating more 2D data can improve generation performance, we evaluate different sampling methods. For this experiment, we train our model using 50% 3D motion and all available 2D motion, and diffuse 1k steps in all methods. The parameter α indicates the time-step at which the motion representation switches from 2D to 3D. The red dashed line represents the standard sampling method.

The results, as shown in Figure 1, reveal that $\alpha = 500$ achieves the lowest FID score, indicating that our model can effectively use abundant 2D motion to enhance 3D generation performance. However, the standard sampling method scores high on R-Precision, as the generation is more precise. Additional results are provided in the supplementary material.

2 Effect of Root-decoupled Diffusion

To address the uncertainty of camera movement for in-the-wild videos, we decouple the root information $r_{3D/2D}$ and generate it based on other pose features. To achieve this, we employ a L_3-layer transformer encoder to encode the root information, which is then decoded with a L_4-layer transformer decoder conditioned on the last L_4 layer 3D/2D decoder outputs. L_3 and L_4 are set to 2 in experiments.

The comparison of generating with and without this technique is illustrated in Figure 2. Without root-decoupled diffusion, the generated motion exhibits random foot sliding while performing pulling-ups, as it learns from 2D motion data. However, since 2D motion sequences captured from videos may not accurately capture the root position, generating global movement based on body pose can lead to more precise results.

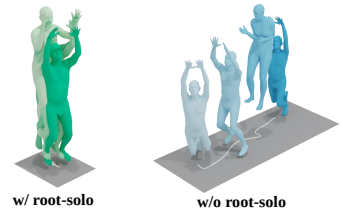

Fig. 2: Visualization of generated motions with and without the root-decoupled diffusion model.

3 Details of User Study

We ask two questions in user studies to assess the vitality and diversity of the motions. The first is "Which motion is more realistic and contains more details?" and the participant is given a generated motion of our method and the compared method to choose. The second is "Which generations are more diverse?" and the participant is given three generated motions of our method and the compared method to choose. We eventually received 135 feedbacks. Considering the response time under 1 minute is invalid, we finally collate 113. Figure 5(a) in the paper shows that most of the time, CrossDiff was preferred over the compared models.

4 Multi-view constraints for generalization issues

In our method, we implicitly introduce 3D consistency during the training process. For each 3D ground truth (GT) motion data, we randomly project it onto a 2D view. Thus, each randomly projected 2D data has a unique corresponding 3D supervision. However, for 2D sequences without 3D data, it is challenging to apply the multi-view consistency approach, as these sequences are not recorded by surrounding cameras, and our focus is on 3D sequential motion rather than static object reconstruction.

We have attempted to implement a multi-view consistency loss (MCL) by explicitly ensuring that multiple views in the same batch produce consistent 3D reconstructions. As shown in Table 1, this explicit constraint did not improve performance. This could be due to the suboptimal generation of multi-view results during the training process, making it less effective than using 3D GT for individual supervision.

5 Foot Skating Ratio evaluation results

We made comparisons with GMD [1] on foot skating ratio. However, since GMD primarily focuses on trajectory control, while our task involves basic text-to-motion, we initially did not attempt to solve the foot skating issue. We tested

Table 1: Evaluation of MCL loss.

Methods	R Precision↑	FID↓	MM Dist↓	DIV→
w MCL	0.718	0.202	3.425	9.287
w/o MCL	**0.730**	**0.162**	**3.358**	**9.577**

our model using the foot skating ratio metric and achieved comparable results to GMD. As in Table 2, Our model outperforms GMD in terms of the second-best parameters.

Table 2: Evaluation of foot skating ratio.

Methods	MDM	GMD(c=5)	GMD(c=10)	Ours
Foot Skating Ratio	0.284	0.199	**0.128**	0.178

References

1. Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: Gmd: Controllable human motion synthesis via guided diffusion models. arXiv preprint arXiv:2305.12577 (2023)