

Unveiling the Impact of Data and Model Scaling on High-Level Control for Humanoid Robots

Yuxi Wei^{1,2} Zirui Wang^{2,3} Kangning Yin^{1,2} Yue Hu⁴
Jingbo Wang^{2*} Siheng Chen^{1*}

¹ Shanghai Jiao Tong University ² Shanghai AI Laboratory ³ Zhejiang University ⁴ University of Michigan
wyx3590236732@sjtu.edu.cn
vfishc.github.io/schur

Abstract—Data scaling has long remained a critical bottleneck in robot learning. For humanoid robots, human videos and motion data are abundant and widely available, offering a free and large-scale data source. Besides, the semantics related to the motions enable modality alignment and high-level robot control learning. However, how to effectively mine raw video, extract robot-learnable representations, and leverage them for scalable learning remains an open problem. To address this, we introduce Humanoid-Union, a large-scale dataset generated through an autonomous pipeline, comprising over 260 hours of diverse, high-quality humanoid robot motion data with semantic annotations derived from human motion videos. The dataset can be further expanded via the same pipeline. Building on this data resource, we propose SCHUR, a scalable learning framework designed to explore the impact of large-scale data on high-level control in humanoid robots. Experimental results demonstrate that SCHUR achieves high robot motion generation quality and strong text-motion alignment under data and model scaling, with 37% reconstruction improvement under MPJPE and 25% alignment improvement under FID comparing with previous methods. Its effectiveness is further validated through deployment in real-world humanoid robot.

I. INTRODUCTION

Scaling up data and model size for scalable learning has led to significant breakthroughs across many research domains, including natural language processing, visual segmentation and detection, as well as image and video generation, and is now being extended to an increasing number of fields[1], [29], [7]. In the context of robot learning, scaling has recently attracted substantial attention[9]. However, for robots, directly collecting data is often complex, difficult, expensive, and time-consuming, which has motivated the search for broader, more automated, and low-cost data sources.

For humanoid robots, whose control and behaviors are highly similar to those of humans, human motion provides a valuable prior. Human motion datasets and related videos are comparatively easier to acquire at large scale, and their associated semantic information can be extracted through automated methods. This enables natural and efficient data expansion and scaling for humanoid robots by leveraging human motion as a foundation. Moreover, incorporating human motion priors facilitates more effective high-level control and cross-modal alignment for humanoid robots. Meanwhile, semantics play a central role as an intermediate modality and as one of the most critical forms of control

*Corresponding authors.

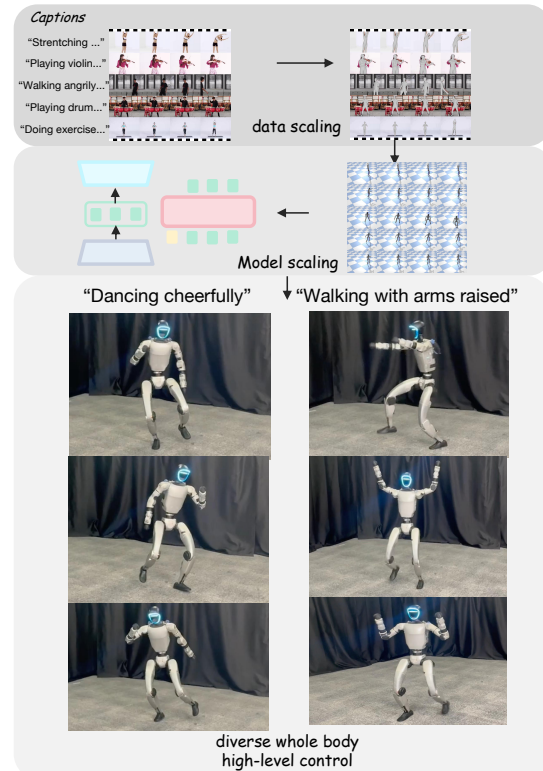


Fig. 1. We present the large-scale, high-quality robot motion dataset **Humanoid-Union** to achieve data scaling. Additionally, we propose a text-based robot motion control method, **SCHUR**, which enables effective scalable learning. By scaling both data and models, our approach can generate complex, diverse and high-quality whole-body robot motions from text, which are executed by a whole-body tracker and successfully deployed on real-world robots.

signals. Achieving robust semantic alignment under scaling is essential for complete high-level control and serves as a foundation for advancing toward diverse control paradigms, including vision-language-action models[2].

