Motion Synthesis with Sparse and Flexible Keyjoint Control

Supplementary Material

In the supplementary materials, we describe the implementation details (Sec. 1) and present ablation results on different sampling strategy (Sec. 2). Please refer to the supplementary video on our project page for additional qualitative results.

1. Further Details

1.1. Model Details

We designed our each model using a diffusion framework based on DDPM [4], incorporating the U-Net architecture introduced by [6]. The training was conducted with the AdamW optimizer [7], where we set the learning rate to 1×10^{-4} and applied a weight decay of 1×10^{-2}. For inference, we utilized classifier-free guidance with a weight $\mathbf{w} = 2.5$. Additional details on the hyperparameters for both the network architecture and diffusion process are summarized in Table 1. Both models were trained using a diffusion process with 50 steps. The key-joint diffusion model is comparatively smaller in size in contrast to the full-body completion diffusion model, which consists of a channel dimension of 128.

Hyperparameter	Key-Joint Model	Full-Body Model
Learning rate	1e-4	1e-4
Optimizer	Adam W	Adam W
Weight decay	1e-2	1e-2
Batch size	64	64
Channels dim	128	512
Channel multipliers	[2, 2, 2, 2]	[2, 2, 2, 2]
Variance scheduler	Cosine [8]	Cosine [8]
Diffusion steps	50	50
Diffusion variance	$\tilde{\beta} = \frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t$	$\tilde{\beta} = \frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t$
EMA weight (β)	0.9999	0.9999
Guidance weight ($\mathbf{w}$)	2.5	2.5

Table 1. Hyperparameters of each model

1.2. Baseline Details

For the baselines OmniControl [10], TLControl [9], MotionLCM [3], and DNO [5], we utilize the officially released checkpoints for evaluation. For CondMDI [2], in order to support arbitrary joint-level control with their strategy, we train the model using global position representations for all joints.

1.3. Goal-Driven Motion Synthesize Task

To capture the spatial relationship between body and target locations, we introduce a body shape encoding. This encoding is represented as a continuous shape feature $\mathbf{b}$, which is derived from a set of key measurements. These measurements include joint-to-joint distances obtained from the T-pose, such as: $[root, head]$, $[left\_shoulder, right\_shoulder]$, $[shoulder, wrist]$, $[left\_pelvis, right\_pelvis]$, $[pelvis, feet]$. Additionally, depth measurements for chest and hip thicknesses are computed by evaluating the distances between their front and rear vertices. These measurements collectively form a compact and continuous representation of the body's proportions. In order to model various motion dynamics within a unified framework, we assign a unique action label to each task, conditioning the network on these labels.

We evaluate our approach to goal-driven motion synthesis task by training a unified network that integrates multiple scenarios. In each scenario, control signals are provided as an initial pose paired with target control joints at the final frame. Specifically, *reaching target hand positions* [1] focuses on controlling the right-hand position, *climbing with rock constraints* [11] involves controlling both hands and feet, and *sitting with hand control* [12] requires controlling both hands at the final frame. Our unified network is trained to address multiple dynamics and control settings simultaneously.

Dataset Description We collect a variety of tasks that require control over different target joints at the final frame and involve multiple motion dynamics. For the scenario of *reaching target hand positions*, we utilize the dataset from [1]. Specifically, we extract sequences of reaching motions and augment them by mirroring the left-hand reaching motions to the right hand, thus generating a total of 3,138 right-hand reaching sequences. To define the goal position, we identify the farthest point reached by the hand from its initial location. In our experimental setup, 2,510 samples are designated for training. In the case of the *climbing with rock constraints* scenario, we leverage the dataset from [11]. We carefully select sequences that depict the subject detaching from one climbing rock and securely reaching for another. This process yields 156 motion samples. For the *sitting with hand control* scenario, we use the dataset from [12]. We extract 160 sequences that begin with the subject in a stable position and end when they are seated in a chair. To ensure consistency across all tasks, we standardize the dataset by aligning the subject's face direction to the $+z$ axis at the initial frame and setting the root position at the origin.

2. Ablation Study on Sampling Strategy

We perform a quantitative evaluation of various sampling strategies for diffusion models. Specifically, we train a keyjoint trajectory model and a full-body completion model using 50-step diffusion models and apply both DDPM and DDIM-based diffusion sampling strategies. The results demonstrate that our method maintains consistent performance even with 5-step diffusion sampling and continues to outperform other baselines, achieving high precision and natural motion quality, as supported by Figure 3 of the main paper.

