RoMo: A Large-Scale, Richly Organized Dataset and Semantic Taxonomy for Human Motion Generation Supplementary Material

Jiahao Zhang[1,2] Joseph Liu[2] Young-Yoon Lee[2] Seonghyeon Moon[2]
Victor Zordan[2] Guy Tevet[3] C. Karen Liu[3] Stephen Gould[1]
Oren Jacob[2] Haomiao Jiang[2] Mubbasir Kapadia[2,4] Yizhak Ben-Shabat[2]
[1]Australian National University [2]Roblox [3]Stanford University [4]Rutgers University
[1]{jiaho.zhang, stephen.gould}@anu.edu.au
[2]{josephliu, ylee, smoon, vbzordan, haomiaojiang, ojacob, mkapadia, ibenshabat}@roblox.com
[3]{guytevet, karenliu}@cs.stanford.edu
https://davidzhang73.github.io/romo-website

Abstract

*This supplementary material provides extensive quantitative and interactive support for the claims presented in our main paper. We strongly encourage readers to engage with the accompanying interactive website and video, which offer a full demo of the **Motion Toolbox**, a Mindmap visualization of the novel semantic taxonomy, 3D rendered visualizations of motion samples from each category, and detailed analysis plots of our adaptive filtering approach. This document extends the model evaluation to include **MoMask++** and **MDM+BERT**, leveraging the taxonomy for a fine-grained, category-based analysis that reveals architecture-specific trade-offs. Furthermore, we present an ablation study quantifying the optimal trade-off in motion dynamics using our Dynamic Score ($S_{Dynamic}$), demonstrating that a refined subset (V_C) significantly improves generative performance (FID of* 17.97 *vs.* 20.63 *for the full set). Finally, we provide full implementation details for reproducibility and demonstrate the power of our hierarchical taxonomy with the **Context-Aware Adaptive Filtering methodology**, which strategically preserves authentic, category-specific motions while eliminating quality artifacts.*

1. Taxonomy-Based Fine-Grained Evaluation

Building upon the comparison of diffusion-based and GPT architectures in the main paper (Sec. 6), we extend our evaluation to include **MoMask++** [1] and **MDM+BERT** [2]. We report the performance of these models in Tab. 1 and Tab. 2 for semantic and physical metrics, respectively. Leveraging RoMo's broad taxonomy allows for a fine-grained analysis that surfaces intriguing trade-offs between these architectures. For example, while MDM demonstrates strong physical fidelity, MoMask++ achieves higher performance on semantic metrics, particularly FID (Fréchet Inception Distance), suggesting it better captures the statistical properties of the true motion distribution.

1.1. Taxonomy-Based Evaluation Insights

Our hierarchical taxonomy exposes a non-uniform model performance across motion categories, validating the need to move beyond aggregated metrics. For instance, MoMask++ significantly outperforms MDM in FID within categories characterized by subtle motions like *Communication*, *Education*, and *Gestures* (Tab. 3). To validate these variances, we evaluated 1,000 motion samples **randomly selected** per category across the top 10 most frequent categories. These results collectively suggest that diffusion-based approaches (MDM) offer better adherence to physical constraints, while transformer-based approaches (MoMask++) excel at capturing the distributional nuances of complex social and expressive motions.

1.2. Addressing Physical Quality Discrepancies

The performance disparities on physical metrics are pronounced (Tab. 4). MoMask++ shows significantly higher values for Jerk (e.g., 332.73 overall vs. 46.77 for MDM) and Acceleration Peaks, indicating less smooth and more artifact-prone sequences. This is typical for discrete tokenization approaches, where the quantization and sequential generation process can lead to accumulated temporal inconsistencies. Conversely, MDM's holistic, non-autoregressive denoising mechanism appears to enforce superior long-