Using human motion as prior for high-level control of humanoid robots has been explored recently [27], [21], [22]. High-level control means generate control signal for low-level executor by conditions like texts. However, existing studies have shown several shortcomings. First, [27], [2] still rely on human motion as a direct prior, yet its distribution differs from robot motion. Although retargeting can transfer human to robot motion, it complicates usage and alters the distribution, thereby limiting modality alignment. Second,

parts of human motion data may not be executable by low-level executors, reducing data quality and limiting execution effectiveness. Third, existing works are limited in a small data scale. The commonly used dataset HumanML3D[4] contains only less than 30 hours of data. They fail to fully utilize the abundant data available currently, lacked corresponding model scaling, and have not fully tapped into the potential of human motion data for humanoid robots.

Therefore, it is essential to generate high-quality and executable robot motion not just human motion while fully leveraging the potential of large-scale human motion data. By expanding the data scale and correspondingly scaling the models, we can achieve more effective control capabilities and improved generalization. Some studies have made preliminary explorations [15], but significant issues remain. First, it lacks investigation into robot motion representation, relying only on root information and DoFs without introducing more effective priors such as keypoints in Euclidean space for additional regularization. Second, it overlooks the advantages of directly generating robot motion and neglects to refine or filter robot motion data, leaving unexecutable or noisy segments unaddressed. Third, the exploration of scalable learning is shallow, with no effective methods or study of model scaling. Finally, they do not demonstrate lower-limb motions on real robots, instead focusing on arm movements, resulting in limited expressiveness.

To fully leverage large-scale human motion priors for humanoid robots and to investigate robot motion representations along with high-level scalable learning, it is crucial to construct a large and high-quality robot motion dataset automatically derived from human videos and motion, together with effective scalable learning strategy. To this end, we introduce the Humanoid-Union dataset and propose SCHUR, a scalable large text-based robot motion generation model.

The Humanoid-Union dataset comprises over 260 hours of source motion data, featuring an automatic pipeline and post-process for data scaling. Human-centric videos from various sources are converted into human motion through video motion capture techniques, while existing high-quality human motion data is also integrated and retargeted to obtain robot motion. Due to the extensive and varied sources of human motion data, which include interactions with scenes, variations in motion amplitude, and potential occlusions or noise, the robot motion is subsequently filtered and processed using a pretrained universal whole-body motion tracker. This ensures that only high-quality data, compliant with robotic kinematics and physical constraints, is retained at the robot motion level. Additionally, the descriptions corresponding to the motions are automatically annotated using a vision-language model, such as GPT-4V [1], enabling the construction of text-robot motion pairs for semantic alignment training and text generation to achieve high-level control.

To fully exploit the Humanoid-Union dataset, we propose SCHUR, an effective scalable learning model designed for high-quality text-based robot motion generation. SCHUR optimizes robot motion representation, and adopts an effective quantization method and related framework with proper

training strategy. Although the root position, root orientation, and degrees of freedom (DoFs) can comprehensively represent the robot's state, we find that relying solely on these representations does not yield well generation results. Therefore, we further incorporate the positions and orientations of bound virtual keypoints calculated through forward kinematics (FK) into the representation. The redundant information provided by these virtual keypoints is similar to human motion style and effectively enhances the generation quality. SCHUR adopts a two-stage approach: first, robot motion is tokenized, and then a LLaMA-based architecture is utilized for motion generation using text prefix. In the tokenization stage, we apply Finite Scalar Quantization (FSQ) VAE [16], which includes a more effective quantization method that significantly improves tokenization results compared to VQ-VAE and is less prone to collapse, thereby supporting larger codebook sizes. In the autoregressive generation stage, a prefix bidirectional attention mask is employed to generate the corresponding motion tokens, which are decoded to obtain the generated motion.

