LLaMo: Scaling Pretrained Language Models for Unified Motion Understanding and Generation with Continuous Autoregressive Tokens

Supplementary Material

1. Implementation Details

In this section, we report all the details of LLaMo's implementation, to support reproducibility. We further trained an 8B model and discuss all our model sizes - 1B, 3B and 8B.

1.1. Motion VAE

We adopt the causal VAE architecture and training losses from MotionStreamer [15]. The first 1K training iterations use a linear warmup learning rate schedule from 0 to 5e-5, followed by 3M iterations with a cosine decay schedule from 5e-5 to 0. We use the AdamW optimizer [10] with [β_1, β_2] = [0.9, 0.95] and a batch size of 256. All VAE training runs use 8 A100 GPUs. For robustness in modeling continuous autoregressive tokens, we sample the variance for each VAE latent from $\mathbf{U}(0, \mathcal{C}_\sigma)$ where $\mathcal{C}_\sigma = 0.01$, instead of predicting it from the network.

1.2. Unified Motion-Language Model

Mixture-of-Transformer. We used Llama3.2-1/3B-Instruct and Llama3.1-8B-Instruct as the base language model to build LLaMo-1B/3B and 8B, respectively. During training, all language-related parameters are frozen, except for the text embeddings of [BOM] and [EOM]. These special text token embeddings are initialized from the mean of the language codebook. . The motion branch transformer parameters are initialized from the text branch transformer. The motion adapter $\mathcal{P}(\cdot)$ is a two-layer MLP with GELU [6] activation and post-RMSNorm [16].

Flow Matching Head. We use the MLP head architecture design from MAR [9], with a hidden dimension of 1536 and 12 layers. Before the output vectors of the Transformer serve as the condition for flow matching, we apply a two-layer MLP with GELU [6] activation and post-RMSNorm [16] as the motion projector. During inference, we use an ODE solver based on Euler integrator with 50 steps.

Motion Generation Exit Head. We follow TransformerTTS [8] and SpeechT5 [1]in using a simple MLP to predict the stop generation signal based on the decoder-only transformer output. The MLP is structured by 5 Linear layers with Swish activation [11]. We adopt a binary classification loss for stop prediction.

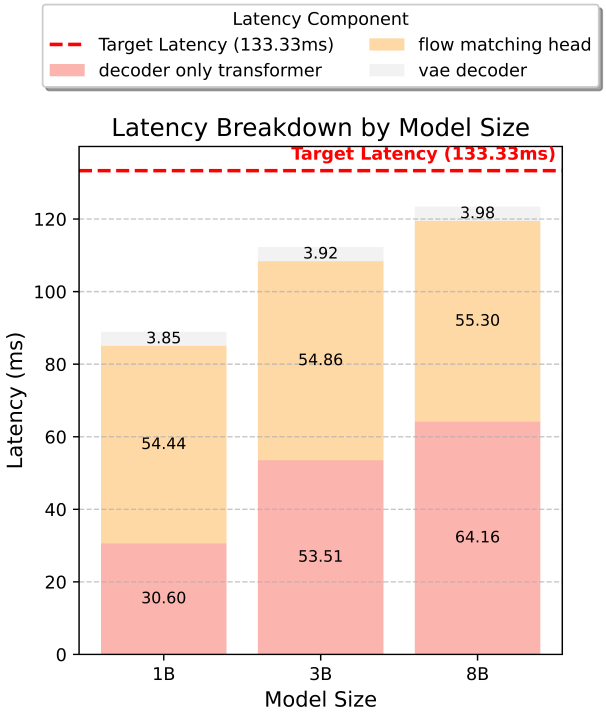


Figure 1. **Token Latency breakdown of Inference.** We compared the inference speed based on different model sizes. With infrastructural optimizations, even 8B model can achieve real-time streaming motion generation. **Our VAE does 4x temporal downsampling. So the 7.5FPS token generation speed equal to 30FPS motion generation speed.**