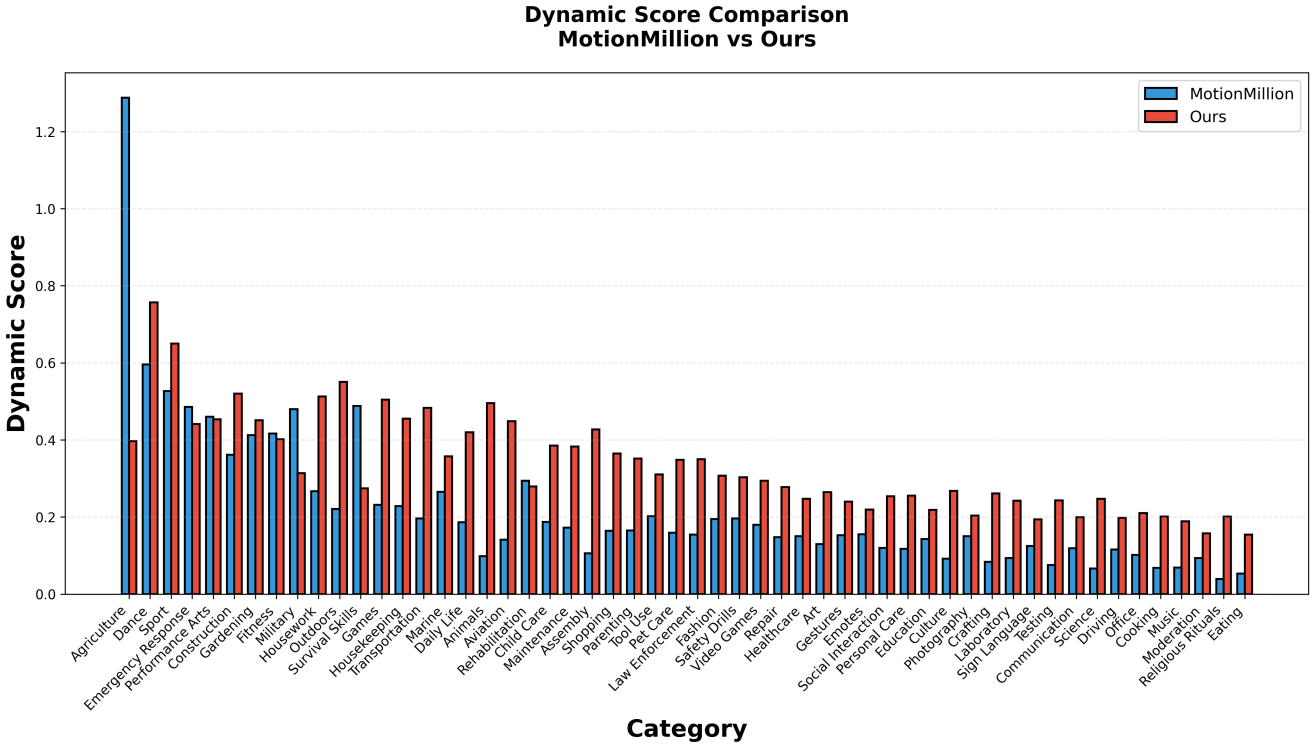

Figure 1. **Dynamic Score analysis.** RoMo demonstrates higher dynamic scores across the majority of categories

Table 1. **Semantic benchmark.** Comparison of text-to-motion generation quality between MoMask++ and MDM trained on the RoMo full dataset.

Method	R Precision			Diversity ↑	FID ↓	Matching Score ↑
	Top1	Top2	Top3			
MDM+BERT	0.5822	0.7578	0.8434	27.59	17.97	12.18
MoMask++	0.5148	0.6910	0.7820	28.17	14.30	12.86

range consistency and physical plausibility, leading to substantially smoother motion.

2. Category-Wise Validation of Superior Motion Dynamics

Building on the global analysis provided in the main paper, we present a comprehensive per-category comparison of Dynamic Scores between our RoMo dataset and MotionMillion in Figure 1. This fine-grained breakdown substantiates our primary finding: RoMo consistently exhibits **higher motion intensity** across the vast majority of categories.

While MotionMillion may achieve higher Dynamic Scores in a few isolated categories, our dataset demonstrates a clear systematic advantage, validating that the substantial global increase in mean Dynamic Score (0.336 for RoMo vs. 0.222 for MotionMillion) is driven not by outliers, but by a systematic improvement in motion quality across the taxonomy. This per-category perspective is essential, as it confirms that RoMo provides a more reliable and vigorous training signal across the entire spectrum of human activities.

