

Appendix

We further discuss our proposed approach with the following supplementary materials:

- Appendix A: Diffusion Preliminary.
- Appendix B: Additional Implementation and Metric Details.
- Appendix C: Detailed ACMDM Model and Patch Size Scaling Results.
- Appendix D: Detailed Text Driven Controllable Motion Generation Results.
- Appendix E: Quantitative Text-to-Motion Generation Results on the KIT Dataset.
- Appendix F: Text Driven Controllable Motion Generation Results with DNO Approach.
- Appendix G: Quantitative Results on Autoregressive Diffusion Models.
- Appendix H: Benefit of Direct Text-to-SMPL-H Mesh Vertices Motion Generation.
- Appendix I: An Explanations on Fully Absolute in Global Space v.s. Joint-Locally Absolute.
- Appendix J: Additional Qualitative Results of ACMDM.
- Appendix K: Computation Resources and Training Time.
- Appendix L: Limitations.

A Diffusion-based Text-to-Motion Generation Formulation.

Diffusion-based text-to-motion models obtains noisy versions of ground-truth motion \mathbf{x}_0 through an interpolation process. Following DDPM [29], the forward process is:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad (1)$$

where $\bar{\alpha}_t$ controls the pace of the diffusion process where $0 = \bar{\alpha}_T < \dots < \bar{\alpha}_0 = 1$ with assumption that $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Alternatively, flow-matching [57] methods define a linear interpolation:

$$\mathbf{x}_t = (1 - t)\mathbf{x}_0 + t\epsilon, \quad (2)$$

with a continuous timestep where $t \in [0, 1)$.

During training, the model predicts a diffusion target, such as \mathbf{x}_0 or ϵ for DDPM-based methods, and velocity \mathbf{v} for flow-matching methods given \mathbf{x}_t and t . The diffusion models are typically optimized using a simple MSE loss between the predicted target and its ground truth counterpart.

During inference, starting from random Gaussian noise \mathbf{x}_T , the model iteratively predicts intermediate states by estimating \mathbf{x}_0 , ϵ , or \mathbf{v} and updates to \mathbf{x}_{t-1} via the learned reverse process: typically solving an SDE function for DDPM-based methods, or an ODE function for flow-matching methods.

B Additional Implementation and Metric Details

AE Model Details All AutoEncoder variants use a 3-block of 3-layer ResNet-based encoder-decoder architecture with a hidden dimension of 512, output latent channel of 4, and a total temporal downsampling factor of 4. All AutoEncoders are trained with a batch size of 256, where each sample contains 64 frames. We train for 50 epochs, and apply learning rate decay by a factor of 20 at the 150,000th iteration.

KIT Dataset Details KIT-ML includes 3,911 motion clips from the KIT and CMU [67] datasets, annotated with 6,278 textual descriptions (1–4 per motion), and downsampled to 12.5 FPS.

Evaluation Metric Details We adopt the more robust and recent evaluation framework proposed in [66], which focuses on essential, animatable dimensions of generated motion. Specifically, we use the following metrics following [25, 66]: (1) R-Precision (Top-1, Top-2, and Top-3 accuracies) and Matching, which measures the semantic alignment between generated motion embeddings and their corresponding captions’ glove embedding; (2) Fréchet Inception Distance (FID), which assesses the statistical similarity between ground truth and generated motion distributions; and (3) MultiModality, which measures the diversity of generated motion embeddings per same text prompt. (4) CLIP-Score, which is the cosine similarity between the generated motion and its caption via CLIP embeddings.

