MotionLab: Unified Human Motion Generation and Editing via the Motion-Condition-Motion Paradigm

Supplementary Material

1. Details of Rectified Flows

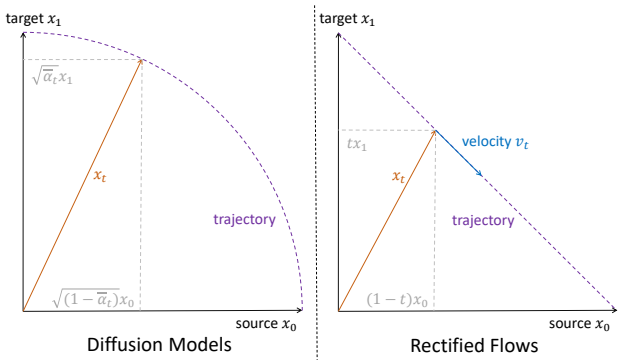

Figure 1. Demonstration of the difference between diffusion models and rectified flows. This difference lies in that the trajectory of diffusion models is based on $x_t = \sqrt{(1-\overline{\alpha_t})}x_0 + \sqrt{\overline{\alpha_t}}\epsilon$, while the trajectory of rectified flows is based on $x_t = (1-t)x_0 + tx_1$. This distinction leads to more robust learning by maintaining a constant velocity, contributing to the model's efficiency [7].

Since the trajectory x_t from p_1 to p_0 should be as straight as possible, it can be reformulated as the linear interpolation between x_0 and x_1, and the velocity field v_t can be treated as a constant, namely:

$$x_t = (1-t)x_0 + tx_1 \tag{1}$$

$$v_t = \frac{dx_t}{dt} = \frac{\partial \varphi_t(x_0, x_1, t)}{\partial t} = x_1 - x_0 \tag{2}$$

Therefore, the training objective can be reformulated as:

$$\mathcal{L}_{RF}(\theta) = \int_0^1 \mathbb{E}_{(x_0,x_1)\sim(p_0,p_1)}[||v_\theta(t,x_t) - (x_1 - x_0)||_2^2]dt \tag{3}$$

After the training of rectified flows is completed, the transfer from x_1 to x_0 can be described via the numerical integration of ODE:

$$x_{t-\frac{1}{N}} = x_t - \frac{1}{N}v_\theta(t, x_t) \tag{4}$$

where N is the discretization number of the interval [0,1].

2. MotionLab Inference

During inference, Classifier-Free Guidance (CFG) [3] is incorporated for both motion generation and motion editing to boost sampling quality and align conditions and target motion.

For all motion generation tasks, we generate target motion M_T with the guidance of arbitrary conditions C:

$$\begin{aligned} v_\theta(M_T, t, C) =& v_\theta(M_T|t, \emptyset) + \\ & \lambda_C[v_\theta(M_T|t, C) - v_\theta(M_T|t, \emptyset)] \end{aligned} \tag{5}$$

where t is the timestep and $\lambda_C > 1$ is a hyper-parameter to control the strength of the corresponding conditional guidance.

For all motion editing tasks, which aim to modify the source motion based on the condition. Hence, we generate the target motion M_T with source motion M_S first and then condition C:

$$\begin{aligned} v_\theta(M_T, t, M_S, C) =& v_\theta(M_T|t, \emptyset, \emptyset) \\ & + \lambda_S[v_\theta(M_T|t, S, \emptyset) - v_\theta(M_T|t, \emptyset, \emptyset)] \\ & + \lambda_C[v_\theta(M_T|t, S, C) - v_\theta(M_T|t, S, \emptyset)] \end{aligned} \tag{6}$$

where $\lambda_S > 1$ is a hyper-parameter to control the strength of source motion guidance.

3. Memory Usage and Time Cost

The maximum memory usage during training is 23 GB for each GPU. The memory usage and the time spent during inference are summarized in the following Table 1.

Metric	text gen	traj. gen	text edit	traj. edit	in-between	style transfer
memory usage (GB)	4.16	4.31	5.83	6.81	4.32	5.74
time spend (AITS)	0.068	0.134	0.160	0.191	0.142	0.152

Table 1. The memory usage and time cost of MotionLab.