Through experiments, we validate the tokenization quality and scaling properties of SCHUR on the Humanoid-Union dataset. As the codebook size increases, there is a continuous improvement in the quality of tokenization, with FSQ demonstrating significantly better scaling capabilities compared to VQ. Additionally, in the second stage of the generator, an increase in model parameters also exhibited markedly improved modality alignment and generation quality. We further confirm the impact of high-quality data and advantage of directly generating robot motion on the success rate and tracking error of the low-level tracker execution, utilizing the tracker deployed on a real-world robot to verify the effectiveness in real-world robot scenarios.

II. RELATED WORKS

A. Human Motion Generation

Human motion generation has been extensively explored. Transformer-based [30], [3] and diffusion-based [23], [20], [26] methods have been proposed to achieve text-to-human-motion generation. Compared with robot motion data, human motion data are easier to obtain through various sources and benefit from numerous existing datasets [4], [13]. However, directly using human motion data for high-level control of humanoid robots reduces modality alignment and complicates the control pipeline. Moreover, robot motion is subject to additional kinematic and physical constraints for deployment in the real world, making some human motion data unsuitable as priors for humanoid robots. In this work, training is performed directly on robot motion data, which are filtered and processed using a universal whole-body tracker. This procedure removes low-quality data while providing physical and kinematic priors, ensuring more effective and realistic learning for humanoid robot control.

B. Humanoid Robot Learning

Extensive research has focused on low-level control policies for humanoid robots. Prior work has explored full-body

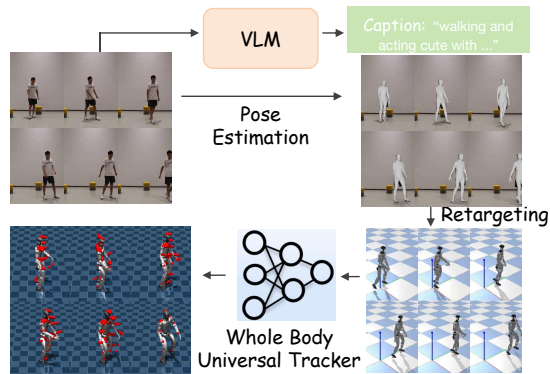


Fig. 2. The data pipeline of **Humanoid-Union**. Pose data is extracted from videos by pose estimation, resulting in the SMPL representation. Corresponding descriptions are generated using vision-language model. The SMPL data is then retargeted to obtain the corresponding robot motion data, which is subsequently filtered and post-processed using a trained universal whole-body tracker.

tracking, achieved loco-manipulation, and implemented text-based control via reinforcement learning (RL) [6], [5], [21], [27], [28]. Nevertheless, RL-based low-level policies are difficult to scale with large datasets due to the constraints of lightweight policy model. Approaches that rely on human motion for robot require additional retargeting, which reduces modality alignment, complicates usage processes, and still fails to fully exploit data scalability. Methods using robot motion data often include infeasible motions and lack effective scalable learning strategies. In contrast, the Humanoid-Union dataset introduced in this work provides high-quality, tracker-processed robot motion data and SCHUR supports a more efficient approach to scalable learning.

III. HUMANOID-UNION DATASET

A. Overview

Diverse and abundant human video and motion data are suitable sources for humanoid robots. However, directly applying human motion introduces large portions of data unsuitable for robots due to scene interactions, joint constraints, and other factors. Moreover, using human motion for high-level control complicates the process and limits modality alignment. To bridge this gap, we require a high-quality, large-scale robot motion dataset executable by robots. Furthermore, to enable high-level control, we provide textual descriptions for each motion, aligning robot motion with text as an essential and general intermediate modality.

We construct the Humanoid-Union dataset, which includes various sources [10], [8], [11] and some self-collected data, resulting in about 260 hours of robot motion data, 170,000 sequences. We create the dataset with 3 steps: i) **video motion capture and description**, see sec III-B. We extract human motion [10] from a vast array of human-centric videos and employ a vision-language model to perform semantic labeling of the sequences, while also integrating existing publicly available human motion data with semantic annotations. ii) **retargeting**, see sec III-C, human motion data is then retargeted to generate robot motion that conforms

