Appendix

This appendix provides more qualitative results (Appendix A), dataset details (Appendix B), user study (Appendix C), and limitations (Appendix D).

A. Qualitative Results

We show more qualitative results from both in-domain and out-of-domain text inputs of several text-motion datasets. First, we show the generated motions of our OMG model from the text inputs in the HumanML3D test set. As illustrated in Fig. B, our model enables realistic and diverse motion generation from complicated natural sentences. Then, we show the out-of-domain generation capability using text inputs of the Mixamo test set and the concurrent MotionX [44] dataset. As illustrated in Fig. C and Fig. D, our model well-handles unseen high-level descriptions of motion traits, like "scary clown" or "imitating snake".

B. Dataset Details

Here we provide the details of the motion-only datasets used at the pre-training stage, as illustrated in Tab. A. We utilize 13 publicly available human motion datasets captured from various motion modalities, such as artist-created datasets [23, 50], marker-based [7, 30, 46, 49, 77], IMU-based [41, 80] and multi-view markerless [10, 40, 42, 93] motion capture datasets, totaling over 22 million frames. Since the majority of motion data is in SMPL format, we apply the retargeting algorithm to standardize them to the SMPL skeleton with rotations and positions of 22 joints, and global translation.

Moreover, we utilize HumanML3D [20] training set to train our motion ControlNet for fair comparisons with previous methods. The dataset consists of 14616 motion clips with 44970 text annotations, totaling 3.1M motion instances, as illustated in Tab. B. We further introduce Mixamo [1] dataset, consisting of abundant artist-created animations and human-annotated descriptions. It is widely used in character animation applications, such as games and VR/AR. We employ it to benchmark the zero-shot performance due to its wide variety of diverse and dynamic motions and complicated and abstract motion trait descriptions.

C. User Study

For the comparisons of the user study presented in Fig. A, we ask the users to "Rate the motion based on how realistic it is" and "Rate the match between motion and prompt". The provided motions are generated from 60 text descriptions, 30 of which are randomly generated from the HumanML3D [20] test set and 30 from Mixamo [1] test set. We invite 20 users, shuffle the order of results from the distinct compared methods, and ask them to complete the rating, as illustrated in Fig. E. As shown in Fig. A, our OMG was preferred over the other state-of-the-art methods in both motion quality and text-motion alignment.

Dataset	Duration (h)	Frame Number	Mocap Modality	Motion Format
HCM [41]	2.9	0.3M	IMU	SMPL
AMASS [49]	62.9	6.8M	Marker	SMPL
EgoBody [93]	0.4	0.04M	RGB-D	SMPL
GRAB [77]	3.8	0.4M	Marker	SMPL
AIST++ [40]	4.0	0.4M	RGB	SMPL
HuMMan [10]	0.9	0.1M	RGB-D	SMPL
InterHuman [42]	13.1	1.4M	RGB	SMPL
CIRCLE [7]	10.0	1.1M	Marker	SMPL
BEAT [46]	76	8.2M	Marker	BVH
LaFan1 [23]	4.6	0.5M	Marker	BVH
Human3.6M [30]	5.0	0.5M	Marker	SMPL
Total Capture [80]	0.8	0.09M	IMU	SMPL
100style [50]	22.1	2.4M	marker	BVH
Total	206.5	22.3M	-	-

Table A. The details of unlabeled motion datasets used at the pre-training stage.

Dataset	Clip Number	Text Number	Duration (h)	Frame Number	Motion Format
HumanML3D [19]	14616	44970	28.59	3.1M	SMPL
Mixamo [1]	2254	2254	2.5	0.3M	FBX

Table B. We use HumanML3D training set at the fine-tuning stage and HumanML3D and Mixamo test set for evaluation.

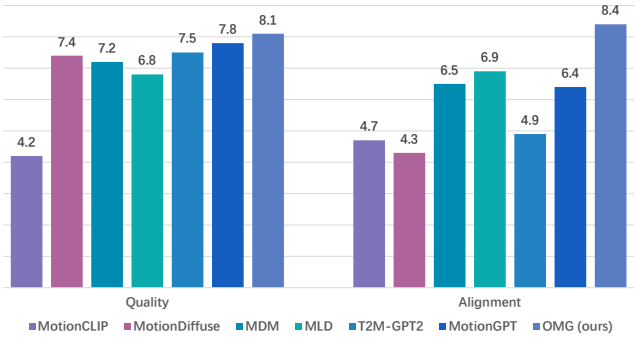

Figure A. **User Study.** We show the average quality rates and the average alignment rates of the compared methods, which indicate human evaluation of both motion quality and text-motion consistency respectively.