4. Additional Quantitative Results

As shown in Table 2, our framework outperforms CondMDI on all settings, illustrating the effectiveness of our framework in motion in-between.

Method	Frames	FID↓	R-precision Top-3↑	Diversity→	Foot skating ratio↓	Keyframe error↓
CondMDI [1]	1	0.1551	0.6787	9.5807	0.0936	0.3739
	5	0.1731	0.6823	9.3053	0.0850	0.1789
	20	0.2253	0.6821	9.1151	0.0806	0.0754
Ours	1	0.7547	0.6681	8.9058	0.0779	0.0875
	5	0.0724	0.9146	9.4406	0.0504	0.0283
	20	0.0288	0.9914	9.5447	0.0216	0.0215

Table 2. Evaluation of motion in-between with CondMDI [1] on HumanML3D [2] dataset.

Also, as shown in Figure 2, our framework also outperforms MCM-LDM on all metrics, demonstrating the effectiveness of our framework in motion style transfer.

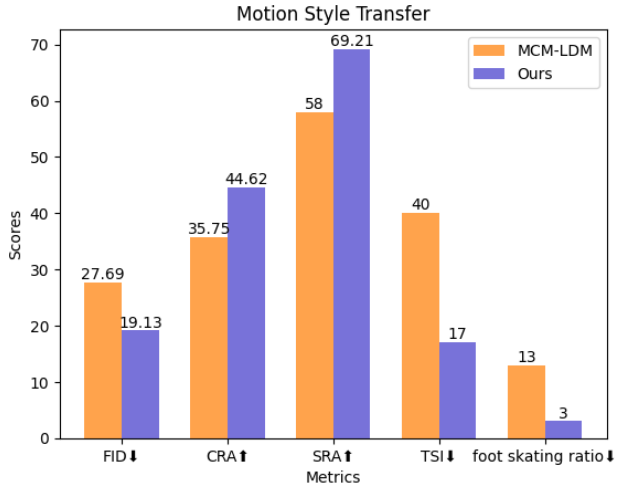

Figure 2. Comparison of the motion style transfer with MCM-LDM [5] on a subset of HumanML3D [2]. This shows that our model has a stronger ability to preserve the semantics of source motion and a stronger ability to learn the style of style motion.

5. Additional Ablation Studies

To further validate the designs in our framework, we perform traditional ablation studies in this section.

To further validate the Aligned ROPE, we also introduce the variant of 3D-Learnable and 3D-ROPE to distinguish the source motion, target motion, and trajectory. As shown in Table 4 and Figure 3, 1D-position encoding is better than 3D-position encoding by avoiding introducing distances between different modalities, and ROPE are better than learnable position encoding by explicit positional encoding. Hence, our 1D-ROPE outperforms all other variants, demonstrating its effectiveness in embedding the position information into tokens.

To further validate the motion curriculum learning, we adopt three variants: removing the masked pre-training and directly supervised fine-tuning in order; with masked pre-training but supervised fine-tuning all tasks together; and introducing masked reconstruction, motion in-between, and trajectory-based motion generation in an orderly manner. As shown in Table 5, motion curriculum learning outperforms all other variants, highlighting the effectiveness of masked pre-training and fine-tuning tasks in order to avoid gradient conflicts between different tasks. Specifically, the variant of masked pre-training in order to demonstrate the necessity of introducing motion in-between and trajectory-based motion generation together, or it will greatly weaken the performance of the model in the latter task.

The explanation of "w/o task instruction modulation" uses 'null' as the instruction for all tasks, rather than learned task tokens or one-hot encoding vectors. We have conducted an additional ablation experiment to examine these situations, which can be suboptimal due to the random initialization of their parameters, as shown in Table 3.

Method	text gen. (FID)	traj. gen. (avg. err.)	text edit (R@1)	traj. edit (R@1)	in-between (avg. err.)	style transfer (CRA)	style transfer (SRA)
w/o task instruction modulation	0.223	0.0401	55.96	70.01	0.0288	40.55	63.91
one-hot encoding	0.187	0.0369	56.18	71.52	0.0287	43.20	66.98
learnable tokens	0.183	0.0356	56.03	71.89	0.0288	41.20	64.98
Ours	**0.167**	**0.0334**	**56.34**	**72.65**	**0.0283**	**44.62**	**69.21**

Table 3. Ablation studies of Task Instruction Modulation.