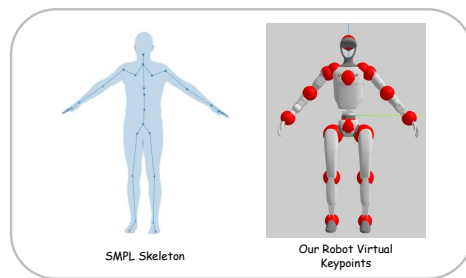


Fig. 3. Comparison between the commonly used SMPL skeleton and our manually bound keypoints reveals that the virtual keypoints we defined exhibit a topology that is closer to SMPL. This alignment allows for a more accurate representation, maintaining a style that is more consistent with human motion.

to kinematic constraints. iii) **whole body tracker filtering and postprocess**, see sec III-D, considering the wide-ranging sources of human-centric data, which may include a significant number of actions involving scenes and object interactions, as well as certain motions that may exceed the current capabilities of robots and the inherent noise in the data processing, we utilize a pretrained universal whole-body tracker for post-processing. Motions that fail during tracker execution are filtered out, and the motions processed by the tracker are used to complete the final post-processing. The data pipeline is shown in Fig.2

B. Video MoCap and Description

Here, we extract human motion from videos and generates the corresponding textual descriptions with 3 steps. First, we detect and track human in the video sequences. Second, we perform pose estimation and refinement for each human motion sequence. Note that these two steps follow the standard procedure in [10], and the output unified human motion data is represented in the SMPL model format [17], including the SMPL rotation parameters θ_{human} and the position $r_{\text{human}}^{\text{pos}}$. Third, we annotate the human motion data with additional textual descriptions via vision-language model [1]. These three steps result in paired multi-modality data, enabling high-level control learning.

C. Motion Retargeting

Motion retargeting focuses on transforming human motion data into robot states. We bind some virtual keypoints to the robot by some fixed topology relationship that mimics the structure of the human skeleton, improving the alignment between human and robot topologies. The retargeting is achieved by mapping human motion data to the robot states including the virtual keypoints. The virtual keypoints also preserve properties such as symmetry, making the resulting motions more human-like in style and also benefits the following model learning as part of representation. Note that, these points are not in a one-to-one correspondence with the robot's physical parts but instead satisfy specific relative relationships. The virtual keypoints is shown in Fig.3

Specifically, motion retargeting is achieved with 4 steps: i) bind virtual keypoints for the robot by some artificial definition of relationship with body parts; ii) recalibrate the

scale of the body parts using a fixed T-pose; iii) based on the inverse kinematics (IK) process, we solve for the robot’s joint positions j_{robot} ; iv) and then employ forward kinematics (FK) to compute the virtual keypoints’ positions k^{pos} and rotations k^{ori} . The height of the robot motion is corrected using a strategy similar to PHC [12], followed by a smoothing post-processing step.

The final results include the robot’s root position $r_{robot}^{pos} \in \mathbb{R}^3$, root orientation $r_{robot}^{ori} \in \mathbb{R}^3$, joint positions for all DoFs $j_{robot} \in \mathbb{R}^d$, and the positions $k^{pos} \in \mathbb{R}^{n \times 3}$ and rotations $k^{ori} \in \mathbb{R}^{n \times 3}$ of virtual keypoints. Here, $d = 29$ is the dimensionality of the DoFs, and $n = 17$ is the number of virtual keypoints. r_{robot}^{ori} and k^{ori} are represented with roll-pitch-yaw.

D. Whole Body Tracker and Postprocessing

We train a whole-body tracker for both advanced data processing and final low-level motion execution. The tracker is trained on the almost full AMASS dataset [13]. After extensive training, the resulting universal tracker is capable of directly imitating whole-body motions, including lower-limb movements. The training pipeline resembles that of ExBody2 [6], where the retargeted motion results, including the virtual keypoints, are used as observations. This tracker demonstrates strong generalization and a high success rate in execution. It is employed to further filter and refine robot motion data by leveraging the execution constraints, discarding sequences that are evidently infeasible while preserving motions better suited for robotic execution. About 10% of motions are filtered. In addition, the tracker introduces physics-informed priors into the motion representations. Finally, we deploy the same tracker in our real-world robot experiments, validating its effectiveness.