Efficient Training and Inference To achieve more efficient training, we utilize DeepSpeed-Zero2 [1] to reduce the redundancy in optimizer states and updating gradients. All the training is under BF16 precision via the AdamW optimizer [10] with [β_1, β_2] = [0.9, 0.95] and a batch size of 128. The LLaMo-1B and LLaMo-3B are trained on 8 A100s. The LLaMo-8B are trained on 16 A100s. To achieve real-time streaming motion generation on a single A100, we adopt serval infrastructural optimizations with few engineer efforts. Specifically, we use KV-cache to reduce the computation of large decoder-only transformer and compile the cuda graph of flow matching sampling loop to remove kernel launch overhead. We profile the cost of inference computation with batch size 4 in Fig. 1 and, as shown, all the models with different sizes can achieve real-time streaming motion generation. Since our motion causal VAE encodes human motion using a 4× temporal downsampling

[1] https://www.deepspeed.ai/tutorials/zero/

factor, the target token latency required to achieve real-time streaming motion generation is $1000/30 \times 4 = 133.33$ ms.

Methods	Motion-to-Text		Text-to-Motion	
	R@3↑	MM-D↓	FID↓	R@3↑
Real Motion	0.9866	0.7016	-	0.9866
LLaMo-1B (our)	0.9393	0.7136	27.361	0.9332
LLaMo-3B (our)	0.9422	0.7132	19.893	0.9594
LLaMo-8B (our)	0.9613	0.7126	18.935	0.9603
Ablations based on LLaMo-3B				
use [15]VAE	0.9221	0.7310	34.002	0.8936
only Stage1	0.9108	0.7443	80.912	0.7615
only Stage1&2	0.9422	0.7132	22.524	0.9521

Table 1. **Ablation of Design Choice.** We evaluate our models on the test split (∼ 30K samples) of our large-scale motion-text dataset. We follow the evaluator design in [5], text-to-motion protocols in [3], and motion-to-text protocols in [4]. We show that a) using a VAE with predicted variance significantly hurts motion generation, and b) our multistage training recipe progressively improves the model performance.

2. More Ablation Studies

In this section, we demonstrate more results and analysis to validate the effectiveness of LLaMo design choices.

Further Scale Up Model Size To explore the scalability of this solution, we design a 8B-version LLaMo based on Llama-3.1-8B-Instruct. Consistent with the findings in MotionMillion [2], as shown in Tab. 1, we observe that scaling the model from 3B to 8B yields negligible improvements in FID and R-precision compared to the transition from 1B to 3B.

Traditional VAE *vs*. Our VAE Although prior works [7, 12–14] have demonstrated that a robust VAE is crucial for continuous autoregressive generation, it remains unclear whether this conclusion generalizes to the motion modality. Therefore, we further evaluate the validity and generalization of this observation in our motion-language setting in Tab. 1. Instead of sampling the variance from predefined distribution, we use the classic network-predicted variance VAE as in [15]. The significant degradation of motion synthesis performance confirms that adding noise to ensure a robust VAE is essential for the continuous autoregressive paradigm. However, we note that motion understanding performance is not affected by the VAE robustness.

Multi-Stage Training Recipe. We further evaluate the effect of our multi-stage training strategy, with results summarized in Tab. 1 for the 3B model setting. Across stages, we observe steady improvements in both motion fidelity and text–motion consistency, indicating that the staged optimization procedure effectively stabilizes the learning dynamics of large models. By decomposing the training process into progressively more specialized phases, our approach mitigates early training instability, facilitates more reliable modality alignment, and ultimately leads to better overall zero-shot performance.

3. Data Curation Details

Annotation Prompt. We include the full Gemini-2.5-Pro prompt used for annotating the human-motion videos in the supplementary materials.