To further validate the choice and combinations of the tasks, we also introduce the variants of different tasks. As shown in Table 6, improper combination of tasks will cause the unified framework to be weaker than the ours specialist models, while our carefully selected combination of all tasks makes our unified framework beat ours specialist models.

6. Representation for Each Modality

We represent the features of all modalities as tokens for the attention mechanism [6]. Specifically, source motion and target motion are represented as $M_S \in \mathbb{R}^{N\times D}$ and $M_T \in \mathbb{R}^{N\times D}$, and we first ignore timestep t here. For the instruction, it is represented as $I \in \mathbb{R}^{1\times 768}$ extracted from the CLIP [4]. For available conditions C, the text is represent as $p \in \mathbb{R}^{77\times 768}$ extracted from the last hidden layer of CLIP, the trajectory is represented as $h \in \mathbb{R}^{N\times J\times 3}$, and the style is represented as $s \in \mathbb{R}^{1\times 512}$ extracted from [8].

7. Instructions for Each Task

As shown in the Table 7, the instructions in the Task Instruction Modulations for each task are presented, which benefits our framework to distinguish different tasks.

8. Classifier Free Guidance for Each Task

As shown in Table 8, strengths of classifier-free guidance for each task are presented, which contribute to the results' quality during sampling. We conduct ablation experiments based on the hyperparameters provided by the baseline and finally obtain the above hyperparameters.

9. 3D Assets

We have borrowed some 3D assets for our video and figure from the Internet, including Dojo Matrix Drunken Wrestlers, Basketball Court, Grandma's Place, DAE Diorama retake – Small farm, DAE Diorama retake – Small farm, Japanese Small Shrine Temple 0002.

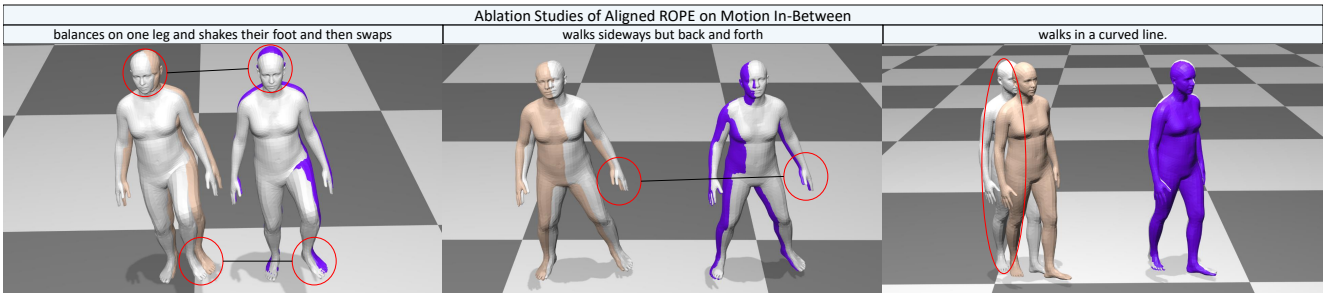

Figure 3. Ablation results of MotionLab on the motion in-between (with text). Beige motion is use 1D-learnable position encoding, purple motion use Aligned ROPE, and gray motions are the poses provided in keyframes, demonstrating the importance of Aligned ROPE.

Method	text gen. (FID)	traj. gen. (avg. err.)	text edit (R@1)	traj. edit (R@1)	in-between (avg. err.)	style transfer (CRA)	style transfer (SRA)
1D-Learnable	0.246	0.0886	45.39	61.99	0.0756	39.40	56.59
3D-Learnable	0.346	0.1865	35.46	53.74	0.1460	36.99	58.81
3D-ROPE	0.241	0.0579	51.34	70.00	0.0354	42.96	62.46
1D-ROPE (ours)	0.167	0.0334	56.34	72.65	0.0273	44.62	69.21

Table 4. Ablation studies of our MotionLab's position encoding on each task.