IV. SCHUR FOR SCALABLE ROBOT MOTION GENERATION

A. Overview

With a large-scale, high-quality robot motion dataset, we focus on effectively leveraging it by aligning robot motion with semantic descriptions. Text, as a key intermediate modality, enables high-level control, and this alignment underpins the Vision-Language-Action (VLA) model. Building on the Humanoid-Union dataset, we propose **SCHUR** (Scalable **H**umanoid **R**obot model), which scales to achieve text-based high-level control of humanoid robots. SCHUR draws on successful experiences from computer vision [19], employing a paradigm of discrete tokenization and autoregressive generation. It consists of two stages. In the first stage, the representation of the motion is tokenized to obtain a compact and structured discrete representation. We improve the representation of robot motion itself, effectively enhancing the quality of the tokenizer, and apply FSQ, a more effective quantization method, which further improves quality while enhancing the model’s scaling capability, reducing the likelihood of collapse. In the second stage, we adopt a decoder-only autoregressive generation approach based on a LLaMA [25] structure, using the tokens from text as prefixes

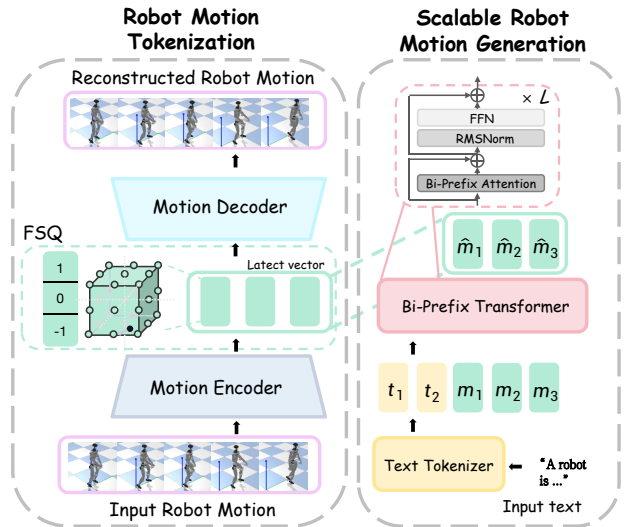


Fig. 4. The framework of **SCHUR** consists of two stages. In the first stage, robot motion is tokenized using an effective representation, with FSQ employed for quantization. Here, we present an example using $L=3$ to illustrate FSQ. The second stage utilizes text tokens as prefix and applies prefix-bidirectional attention to generate motion tokens in an autoregressive manner.

to autoregressively generate the robot motion tokens from the first stage. The overview of SCHUR is in Fig.4

B. Robot Motion Representation

Robot motion itself can be accurately and completely represented through the root position r_{robot}^{pos} , root orientation r_{robot}^{ori} , and degrees of freedom (DoFs) j_{robot} , which is the approach taken in [15]. However, we find that such a representation does not yield sufficiently effective tokenization or satisfactory generation results, the related results can be found in Tab. I. To address this, we introduce a more enriched representation. During the retargeting process, the virtual keypoints we bind can effectively restore the human motion style and additionally provide a representation of the motion within the 3D Euclidean space. We combine the positions k^{pos} and orientations k^{ori} of the virtual keypoints into the representation, resulting in a final representation $e = \text{concat}(r_{robot}^{pos}, r_{robot}^{ori}, j_{robot}, k^{pos}, k^{ori})$. Experimental results demonstrate that this redundant representation of robot motion is highly effective for tokenization.

C. Scalable and Efficient Motion Tokenizer

The preliminary representation of robot motion must be tokenized into discrete tokens for subsequent generation stage. Effective tokenization also facilitates broader applications in modality alignment and high-level control, such as within VLA frameworks. Previous approaches typically adopt VQ-VAE for this stage, and discrete tokenization has also achieved success in many generative tasks. However, as the scale of data increases, VQ-VAE requires a correspondingly larger codebook to capture the expanded distribution. Due to the quantization mechanism of conventional VQ, particularly the backpropagation process, such methods are prone to model collapse, making it difficult to achieve

TABLE I
ABLATION OF TOKENIZER

Metric	Human Motion	20% Data	Naive Repre	Ours
MPJPE	0.0389	0.0483	0.0406	0.0326
MPKPE	0.0401	0.0497	0.0454	0.0372
L1 Loss	0.0413	0.0512	0.0463	0.0380

sustainable scaling. To overcome these limitations, we adopt Finite Scalar Quantization (FSQ), a more robust quantization strategy that mitigates collapse while enabling more effective tokenization and continued scaling.