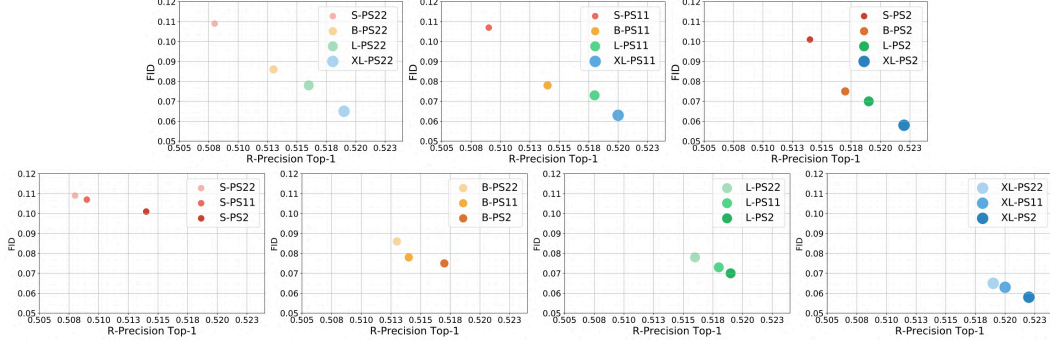


Figure A1: **Model and patch size scaling results of ACMDM.** Top row: FID and R-Precision Top 1 are compared while holding patch size constant. Bottom row: Results are shown while holding model size constant. Our model exhibits strong scalability with increasing model capacity and decreasing patch size.

For trajectory-control-specific evaluations, following [42], we additionally report the following metrics: Diversity, which measures variability within the generated motions; Foot Skating Ratio, which indicates the physical plausibility of the motion by quantifying slippage artifacts; Trajectory Error, Location Error, and Average Joint Error, which evaluate the accuracy of controlled joint positions at keyframes. To assess performance under varying supervision levels, we report the average results over five different control sparsity levels—using randomly sampled 1%, 2%, 5%, 25%, and 100% of the ground-truth keyframes as control inputs. During training, control keyframe intensities are randomly sampled.

For mesh vertex generation, we also include Laplacian Surface Distance (LSD) to assess the quality of the generated mesh that preserves the structural shape of the ground-truth T-pose.

For all ACMDM evaluations, we convert absolute joint coordinates or extract joints from generated mesh vertices and process them into essential HumanML3D evaluation features for consistent and fair comparisons across all ACMDM and baseline methods.

C Detailed ACMDM Model and Patch Size Scaling Results

In Figure 4 of the main paper, we visualize the scalability of ACMDM across different model and patch sizes. Here, we provide the complete table of results in Table A1 and further visualization results in Figure A1. In Table A1, we cover all ACMDM variants with and without classifier-free guidance. These results further underscore that our proposed formulation scales effectively, consistently benefiting from increased model capacity and finer spatial resolution to achieve strong improvements in motion generation quality.

D Detailed Text Driven Controllable Motion Generation Results

In the main paper, we presented a summarized version of the controllable motion generation comparison results. In Table A2, we provide the complete evaluation table across all joints, following the protocol of OmniControl [103]. Compared to prior methods, our approach not only achieves superior performance on every controlled joint but also enables significantly faster inference (2.51 AITS v.s. 81.0 for OmniControl), demonstrating both efficiency and effectiveness of our proposed formulation.

E Quantitative Text-to-Motion Generation Results on the KIT Dataset

In Table A3, we present quantitative text-to-motion generation results on the KIT dataset, comparing ACMDM against state-of-the-art baselines. Notably, even our smallest variant, ACMDM-S-PS22, already surpasses all prior methods across key evaluation metrics such as FID, R-Precision, Matching Score, and CLIP-Score, further demonstrating the effectiveness of our proposed formulation.

Table A1: **ACMDM model and patch size scaling results** on the HumanML3D dataset grouped by model size and patch size. We present all ACMDM variants’ performances with and without CFG.