Sampling	Frame Select	FID ↓	Control Err. (m) ↓	R-precision (Top-3) ↑	Div. →	Foot Skating ↓
-	-	0.002	0.000	0.797	9.503	0.000
DDPM (50)	$r = 1$	0.127	0.019	0.681	9.518	0.071
	$r = 2$	0.128	0.019	0.680	9.539	0.070
	$r = 5$	0.148	0.024	0.681	9.554	0.069
	$r = 10$	0.171	0.027	0.678	9.402	0.074
	$r = 20$	0.195	0.033	0.677	9.575	0.064
	$r = 30$	0.224	0.036	0.673	9.674	0.061
	$r = 60$	0.263	0.044	0.659	9.627	0.062
DDIM (10)	$r = 1$	0.136	0.019	0.678	9.559	0.075
	$r = 2$	0.141	0.020	0.681	9.574	0.074
	$r = 5$	0.158	0.022	0.682	9.599	0.073
	$r = 10$	0.186	0.025	0.675	9.632	0.069
	$r = 20$	0.222	0.030	0.681	9.644	0.067
	$r = 30$	0.254	0.037	0.673	9.621	0.063
	$r = 60$	0.297	0.041	0.661	9.615	0.065
DDIM (5)	$r = 1$	0.127	0.019	0.678	9.528	0.072
	$r = 2$	0.140	0.019	0.686	9.573	0.073
	$r = 5$	0.164	0.022	0.678	9.582	0.072
	$r = 10$	0.202	0.025	0.674	9.543	0.065
	$r = 20$	0.236	0.031	0.665	9.521	0.069
	$r = 30$	0.338	0.037	0.655	9.356	0.065
	$r = 60$	0.658	0.045	0.622	9.147	0.073

Table 2. Quantitative evaluation of various sampling strategies for diffusion models.

3. Additional Results

3.1. Results on Challenging or Unseen Scenarios

We further conduct experiments using manually specified control signals in challenging forms (e.g., S-curves or straight lines), which are unseen and difficult scenarios. These results highlight the robustness and generalization capability of our method (Figure 1).

3.2. Results on More Expressive Prompts

We provide results for more expressive and complex prompts in Figure 2. Our method generates rich, detailed, and vivid motions, demonstrating its ability to capture expressive textual descriptions while faithfully adhering to sparse control signals such as "flying like an airplane," "extending arms," "walks backward," and "walking up the stairs".

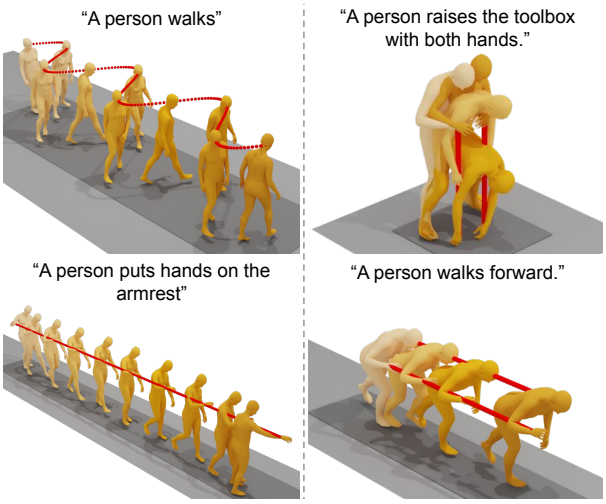

Figure 1. Qualitative results on manually specified challenging control signals, such as S-curves and straight paths. Our method successfully follows the intended trajectories, demonstrating strong generalization to unseen and difficult motion constraints.

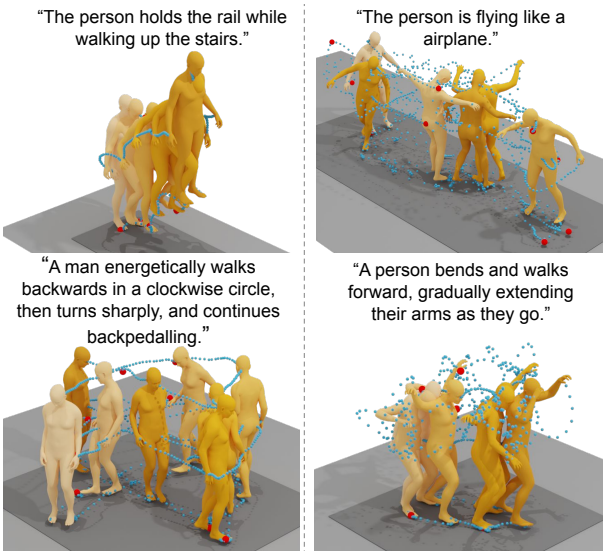

Figure 2. Motion generation results for expressive prompts. Our method produces vivid and diverse motions in response to complex textual descriptions while respecting the given sparse control signals.

References

[1] Joao Pedro Araujo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Deepak Gopinath, Jiajun Wu, Alexander Clegg, and C. Karen Liu. Circle: Capture in rich contextual environments, 2023. 1

[2] 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

[3] Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. In *ECCV*, pages 390–408, 2025. 1

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 1

[5] Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Suwajanakorn, and Siyu Tang. Optimizing diffusion noise can serve as universal motion priors. In *arxiv:2312.11994*, 2023. 1

[6] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2151–2162, 2023. 1

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 1

[8] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021. 1

[9] Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. Tlcontrol: Trajectory and language control for human motion synthesis. *arXiv preprint arXiv:2311.17135*, 2023. 1

[10] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. *arXiv preprint arXiv:2310.08580*, 2023. 1

[11] Ming Yan, Xin Wang, Yudi Dai, Siqi Shen, Chenglu Wen, Lan Xu, Yuexin Ma, and Cheng Wang. Cimi4d: A large multimodal climbing motion dataset under human-scene interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12977–12988, 2023. 1

[12] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. 2022. 1