Avatars Grow Legs: Generating Smooth Human Motion from Sparse Tracking Inputs with Diffusion Model — Supplementary Material —

1. Extra Ablation Experiments

We first show extra ablation experiments of our method on AMASS [5] dataset following the protocol proposed in [2]. Then we show extra qualitative results and comparison between our method and the state-of-the-art method [2].

Additional Losses In addition to $\mathcal{L}_{dm}$, we explore three other geometric losses during the training like previous works [6,8]:

$$\mathcal{L}_{pos} = \frac{1}{N}\sum_{i=1}^{N} \parallel \text{FK}(y_0^i) - \text{FK}(\hat{x}_0^i) \parallel_2^2 \tag{1}$$

$$\mathcal{L}_{vel} = \frac{1}{N-1}\sum_{i=1}^{N-1} \parallel (\text{FK}(y_0^{i+1}) - \text{FK}(y_0^i)) - (\text{FK}(\hat{x}_0^{i+1}) - \text{FK}(\hat{x}_0^i)) \parallel_2^2 \tag{2}$$

$$\mathcal{L}_{foot} = \frac{1}{N-1}\sum_{i=1}^{N-1} \parallel (\text{FK}(y_0^i) - \text{FK}(\hat{x}_0^i)) \cdot m_i \parallel_2^2, \tag{3}$$

where $\text{FK}(\cdot)$ is the forward kinematics function which takes local human joint rotations as input and outputs these joint positions in the global coordinate space. Here $y_0^{1:N} = x_0^{1:N}$ is the original data without noise. $\mathcal{L}_{pos}$ represents the position loss the of joints, $\mathcal{L}_{vel}$ represents the velocity loss of the joints in 3D space, and $\mathcal{L}_{foot}$ is the foot contact loss, which enforces static feet when there is no feet movement. $m_i \in \{0, 1\}$ denotes the binary mask and equals to 0 when the feet joints have zero velocity.

We train our model with different combinations of extra losses, setting their weights equal to 1. As shown in Table I. In contrast to previous works [2], the extra geometric losses do not bring additional performance to our diffusion model. Our model can achieve good results when trained solely with the denoising objective function $\mathcal{L}_{dm}$ from Eq. (4).We hypothesize that the lack of improvement in AGRoL's performance with additional losses is due to the intricacies of the reverse diffusion process. This process may not synergize effectively with extra geometrical losses without appropriate adjustments.

$\mathcal{L}_{pos}$	$\mathcal{L}_{vel}$	$\mathcal{L}_{foot}$	MPJRE	MPJPE	MPJVE	Hand PE	Upper PE	Lower PE	Root PE	Jitter
			2.66	**3.71**	**18.59**	**1.31**	**1.55**	**6.84**	**3.36**	**7.26**
		✓	2.83	4.07	20.66	1.58	1.70	7.49	3.66	9.20
✓			2.81	4.06	21.85	1.75	1.73	7.43	3.72	12.16
✓	✓		2.73	3.92	20.55	1.72	1.68	7.15	3.52	10.16
✓	✓	✓	2.89	4.16	20.58	1.73	1.76	7.63	3.81	8.98

Table I. Ablation of the additional losses used during training.

Sampling Steps In Table II we ablate the number of sampling steps T during training. Surprisingly, even when training with merely 10 sampling steps, the model can achieve decent performance. Although we notice that the model converges to a worse local minimum when only a few sampling steps is used. To achieve the best results, more sampling steps is required.

# Sampling Steps	MPJRE	MPJPE	MPJVE	Hand PE	Upper PE	Lower PE	Jitter
10	2.69	3.70	19.41	1.47	1.55	6.80	7.63
100	2.65	3.62	18.74	1.33	1.52	6.66	6.71
1000 (Ours)	2.66	3.71	18.59	1.31	1.55	6.84	7.26

Table II. Ablation of the number of sampling steps during training the AGRoL model. The results become worse when the number of sampling steps is too small. More sampling steps is beneficial during training the network.

Input/Output length The proposed AGRoL model takes a sequence of sparse tracking signals as input and predicts the full body motion of the same length. In Table III we ablate the input & output length N of the AGRoL model. Our model benefits from longer input sequences, especially decreasing the mean per joint velocity error and jitter. But the performance saturates after the length of $N = 196$. In Table IV we further compare our method with AvatarPoser [2] by varying its input length. Note that with longer input sequences our model can achieve significantly lower errors on velocity-related metrics like MPJVE and jitter, while AvatarPoser still has large MPJVE and jitter even with longer input length, thus failing to fully leverage the temporal information to generate smooth motions.