Size	Transformer	Patch	CFG	FID ↓	R-Precision ↑			Matching↓	CLIP-score↑
					Top 1	Top 2	Top 3		
S	8 head 512 dim	22	✗ ✓	0.178±.009 0.109±.005	0.399±.002 0.508±.002	0.577±.003 0.701±.003	0.682±.003 0.798±.003	3.938±.013 3.253±.010	0.558±.001 0.639±.001
		11	✗ ✓	0.153±.010 0.107±.004	0.415±.002 0.509±.003	0.596±.002 0.704±.002	0.698±.003 0.799±.002	3.826±.010 3.251±.008	0.571±.001 0.642±.001
		2	✗ ✓	0.149±.009 0.101±.005	0.424±.003 0.514±.003	0.606±.003 0.707±.002	0.707±.003 0.802±.001	3.764±.011 3.227±.009	0.578±.001 0.644±.001
B	12 head 768 dim	22	✗ ✓	0.145±.011 0.086±.004	0.435±.002 0.513±.003	0.618±.003 0.707±.003	0.719±.003 0.801±.003	3.697±.013 3.214±.010	0.589±.001 0.646±.001
		11	✗ ✓	0.144±.009 0.078±.003	0.448±.003 0.514±.002	0.633±.002 0.709±.003	0.731±.002 0.802±.002	3.627±.012 3.211±.009	0.597±.001 0.647±.001
		2	✗ ✓	0.141±.010 0.075±.004	0.446±.003 0.517±.002	0.634±.002 0.710±.003	0.733±.002 0.803±.003	3.613±.010 3.209±.008	0.598±.001 0.648±.001
L	16 head 1024 dim	22	✗ ✓	0.181±.009 0.078±.003	0.447±.003 0.516±.003	0.630±.003 0.709±.003	0.731±.003 0.803±.002	3.628±.011 3.210±.009	0.601±.001 0.648±.001
		11	✗ ✓	0.175±.007 0.073±.003	0.451±.002 0.518±.002	0.637±.003 0.710±.003	0.738±.003 0.803±.003	3.591±.015 3.208±.011	0.604±.001 0.649±.001
		2	✗ ✓	0.171±.009 0.070±.003	0.459±.003 0.519±.003	0.643±.004 0.711±.003	0.743±.004 0.804±.003	3.556±.013 3.207±.010	0.608±.001 0.650±.001
XL	20 head 1280 dim	22	✗ ✓	0.194±.009 0.065±.004	0.461±.002 0.519±.003	0.645±.003 0.711±.003	0.746±.003 0.805±.003	3.542±.013 3.209±.009	0.611±.001 0.650±.001
		11	✗ ✓	0.181±.008 0.063±.004	0.462±.002 0.520±.003	0.650±.003 0.712±.002	0.750±.003 0.806±.002	3.532±.011 3.206±.009	0.611±.001 0.651±.001
		2	✗ ✓	0.173±.008 0.058±.004	0.467±.002 0.522±.002	0.655±.003 0.713±.002	0.757±.003 0.807±.002	3.521±.010 3.205±.008	0.613±.001 0.652±.001

F Text Driven Controllable Motion Generation Results with DNO Approach

In Table A4, we demonstrate that our absolute coordinate formulation also supports input noise optimization following DNO [41] for text-driven controllable motion generation. However, we strongly discourage using this approach due to its heavy time cost (27.8 AITS) from multi-round optimization and high computational burden from gradient accumulation over 10 iterations with the Euler ODE Solver. Employing higher-order solvers like Euler-50 or DOPRI-5 would further increase both gradient steps and inference time, making the method impractical.

G Quantitative Results on Autoregressive Diffusion Models.

Our absolute coordinate formulation is not limited to standard diffusion models, it also generalizes well to autoregressive (AR) diffusion approaches. In Table A5, we report results using three AR variants: (1) Masked AR, which predicts masked latent segments conditioned on previous unmasked motion; (2) Prefix AR, which generates future motion autoregressively from a fixed-length 20-frame prefix; and (3) Noisy Conditioned AR, which is trained on noisy versions of arbitrary-length prefixes and performs inference with clean prefixes. Across all AR variants, our absolute coordinate formulation consistently achieves strong performance across evaluation metrics, highlighting the flexibility and effectiveness of our approach.

H Benefit of Direct Text-to-SMPL-H Mesh Vertices Motion Generation

Compared to the common pipeline of generating joints followed by mesh fitting, direct SMPLH-H mesh generation produces more natural mesh motion without jittering body parts. It can implicitly model nuanced hand and dynamic flesh movements and help to prevent self-penetration. We provide qualitative visualizations in Appendix J to further illustrate these advantages.