Data Filtering. During VLM-based annotation, we instruct Gemini to identify videos depicting static or near-static human motions. To further remove under-expressive sequences from a kinematic perspective, we apply an additional heuristic: a motion sequence is filtered out if the velocities of all end-effectors remain below 5 cm/s. This threshold corresponds to natural micro-movement during human quietly standing. Combining semantic–kinematic filtering, we ensures the dataset for motion head fine-tuning primarily contains expressive motion patterns.

4. Zero-shot Text-to-Motion Generation

In this section, we present additional results demonstrating the zero-shot motion generation capabilities of our large unified model.

User Study versus MotionMillion [2]. We conducted an A/B human evaluation study with 14 participants to assess motion generation quality across three dimensions: **Physical Plausibility**, **Motion Smoothness**, and **Text Alignment**. In this study, participants were shown motions generated by each model for the same text prompt, without knowing which model produced which motion. As shown in Fig. 2, our model achieves substantially better performance than the current state-of-the-art, MotionMillion [2], across all metrics. Leveraging high-fidelity continuous motion representations, LLaMo produces noticeably smoother and more physically plausible human motions compared with MotionMillion. Furthermore, our model demonstrates superior text–motion alignment, even though both MotionMillion and LLaMo employ comparable parameter budgets for text tokens (see Tab. 2). This highlights the effectiveness of our strategy around retaining strong native language capabilities in the underlying LLM while enabling high-quality motion generation.

Generalization to Unseen Languages. While studying the zero-shot capabilities of LLaMo, we came across an

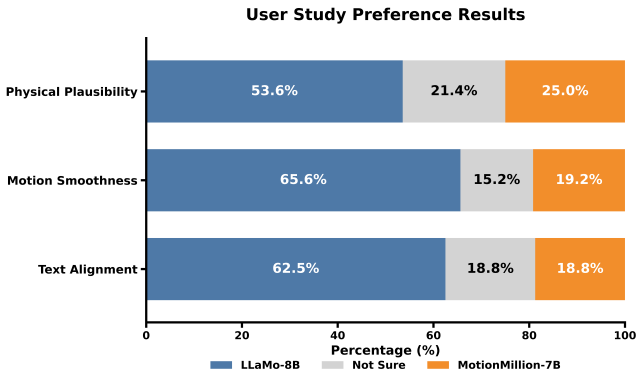


Figure 2. **User Study of Zero-shot Text-to-Motion Generation.** We use the prompts from MotionMillion-Eval [2] to evaluate our model against MotionMillion [2]. Results show that users significantly prefer our model across all the evaluation axes.

Methods	Motion Activated #Params	Text Activated #Params	Total #Params
MotionMillion-3B	3B	4.2B	4.2B
MotionMillion-7B	7B	8.2B	8.2B
LLaMo-1B	1B	1B	2B
LLaMo-3B	3B	3B	6B
LLaMo-8B	8B	8B	16B

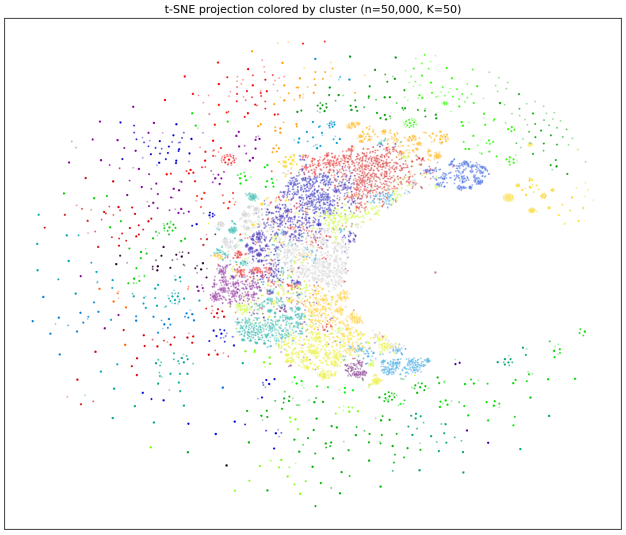
Table 2. **Parameters Comparison for Each Modality.** MotionMillion-7B has similar text token activated parameters with LLaMo-8B, which indicates similar language modeling capacity.