Method	text gen. (FID)	traj. gen. (avg. err.)	text edit (R@1)	traj. edit (R@1)	in-between (avg. err.)	style transfer (CRA)	style transfer (SRA)
random selection based on FID	2.236	0.1983	28.56	36.61	0.1682	26.61	34.23
removing the masked pre-training	0.861	0.0932	44.99	63.92	0.0639	39.63	57.59
supervised fine-tuning all tasks together	1.331	0.1317	38.19	55.22	0.1143	36.60	50.59
masked pre-training in order	0.256	0.0423	56.33	69.31	0.0264	42.67	64.39
motion curriculum learning (ours)	0.167	0.0334	56.34	72.65	0.0273	44.62	69.21

Table 5. Ablation studies of our MotionLab's motion curriculum learning on each task.

Task						Metric						
text gen.	traj. gen	text edit	traj. edit	in-between	style transfer	text gen. (FID)	traj. gen. (avg. err.)	text edit (R@1)	traj. edit (R@1)	in-between (avg. err.)	style transfer (CRA)	style transfer (SRA)
ours specialist models						0.209	0.0398	41.44	59.86	0.0371	43.53	67.55
✓	×	×	×	×	✓	0.240	-	-	-	-	41.23	65.53
✓	×	✓	×	×	×	0.235	-	52.79	-	-	-	-
✓	✓	×	×	✓	×	0.176	0.0364	-	-	0.0297	-	-
✓	✓	✓	✓	✓	×	0.171	0.0344	55.10	72.20	0.0287	-	-
✓	✓	✓	✓	✓	✓	0.167	0.0334	56.34	72.65	0.0273	44.62	69.21

Table 6. Ablation studies of our MotionLab's task combinations.

Task	Instruction
unconditional generation	"reconstruct given masked source motion."
masked source motion generation	"reconstruct given masked source motion."
reconstruct source motion	"reconstruct given masked source motion."
trajectory-based generation (without text)	"generate motion by given trajectory."
in-between (without text)	"generate motion by given key frames."
style-based generation	"generate motion by given style."
trajectory-based editing	"edit source motion by given trajectory."
text-based editing	"edit source motion by given text."
style transfer	"generate motion by the given style and content."
in-between (with text)	"generate motion by given text and key frames."
trajectory-based generation (with text)	"generate motion by given text and trajectory."
text-based generation	"generate motion by given text."

Table 7. Instructions in the Task Instruction Modulations for each task.

Task	Source Motion Guidance	Condition Guidance
trajectory-based generation (without text)	–	1.5
in-between (without text)	–	1.5
text-based generation	-	5.75
style-based generation	-	1.5
trajectory-based editing (without text)	2.25	2.25
text-based editing	2.25	2.25
style transfer	1.5	1.5
in-between (with text)	–	1.75
trajectory-based generation (with text)	–	1.75
trajectory-based editing (with text)	2	2

Table 8. Strength of classifier free guidance for each task.

References

[1] Setareh Cohan, Guy Tevet, Daniele Reda, Xue Bin Peng, and Michiel van de Panne. Flexible motion in-betweening with diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–9, 2024. 1

[2] 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 (CVPR)*, pages 5152–5161, 2022. 1, 2

[3] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages

8748–8763. PMLR, 2021. 2

[5] Wenfeng Song, Xingliang Jin, Shuai Li, Chenglizhao Chen, Aimin Hao, Xia Hou, Ning Li, and Hong Qin. Arbitrary motion style transfer with multi-condition motion latent diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 821–830, 2024. 2

[6] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2

[7] Wenliang Zhao, Minglei Shi, Xumin Yu, Jie Zhou, and Jiwen Lu. Flowturbo: Towards real-time flow-based image generation with velocity refiner. *arXiv preprint arXiv:2409.18128*, 2024. 1

[8] Lei Zhong, Yiming Xie, Varun Jampani, Deqing Sun, and Huaizu Jiang. Smoodi: Stylized motion diffusion model. In *European Conference on Computer Vision*, pages 405–421. Springer, 2025. 2