D. Limitations

As the trial to explore realistic open-vocabulary motion generation, the proposed OMG still has limitations as follows.

Motion space. Our method still relies on the training motion manifold and cannot generate motions that are beyond the scope of the training data, such as flying, yoga, or swimming.

Precise control. Our method does not explicitly model the temporal order and inclusion relations of sub-motions, which are unable to handle precise control, such as picking an object or reaching a goal.

Physically implausible. Our method does not explicitly

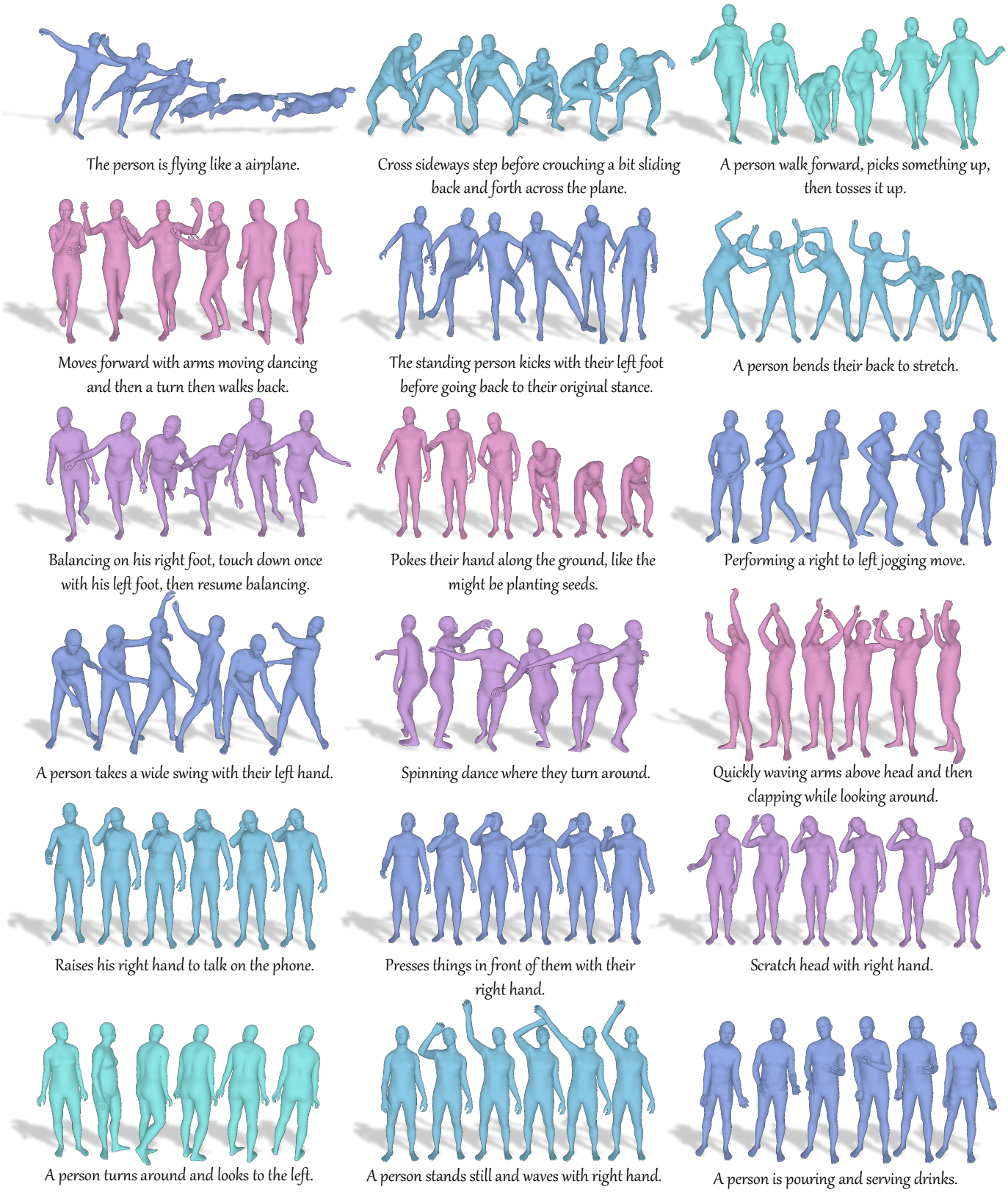

Figure B. Qualitative results on HumanML3D test set.

Figure C. Qualitative results on Mixamo test set.

Figure D. Qualitative results on Motion-X dataset.