Table A2: **Quantitative text-conditioned motion generation with spatial control signals and upper-body editing on HumanML3D.** In the first section, methods are trained and evaluated solely on pelvis controls. In the middle section, methods are trained on all joints and evaluated separately on each controlled joint. The last section presents upper-body editing results. **bold face / underline** indicates the best/2nd results.

Controlling Joint	Methods	AITS↓	Classifier Guidance	FID↓	R-Precision Top 3	Diversity→	Foot Skating Ratio.↓	Traj. err.↓	Loc. err.↓	Avg. err.↓
	GT	—	-	0.000	0.795	10.455	-	0.000	0.000	0.000
Train On Pelvis	MDM [92]	16.34	✗	1.792	0.673	9.131	0.1019	0.4022	0.3076	0.5959
	PriorMDM [84]	20.19	✗	0.393	0.707	9.847	0.0897	0.3457	0.2132	0.4417
	GMD [42]	137.63	✓	0.238	0.763	10.011	0.1009	0.0931	0.0321	0.1439
	OmniControl [103]	81.00	✓	<u>0.081</u>	<u>0.789</u>	<u>10.323</u>	0.0547	<u>0.0387</u>	<u>0.0096</u>	<u>0.0338</u>
	MotionLCM V2+CtrlNet [14]	0.066	✗	3.978	0.738	9.249	0.0901	0.1080	0.0581	0.1386
	ACMDM-S-PS22+CtrlNet	<u>2.51</u>	✗	0.067	0.805	10.481	<u>0.0591</u>	0.0075	0.0010	0.0100
Pelvis	OmniControl [103]	81.00	✓	0.135	0.790	<u>10.314</u>	0.0571	<u>0.0404</u>	<u>0.0085</u>	<u>0.0367</u>
	MotionLCM V2+CtrlNet [14]	0.066	✗	4.726	0.713	9.209	0.1162	0.1617	0.0841	0.1838
	ACMDM-S-PS22+CtrlNet	<u>2.51</u>	✗	0.075	0.805	10.536	<u>0.0603</u>	0.0081	0.0011	0.0134
Left foot	OmniControl [103]	81.0	✓	<u>0.093</u>	0.794	<u>10.338</u>	<u>0.0692</u>	<u>0.0594</u>	<u>0.0094</u>	<u>0.0314</u>
	MotionLCM V2+CtrlNet [14]	0.066	✗	4.810	0.706	9.158	0.1047	0.2607	0.1229	0.2304
	ACMDM-S-PS22+CtrlNet	<u>2.51</u>	✗	0.063	0.800	10.542	0.0590	0.0186	0.0034	0.0240
Right foot	OmniControl [103]	81.00	✓	<u>0.137</u>	<u>0.798</u>	<u>10.241</u>	<u>0.0668</u>	<u>0.0666</u>	<u>0.0120</u>	<u>0.0334</u>
	MotionLCM V2+CtrlNet [14]	0.066	✗	4.756	0.705	9.303	0.1026	0.2459	0.1127	0.2278
	ACMDM-S-PS22+CtrlNet	<u>2.51</u>	✗	0.071	0.803	10.591	0.0583	0.0205	0.0030	0.0251
Head	OmniControl [103]	81.00	✓	<u>0.146</u>	<u>0.796</u>	<u>10.239</u>	0.0556	<u>0.0422</u>	<u>0.0079</u>	<u>0.0349</u>
	MotionLCM V2+CtrlNet [14]	0.066	✗	4.580	0.715	9.278	0.1138	0.1971	0.0977	0.2136
	ACMDM-S-PS22+CtrlNet	<u>2.51</u>	✗	0.081	0.805	10.520	<u>0.0598</u>	0.0051	0.0009	0.0152
Left wrist	OmniControl [103]	81.00	✓	<u>0.119</u>	<u>0.783</u>	<u>10.217</u>	0.0562	<u>0.0801</u>	<u>0.0134</u>	<u>0.0529</u>
	MotionLCM V2+CtrlNet [14]	0.066	✗	4.103	0.726	9.188	0.1167	0.3965	0.1912	0.3150
	ACMDM-S-PS22+CtrlNet	<u>2.51</u>	✗	0.065	0.804	10.480	<u>0.0604</u>	0.0085	0.0014	0.0206
Right wrist	OmniControl [103]	81.00	✓	<u>0.128</u>	<u>0.792</u>	<u>10.309</u>	<u>0.0601</u>	<u>0.0813</u>	<u>0.0127</u>	<u>0.0519</u>
	MotionLCM V2+CtrlNet [14]	0.066	✗	4.051	0.725	9.242	0.1176	0.3822	0.1806	0.3079
	ACMDM-S-PS22+CtrlNet	<u>2.51</u>	✗	0.066	0.802	10.484	0.0599	0.0091	0.0016	0.0201
Average	OmniControl [103]	81.00	✓	<u>0.126</u>	<u>0.792</u>	<u>10.276</u>	<u>0.0608</u>	<u>0.0617</u>	<u>0.0107</u>	<u>0.0404</u>
	MotionLCM V2+CtrlNet [14]	0.066	✗	4.504	0.715	9.230	0.1119	0.2740	0.1315	0.2464
	ACMDM-S-PS22+CtrlNet	<u>2.51</u>	✗	0.070	0.803	10.526	0.0596	0.0117	0.0019	0.0197
	Methods	AITS↓	Classifier Guidance	FID↓	R-Precision Top 1	R-Precision Top 2	R-Precision Top 3	Matching↓	Diversity→	-
UpperBody Edit	MDM [92]	16.34	✗	1.918	0.359	0.556	0.654	4.793	9.210	-
	OmniControl [120]	81.00	✓	<u>0.909</u>	<u>0.428</u>	<u>0.614</u>	<u>0.722</u>	<u>3.694</u>	<u>10.207</u>	-
	MotionLCM V2+CtrlNet [120]	0.066	✗	3.922	0.404	0.592	0.692	5.610	9.309	-
	ACMDM-S-PS22+CtrlNet	<u>2.51</u>	✗	0.076	0.532	0.719	0.820	3.098	10.586	-