Input & Output Length	MPJRE	MPJPE	MPJVE	Hand PE	Upper PE	Lower PE	Jitter
41	2.59	3.64	23.24	1.28	1.50	6.73	13.67
98	2.61	3.70	20.71	1.58	1.57	6.76	10.59
196 (Ours)	2.66	3.71	18.59	1.31	1.55	6.84	7.26
256	2.81	3.81	19.05	1.27	1.57	7.03	7.76

Table III. Ablation of the input & output length of the AGRoL model. Our model can benefit from larger input length.

Methods	Input Length	MPJRE	MPJPE	MPJVE	Hand PE	Upper PE	Lower PE	Jitter
AvatarPoser [2]	41	3.08	4.18	27.70	2.12	1.81	7.59	14.49
AGRoL	41	2.59	3.64	23.24	1.28	1.50	6.73	13.67
AvatarPoser [2]	196	3.05	4.20	28.71	1.61	1.70	7.82	16.96
AGRoL (Ours)	196	2.66	3.71	18.59	1.31	1.55	6.84	7.26

Table IV. Comparison between AGRoL and AvatarPoser [2] while varying the number of input frames. Our method can benefit from longer inputs and generate smoother motion. In contrast, AvatarPoser fails to gain consistent improvement from longer input sequences and even degrades in some metrics, including MPJVE, Lower PE, and Jitter.

Number of blocks in the MLP network In Table V we evaluate the impact of varying the number of blocks (described in Sect. 3.2) in the MLP network. The model's performance consistently improves as more blocks are added. However, the performance gains approach a plateau when more than 12 blocks are used.

#Blocks	#Params	MPJRE	MPJPE	MPJVE	Hand PE	Upper PE	Lower PE	Root PE	Jitter
2	2.08M	4.54	7.39	34.38	3.10	3.13	13.55	7.28	22.03
6	4.35M	2.89	4.12	19.51	1.38	1.69	7.62	3.72	6.29
12 (Ours)	7.48M	**2.66**	3.71	18.59	1.31	1.55	6.84	3.36	7.26
24	14.53M	2.73	**3.61**	**18.33**	**1.08**	**1.50**	**6.65**	**3.28**	**7.23**

Table V. Ablation study of the number of blocks in the proposed MLP network.

Predicting noise Our diffusion model AGRoL follows [7] and directly predicts the clean signal $\hat{x}_0^{1:N}$ in contrast to the original Denoising Diffusion Probabilistic Model (DDPM) formulation [1], where the model predicts residual noise $\epsilon_\theta(x_t, t)$ at every step. In this subsection, we further discuss the experiment presented in Table 3 of the main paper, where we implemented a version of AGRoL model ("AGRoL - pred noise") that predicts the residual noise $\epsilon_\theta(x_t, t)$. Similar to [7], we

also find it better to predict the unnoised $\hat{x}_0^{1:N}$ directly, which is demonstrated by the results in Table VI. Since our simple MLP network (see Sect. 3.2 of the main paper) can already produce reasonable estimations of the full body motion using only one forward pass (see Table 1 in the main paper), we hypothesize that the DDPM formulation of Ramesh et al. [7] allows to exploit the full capacity of the network *at every sampling step*, in contrast to the original formulation of [1].

Method	MPJRE	MPJPE	MPJVE	Hand PE	Upper PE	Lower PE	Root PE	Jitter
AGRoL - pred noise ϵ_θ	5.41	8.88	28.67	4.38	3.91	16.06	8.76	9.80
AGRoL (Ours)	**2.66**	**3.71**	**18.59**	**1.31**	**1.55**	**6.84**	**3.36**	**7.26**

Table VI. Ablating different formulations of the diffusion model: Predicting clean signal directly (Ours) *vs* predicting noise $\epsilon_\theta(x_t, t)$. The AGRoL model that learns to predict clean body motion at every diffusion step is substantially better in every metric.