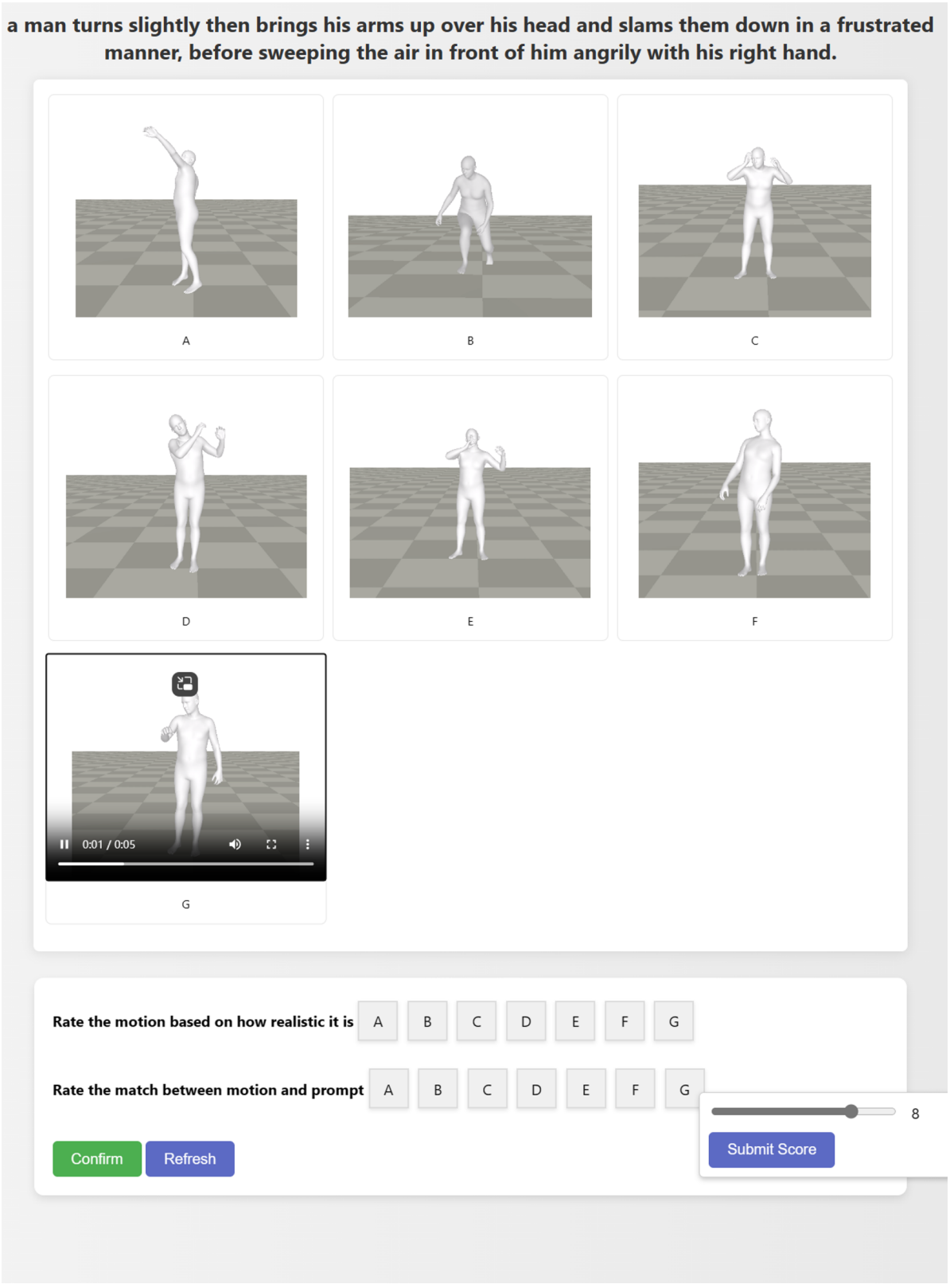

Figure E. **User Study.** We ask 20 users to rate the motion quality and text-motion consistency of 60 results generated from each method. The rating range is from 1 to 10.

model physical dynamics, which leads to physically implausible motion generation. Recent physics-based motion control [24, 86, 87] approaches use reinforcement learning to control human characters in a physically simulated environment, achieving impressive motion quality. It's interesting to introduce physics into the conditional generative model pipeline.

Maximum length. Same as most motion generation methods, our method can generate arbitrary length results but still under the max-length in the dataset. It's interesting to model a non-stop human motion in temporal consistency.

Full-body dynamics. Our method focuses on articulated human bodies. How to model the full-body dynamics including the face, eyes, hands, and even toes, which enables complicated interactions with our complex physical world, remains a huge challenge.