interesting emergent behavior: *We notice that LLaMo is able to generate motion from prompts in languages beyond English, even though our training data only had English language-motion data.* We highlight this intriguing emergent behavior as a qualitative observation by showing some examples in the supplementary. Please open the results webpage in our supplementary materials and allow 1-2 minutes for the webpage to load the videos. You can also click on the thumbnails / black tiles if they are not loaded.

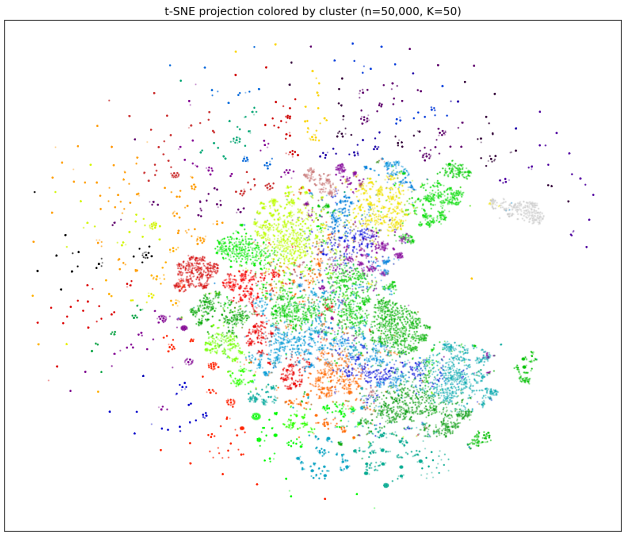
Dataset Comparison To demonstrate the semantic limitation about existing open-source large text-motion dataset, we visualize the text embedding distributions based on t-SNE. To extract the robust embeddings related to the human motion within the massive text annotation, we use Qwen3-Embedding-0.6B for sentence encoding. The instruction prompt we used is shown as followed.

```
Instruct: Given a text description
of human motion, encode the semantic
meaning of the described body
movement, action, and style\nQuery:
```

The t-SNE visualization in Fig. 3 reveals that MotionMillion exhibits pronounced inter-cluster gaps and notably smaller intra-cluster spread compared to our internal dataset. This suggests that, despite its large scale, MotionMillion suffers from limited semantic diversity: samples are concentrated around a narrow set of motion patterns, resulting in substantial semantic redundancy. In contrast, our internal dataset demonstrates more uniform coverage across the embedding space, indicating broader and more balanced motion knowledge.

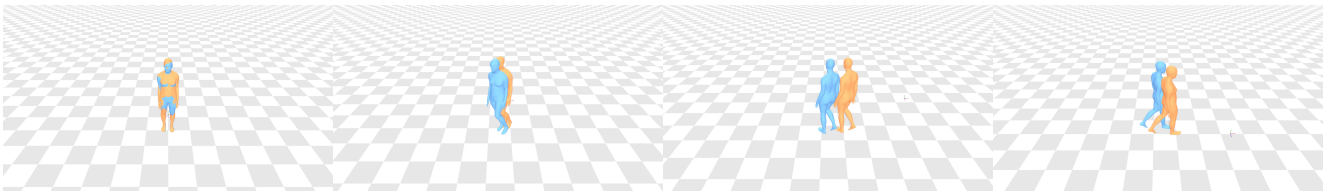


(a) MotionMillion text embedding cluster projection.