2. Extra Datasets

In addition to the AMASS [5] dataset, we also evaluate the performance of our approach on AIST++ [3] dataset. AIST++ dataset contains in total 5.1 hours of dancing movements performed by professional dancers. The dataset has 10 genres of dances, including some dances containing complicated movements like breakdancing, jazz etc. We follow the train/test splits proposed in [3]. The global rotation and translation of the hands and head are calculated using the SMPL human model [4] with the provided model parameters. Compared to the AMASS dataset, which contains mostly everyday life motions, the motions in the AIST++ dataset are much more diverse and challenging. As shown in Table VII, the AGRoL achieves superior performance in all the metrics and produces smoother motions compared to the AvatarPoser and the predictive MLP model. While there is still room for improvement on such a challenging dataset, the proposed AGRoL method significantly reduces the MPJVE, Jitter and lower body positional error (Lower PE) compared to the AvatarPoser.

Method	MPJRE	MPJPE	MPJVE	Hand PE	Upper PE	Lower PE	Root PE	Jitter
AvatarPoser	4.37	9.11	97.24	4.31	3.32	17.47	8.11	65.18
MLP (Ours)	3.63	7.33	74.90	3.86	2.69	14.03	5.48	47.16
AGRoL (Ours)	**3.56**	**6.83**	**65.58**	**2.17**	**2.04**	**13.74**	**4.91**	**41.95**
GT	0	0	0	0	0	0	0	30.48

Table VII. Comparison of our approach with the competitors on AIST++ [3] dataset.

3. Extra Qualitative Results

In Figure I we show extra qualitative comparisons between our method and AvatarPoser [2]. Please refer to the videos in the supplementary material[1] for more qualitative results. As shown in the video, our method reconstructs the full body poses more accurately, it can generate smoother motions and alleviate the jittering issue compared to AvatarPoser.

Real inputs Additionally, in the supplementary material[1], we include a video showcasing AGRoL inference on real VR inputs from the Quest HMD. Our model exhibits a reasonable generalization to the real-world inputs and is capable of generating smooth and precise motions.

Failure cases We also demonstrate failure cases for the proposed AGRoL model in Figure II. We can see that our method fails when we test it on irregular poses, that were not well covered in the training set, or when the lower body pose does not have strong correlation with the upper body. For example, during the break dance motion (Fig. II, bottom row) the upper body may stay static, while legs move which makes it very challenging to predict legs accurately. Increasing the size and diversity of the training set plus incorporating extra physical or geometrical priors to prevent floor penetration could be a potential solution for the failure cases. We plan to further investigate it in future work.

[1] https://dulucas.github.io/agrol/.

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 2, 3

[2] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. *ECCV*, 2022. 1, 2, 3, 5, 6

[3] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021. 3

[4] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM TOG*, 34(6):1–16, 2015. 3

[5] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, pages 5442–5451, Oct. 2019. 1, 3, 6

[6] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *CVPR*, pages 10985–10995, 2021. 1

[7] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 3

[8] Mingyi Shi, Kfir Aberman, Andreas Aristidou, Taku Komura, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *ACM TOG*, 40(1):1–15, 2020. 1

(a) GT **(b)** AvatarPoser [2] **(c)** AGRoL (Ours)

Figure I. Qualitative comparison on test sequences from AMASS dataset. We visualize the predicted skeletons and human body meshes: **(a)** Ground truth skeletons in blue, **(b)** AvatarPoser predictions in red, **(c)** AGRoL predictions in green. It can be seen that our predicted motion is more accurate compared to the predictions of AvatarPoser, especially in the lower body. RGB axes illustrate the location and orientation of the head and hands provided as input to the models. *Please refer to our project page for more qualitative results.*

(a) GT **(b)** AvatarPoser [2] **(c)** AGRoL (Ours)

Figure II. Failure test cases. We visualize the predicted skeletons and human body meshes: **(a)** Ground truth skeletons in blue, **(b)** AvatarPoser predictions in red, **(c)** AGRoL predictions in green. The models were trained on the subset of AMASS [5] following the protocol of [2]. In the figures, the RGB axes indicate the position and orientation of the head and hands, which are used as input to the models. We can observe that our method struggles with (i) irregular poses that were not well-represented in the training set, resulting in inaccurate poses and floor penetration, and (ii) when the lower body pose is not strongly correlated with the upper body, as seen in the break dance motion in the bottom row. To address these failure cases, we could consider increasing the diversity of training motions and incorporating additional physical constraints.