3. Optimizing Training Signal via Dynamic Score Ablation

To quantify the impact of motion quality on generative performance, we conduct an ablation study across five mutually exclusive data partitions defined by our dynamic score ($S_{Dynamic}$). Starting from the complete dataset (`All`), we define four distinct subsets (V_A, V_B, V_C, V_D) based on their dynamic range, testing MDM performance on each.

Table 2. **Physical benchmark.** Comparison of physical motion artifacts and dynamic characteristics between MoMask++ and MDM.

Method	Foot Skating ↓	Jerk ↓	Ground Penetration ↓	Floating ↑	Acceleration Peaks ↑	Dynamic Score ↑
MDM+BERT	0.0011	46.77	0.00	0.0146	0.8374	0.2132
MoMask++	0.0022	332.73	2.1e−05	0.0020	4.7223	0.3887

Table 3. **Taxonomy-based semantic evaluation.** Fine-grained breakdown of generation quality comparing MDM and MoMask++ across the top 10 categories.

Category	Method	R Precision			Diversity ↑	FID ↓	Matching Score ↑
		Top1	Top2	Top3			
Overall	MDM	0.5822	0.7578	0.8434	27.59	17.97	12.18
	MoMask++	0.5148	0.6910	0.7820	28.17	14.30	12.86
Communication	MDM	0.3771	0.5363	0.6393	21.60	28.97	9.77
	MoMask++	0.3152	0.4635	0.5621	23.13	14.64	11.08
Daily Life	MDM	0.4782	0.6567	0.7537	26.55	27.80	12.15
	MoMask++	0.4066	0.5783	0.6760	26.56	21.27	12.91
Dance	MDM	0.3514	0.5111	0.6105	22.88	32.02	12.56
	MoMask++	0.3088	0.4459	0.5344	21.48	37.09	12.67
Education	MDM	0.3477	0.5221	0.6349	22.06	31.56	11.02
	MoMask++	0.3080	0.4641	0.5791	23.82	17.59	12.07
Fitness	MDM	0.5777	0.7525	0.8350	27.35	23.78	13.01
	MoMask++	0.5098	0.6895	0.7764	26.05	19.36	12.99
Gestures	MDM	0.4098	0.5723	0.6725	21.81	25.13	10.38
	MoMask++	0.3568	0.5125	0.6125	23.47	14.66	11.39
Housework	MDM	0.3431	0.5093	0.6107	24.38	27.58	12.98
	MoMask++	0.3139	0.4832	0.5900	23.40	24.69	12.96
Outdoors	MDM	0.4262	0.6125	0.7329	26.76	28.36	12.63
	MoMask++	0.3902	0.5689	0.6742	26.20	32.82	13.06
Pet Care	MDM	0.3807	0.5834	0.7004	24.72	33.23	11.56
	MoMask++	0.3018	0.4846	0.6096	24.90	30.54	12.56
Sport	MDM	0.4891	0.6820	0.7787	28.15	20.74	13.73
	MoMask++	0.4207	0.6158	0.7199	26.60	31.80	13.57

3.1. Dynamic Score Partitions

Starting from the full dataset (`All`), we define four subsets by increasing the lower-bound threshold for inclusion: V_A ($S_{Dynamic} \geq 0.05$), V_B ($S_{Dynamic} \geq 0.1$), V_C ($S_{Dynamic} \geq 0.15$), and the highly-dynamic set V_D ($S_{Dynamic} \geq 0.5$).

3.2. Semantic and Physical Insights

Results in Tab. 5 demonstrate that training on subset $\mathbf{V_C}$ yields the optimal semantic trade-off, achieving the best FID (17.97) compared to the full dataset (`All`, 20.63). This suggests that aggressively filtering sequences with lower dynamic scores effectively removes static noise that dilutes model performance.

However, training exclusively on the lowest dynamic range tested, $\mathbf{V_D}$, results in a degraded FID (28.72) and significant physical instability (Jerk 113.35 in Tab. 6). This highlights that training solely on motion below the optimal dynamic threshold degrades motion smoothness and generation fidelity. It is important to note that while intermediate filtering (e.g., $\mathbf{V_C}$) improves numerical metrics, highly aggressive filtering inevitably discards subtle, fine-grained interactions that are essential for a truly comprehensive and diverse motion distribution.