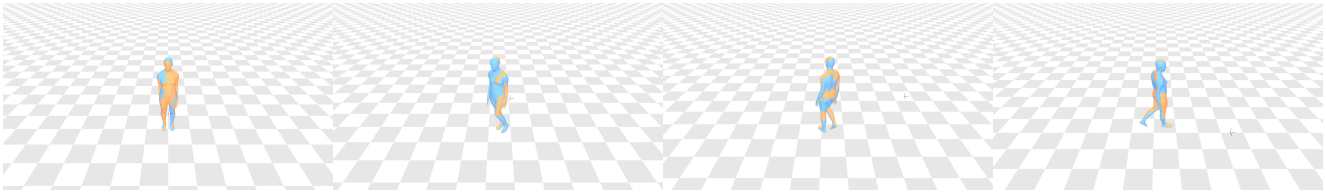


(b) Internal dataset text embedding cluster projection.

Figure 3. Semantic distribution visualization. We apply t-SNE to project the learned embeddings from both datasets into a shared 2D space. K-means clustering is performed independently on each dataset. For clarity, we subsample data points according to the density of each cluster.



(a) Motion Reconstruction by FSQ in MotionMillion.



(b) Motion Reconstruction by sigma causal TAE in LLaMo.

Figure 4. Qualitative Comparison about Motion Reconstruction. The blue motion is ground truth motion and the orange motion is the reconstructed motion.

5. Motion Auto-encoder Comparison

We present a qualitative comparison between the FSQ tokenizer adopted in MotionMillion [2] and the sigma-causal TAE employed in LLaMo. As shown in Fig. 4, the discrete tokenization in FSQ introduces inherent information loss during the quantization step, resulting in noticeable reconstruction artifacts such as jittery motion and loss of fine-grained details. In contrast, our continuous latent representation preserves higher fidelity to the original motion, enabling more accurate and temporally coherent reconstruction.

References

[1] Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 5723–5738, 2022. 1

[2] Ke Fan, Shunlin Lu, Minyue Dai, Runyi Yu, Lixing Xiao, Zhiyang Dou, Junting Dong, Lizhuang Ma, and Jingbo Wang. Go to zero: Towards zero-shot motion generation with million-scale data. *arXiv preprint arXiv:2507.07095*, 2025. 2, 3, 4

[3] 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*, pages 5152–5161, 2022. 2

[4] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pages 580–597. Springer, 2022. 2

[5] Chuan Guo, Inwoo Hwang, Jian Wang, and Bing Zhou. Snapmogen: Human motion generation from expressive texts. *arXiv preprint arXiv:2507.09122*, 2025. 2

[6] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 1

[7] Guolin Ke and Hui Xue. Hyperspherical latents improve continuous-token autoregressive generation. *arXiv preprint arXiv:2509.24335*, 2025. 2

[8] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural speech synthesis with transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6706–6713, 2019. 1

[9] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37:56424–56445, 2024. 1

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[11] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017. 1

[12] Chenze Shao, Darren Li, Fandong Meng, and Jie Zhou. Continuous autoregressive language models. *arXiv preprint arXiv:2510.27688*, 2025. 2

[13] Yutao Sun, Hangbo Bao, Wenhui Wang, Zhiliang Peng, Li Dong, Shaohan Huang, Jianyong Wang, and Furu Wei. Multimodal latent language modeling with next-token diffusion. *arXiv preprint arXiv:2412.08635*, 2024.

[14] NextStep Team, Chunrui Han, Guopeng Li, Jingwei Wu, Quan Sun, Yan Cai, Yuang Peng, Zheng Ge, Deyu Zhou, Haomiao Tang, et al. Nextstep-1: Toward autoregressive image generation with continuous tokens at scale. *arXiv preprint arXiv:2508.10711*, 2025. 2

[15] Lixing Xiao, Shunlin Lu, Huaijin Pi, Ke Fan, Liang Pan, Yueer Zhou, Ziyong Feng, Xiaowei Zhou, Sida Peng, and Jingbo Wang. Motionstreamer: Streaming motion generation via diffusion-based autoregressive model in causal latent space. *arXiv preprint arXiv:2503.15451*, 2025. 1, 2

[16] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019. 1