Table A3: **Quantitative text-to-motion evaluation on KIT dataset.** We repeat the evaluation 20 times and report the average with 95% confidence interval. We use **bold face / underline** to indicate the best/2nd results.

Methods	R-Precision↑			FID↓	Matching↓	MModality↑	CLIP-score↑
	Top 1	Top 2	Top 3				
MDM [92]	0.333±.012	0.561±.009	0.689±.009	0.585±.043	4.002±.033	1.681±.107	0.605±.007
MotionDiffuse [120]	0.344±.009	0.536±.007	0.658±.007	3.845±.087	4.167±.054	<u>1.774±.217</u>	0.626±.006
ReMoDiffuse [121]	0.356±.004	0.572±.007	0.706±.009	1.725±.053	3.735±.036	1.928±.127	0.665±.005
MARDM [66]-ε	0.375±.006	0.597±.008	0.739±.006	0.340±.020	3.489±.018	1.479±.078	0.681±.003
MARDM [66]-v	<u>0.387±.006</u>	<u>0.610±.006</u>	<u>0.749±.006</u>	<u>0.242±.014</u>	<u>3.374±.019</u>	1.312±.053	<u>0.692±.002</u>
ACMDM-S-PS22	0.391±.005	0.615±.005	0.752±.006	0.237±.010	3.368±.019	1.267±.063	0.696±.002

I An Explanations on Fully Absolute in Global Space v.s. Joint-Locally Absolute

In HumanML3D [25], absolute joint coordinates are represented as a 22×3 array per frame. Flattening this to a 66-dimensional vector and computing mean/std normalization per channel leads to a different outcome than computing statistics directly over the three XYZ channels. This is because in the flattened format, after Z-Normalization, even the same value on the same axis but from different joints can correspond to entirely different spatial positions in global space. In contrast, our formulation performs z-normalization directly across the XYZ channels in the global coordinate space, where the

Table A4: **Quantitative results of DNO-style input noise optimization with ACMDM** for text-conditioned motion generation under spatial control on HumanML3D dataset.