4. Context-Aware Adaptive Filtering

To demonstrate the unique utility of our hierarchical taxonomy, we present an adaptive filtering approach that takes advantage of category-specific semantic context to substantially improve the quality of the dataset while preserving critical motion diversity. We focus on the **foot skating ratio** metric—a measure of foot sliding artifacts common in motion capture data.

Traditional **global filtering** removes all samples with high foot skating values across the entire dataset. For instance, removing the top 15% of foot skating ratio samples

Table 4. **Taxonomy-based physical evaluation.** Fine-grained breakdown of physical motion quality and artifacts comparing MDM and MoMask++ across the top 10 categories.

Category	Method	Foot Skating ↓	Jerk ↓	Ground Penetration ↓	Floating ↑	Accel. Peaks ↑	Dynamic Score ↑
Overall	MDM	0.0011	46.77	0.0e+0	0.0146	0.8374	0.2132
	MoMask++	0.0022	332.73	2.1e−5	0.0020	4.7223	0.3887
Communication	MDM	0.0009	32.51	0.0e+0	0.0217	0.4239	0.0959
	MoMask++	0.0027	217.11	1.8e−6	0.0043	2.5882	0.2010
Daily Life	MDM	0.0036	49.18	0.0e+0	0.0054	1.0127	0.3941
	MoMask++	0.0019	320.67	5.1e−6	0.0010	4.6117	0.5264
Dance	MDM	0.0004	72.88	0.0e+0	0.0037	1.9088	0.3031
	MoMask++	0.0005	510.93	6.4e−6	0.0004	8.3472	0.5850
Education	MDM	0.0001	30.52	0.0e+0	0.0252	0.4201	0.1055
	MoMask++	0.0009	192.41	3.7e−7	0.0051	2.3672	0.1986
Fitness	MDM	0.0000	60.89	0.0e+0	0.0073	1.1000	0.2447
	MoMask++	0.0003	432.76	1.6e−8	0.0008	6.1700	0.4781
Gestures	MDM	0.0003	33.17	0.0e+0	0.0323	0.4644	0.1048
	MoMask++	0.0013	223.13	1.4e−6	0.0062	2.8128	0.2033
Housework	MDM	0.0029	40.42	0.0e+0	0.0121	0.7514	0.1924
	MoMask++	0.0023	345.09	1.4e−6	0.0013	5.6696	0.4160
Outdoors	MDM	0.0030	51.65	0.0e+0	0.0089	1.1816	0.4572
	MoMask++	0.0072	341.00	1.1e−5	0.0016	5.0046	0.5940
Pet Care	MDM	0.0100	51.41	1.0e−4	0.0188	0.8560	0.3216
	MoMask++	0.0225	306.83	9.3e−6	0.0027	3.9456	0.4358
Sport	MDM	0.0006	62.34	0.0e+0	0.0075	1.3625	0.3428
	MoMask++	0.0008	451.61	5.5e−6	0.0007	7.0758	0.6317

Table 5. **Ablation study on motion dynamics.** Evaluation of MDM performance across different data partitions defined by motion dynamic ($S_{Dynamic}$). The subsets range from high-dynamics (V_A) to low-dynamics (V_D). Results show that selective training (V_C) yields better semantic alignment (FID) than using the complete dataset (All).

Method	R Precision			Diversity ↑	FID ↓	Matching Score ↑
	Top1	Top2	Top3			
All	0.5893	0.7652	0.8441	27.67	20.63	12.06
VA	0.5754	0.7564	0.8473	27.66	19.61	12.18
VB	0.5863	0.7521	0.8414	27.82	18.70	12.24
VC	0.5822	0.7578	0.8434	27.59	17.97	12.18
VD	0.4660	0.6400	0.7441	28.79	28.72	14.26

results in $123,178$ filtered samples. However, this indiscriminate approach fails to recognize that high foot skating values represent **authentic motion characteristics** in categories such as skateboarding, skiing, and snowboarding, while indicating quality issues in sedentary activities like office work or crafting.