Controlling Joint	Methods	AITs↓	FID↓	R-Precision Top 3	Foot Skating Ratio↓	Traj. err.↓	Loc. err.↓	Avg. err.↓
Pelvis	ACMDM-S-PS22+DNO	27.8	0.151	0.802	0.0610	0.0027	0.0002	0.0089
Left foot	ACMDM-S-PS22+DNO	27.8	0.147	0.799	0.0602	0.0082	0.0003	0.0133
Right foot	ACMDM-S-PS22+DNO	27.8	0.153	0.800	0.0597	0.0086	0.0003	0.0138
Head	ACMDM-S-PS22+DNO	27.8	0.138	0.801	0.0591	0.0025	0.0002	0.0084
Left wrist	ACMDM-S-PS22+DNO	27.8	0.149	0.799	0.0600	0.0076	0.0004	0.0138
Right wrist	ACMDM-S-PS22+DNO	27.8	0.143	0.798	0.0598	0.0081	0.0004	0.0142
Average	ACMDM-S-PS22+DNO	27.8	0.147	0.800	0.0600	0.0034	0.0003	0.0121

Table A5: **Quantitative results of autoregressive diffusions using our absolute coordinate formulation** on the HumanML3D dataset. Our approach consistently performs well across AR variants.

Model & Patch Size	AR Method	First Prefix Type	FID↓	R-Precision↑			Matching↓	CLIP-score↑
				Top 1	Top 2	Top 3		
ACMDM S-PS2	Prefix AR	Generated	0.117±.006	0.496±.002	0.690±.002	0.786±.003	3.354±.008	0.634±.002
	Prefix AR	GT	0.042±.002	0.504±.003	0.700±.003	0.798±.003	3.212±.007	0.640±.001
	Noisy Cond AR	Generated	0.115±.006	0.497±.003	0.690±.004	0.788±.003	3.343±.010	0.636±.003
	Masked AR	Generated	0.111±.005	0.509±.003	0.702±.003	0.799±.003	3.250±.009	0.643±.002

1176 same numeric value consistently refers to the same physical dimension, enhancing spatial coherence
1177 and global awareness during model training.

1178 J Additional Qualitative Results of ACMDM

1179 We provide comprehensive video visualizations hosted on a locally-run, anonymous HTML page
1180 to further demonstrate the effectiveness of our approach. These visualizations include detailed
1181 comparisons with state-of-the-art text-to-motion generation baselines, showcasing that our method
1182 produces more realistic and semantically aligned motions. We also present side-by-side comparisons
1183 with existing text-driven controllable motion generation methods, highlighting that our approach not
1184 only achieves higher accuracy but also enables significantly faster inference. Additional visualizations
1185 illustrate our method’s ability to generate diverse and contextually appropriate motions that accurately
1186 follow control signals, including spatial editing scenarios. Furthermore, we demonstrate the benefits
1187 of directly generating SMPL-H mesh vertices. Compared to the common pipeline of generating joints
1188 followed by mesh fitting, our direct mesh generation results in more natural and expressive human
1189 motion, including subtle details like soft tissue and flesh dynamics. We showcase additional examples
1190 to highlight the quality and realism of our generated mesh-based motions.

1191 K Computation Resources and Training Time

1192 All ACMDM models were trained using either NVIDIA RTX 4090 or H200 GPUs, depending on
1193 model size. Smaller variants, such as ACMDM-S-PS22, were trained on a single RTX 4090 GPU
1194 and required approximately 8 hours of training. In contrast, the largest variant, ACMDM-XL-PS2,
1195 was trained on an H200 GPU and took approximately 2 days to complete.

1196 L Limitations

1197 While ACMDM demonstrates strong scalability and performance, scaling to larger models demands
1198 substantial computational resources and extended training times.