Our adaptive filtering approach addresses this semantic ambiguity through the following steps:

1. **Category Identification:** We identify specific "expected skating" categories where foot skating is natural.
2. **Authenticity Preservation:** We preserve 100% of samples in these natural categories, regardless of their foot skating values.
3. **Contextual Filtering:** We apply **per-subcategory** 15% **filtering** to all other categories, removing only the worst samples relative to their specific contextual distribution.

This taxonomy-enabled methodology preserves $\mathbf{62,733}$ high-quality authentic samples that would have been incorrectly removed by global filtering, while simultaneously identifying $\mathbf{45,385}$ context-specific quality issues that global filtering missed. This result demonstrates how our taxonomy provides the essential semantic structure needed to distinguish between authentic motion and quality artifacts, enabling a robust balance between quality improvement and diversity preservation.

The same adaptive filtering paradigm can be universally

Table 6. **Impact of motion dynamics on physical quality.** Physical metrics for MDM trained on different dynamic score subsets.

Method	Foot Skating ↓	Jerk ↓	Ground Penetration ↓	Floating ↑	Acceleration Peaks ↑	Dynamic Score ↑
All	1.70e-03	43.98	0.00e+00	0.0167	0.8171	0.2138
VA	9.61e-04	47.69	5.99e-06	0.0159	0.8859	0.2311
VB	1.57e-03	45.13	3.09e-05	0.0161	0.8558	0.2120
VC	1.10e-03	46.77	0.00e+00	0.0146	0.8374	0.2132
VD	4.23e-04	113.35	6.94e-04	2.85e-03	1.37	0.2711

applied to other evaluation metrics:

- **Floating (vertical displacement):** Expected in jumping and acrobatic categories but signals errors elsewhere.
- **Ground Penetration:** Natural for digging and excavation activities but represents artifacts in most other contexts.

This flexibility demonstrates how the taxonomy enables researchers to design custom quality control pipelines tailored to downstream tasks. We leave the exploration and evaluation of advanced adaptive filtering strategies within text-to-motion generation for future work.

5. Implementation Details

In the main paper we demonstrated the performance of several SOTA architectures (diffusion and GPT) when trained on our dataset. For reproducibility, we provide the implementation details for the different models.

MDM Training Parameters We train a Motion Diffusion Model using a transformer-based denoising architecture with 8 layers, 4 attention heads, latent dimension 512, and feedforward dimension 1024. Text conditioning is provided through a frozen BERT encoder with maximum sequence length of 128 tokens. The diffusion process employs 50 denoising steps during training. Motion sequences are processed with a maximum length of 224 frames at 30 FPS, corresponding to approximately 7.5 seconds of motion. We train with batch size 256 for 750,000 iterations using exponential moving average (EMA) of model weights without weight decay regularization.

MoMask++ Training Parameters We train MoMask++ on our ITW dataset, with adaptations for our hardware and dataset characteristics. In Stage 1, we train a hierarchical residual VQ-VAE with 6 quantizers across 4 scales (8, 4, 2, 1) using a codebook size of 512 and code dimension of 512. The encoder-decoder architecture uses a width of 512, depth of 3, and temporal downsampling factor of 2. We train for 1000 epochs with batch size 512, learning rate 3e-4, and commitment loss weight 0.02. Critically, we disable gradient clipping to prevent codebook collapse, following recent findings in motion tokenization. In Stage 2, we train a bidirectional masked transformer with 8 layers, 6 attention heads, latent dimension 384, and feedforward dimension 1024. Text conditioning is provided via a frozen T5-base encoder (768-dim embeddings). We train for 500 epochs with batch size 256, applying linear learning rate scaling to 8e-4 (4× the base rate of 2e-4). Both stages use classifier-free guidance with 10% conditioning dropout and a 2000-step warmup schedule with multi-step learning rate decay at milestones [100k, 150k] with gamma 0.3. All experiments are conducted on a single NVIDIA A100-80GB GPU.

References

[1] Inwoo Hwang, Jian Wang, Bing Zhou, et al. Snapmogen: Human motion generation from expressive texts. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 1

[2] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 1