Conventional VQ-VAE. VQ-VAE can be employed to learn a discretized representation of robot motion, following the standard encoder–decoder paradigm. The input motion e is first encoded into a latent vector z , which is then quantized by the operator Q to obtain a discrete latent vector \hat{z} . Finally, \hat{z} is decoded back into the reconstructed robot motion \hat{e} .

In the conventional VQ-VAE formulation, a codebook $C = \{c_i\}_{i=1}^S$ with size S is learned for quantization. The quantization process can be expressed as:

$$\hat{z} = Q(z; C) = \arg \min_{c_k} \|z - c_k\|_2^2.$$

The optimization simultaneously considers the reconstruction quality of the decoder and the learning of the codebook. The overall loss function is written as:

$$\mathcal{L} = \|e - \text{Dec}(Q(z; C))\|_2^2 + \alpha \|z - \text{sg}(\hat{z})\|_2^2,$$

where Dec denotes the decoder, α is a weight parameter, and $\text{sg}(\cdot)$ indicates the stop-gradient operation.

To stabilize training, techniques such as exponential moving average (EMA) and codebook reset, originally proposed for VQ-VAE in human motion modeling [30], are also adopted in our framework. However, in order to handle larger-scale datasets and achieve higher compression quality, it is necessary to increase the codebook size. A common issue with conventional VQ-VAE is codebook collapse when the codebook becomes large, since the $\arg \min$ operation in VQ often tends to select only a subset of codes for frequent updates, leaving other codes underutilized. To address this, we adopt a more effective quantization strategy, named FSQ.

Finite Scalar Quantization(FSQ). To address the potential collapse caused by the $\arg \min$ operation, we adopt FSQ, a more effective quantization method that avoids using $\arg \min$. In FSQ, the quantization process can be written as:

$$\hat{z} = Q(z) = \text{round}(f(z)),$$

where $f(\cdot)$ is a bounding function, which in our implementation is set as the sigmoid function. Unlike VQ, FSQ does not involve the $\arg \min$ operation, thus alleviating the collapse problem inherent in VQ.

FSQ essentially quantizes each element of the latent vector into one of L unique integers, and the number of different combinations and the codebook size can be expressed as: $|C| = \prod_{i=1}^d L_i$, where d is the dimensionality of the latent vector. Since the rounding operation in the quantization

TABLE II
COMPARISON AND ABLATION OF TEXT-BASED GENERATION

Metric	Human Motion	20% Data	MDM	Ours
FID	16.9	23.5	20.8	12.6
R@1	0.688	0.653	0.667	0.719
R@2	0.717	0.689	0.692	0.754
R@3	0.755	0.707	0.713	0.812

TABLE III
TRACKING PERFORMANCE UNDER DIFFERENT HIGH-LEVEL CONTROL

Metric	Dataset	Human Motion	W. raw data	Ours
Success Rate	0.911	0.774	0.762	0.907
MPJPE	0.0874	0.1183	0.1095	0.0893
MPKPE	0.0867	0.1112	0.1028	0.0878

process is non-differentiable, we also employ a similar stop-gradient technique.

The optimization in FSQ does not require additional regularization or optimization strategies; it directly optimizes the reconstruction error from the decoder. Therefore, the final loss function is written as:

$$\mathcal{L} = \|e - \text{Dec}(f(z) + \text{sg}(\text{round}(f(z)) - f(z)))\|_2^2.$$

FSQ does not require modifications to the encoder or decoder, only changes to the quantization process itself. As a result, FSQ achieves significantly better reconstruction quality and scaling properties.

D. Autoregressive Text-Based Generation

Using the high-quality robot motion tokens obtained through FSQ, we aim to align with the text modality, facilitating the generation of robot motion from text. The overall generation process adopts a decoder-only transformer architecture similar to LLaMA. The robot motion tokens are already acquired from the previous stage, while the text tokens are obtained from a pre-trained T5-XL large language model [18], which provides word-level tokens as the sequence prefix. These text tokens are then used for autoregressive generation of robot motion tokens.

It is worth noting that since text tokens do not require prediction, we modify the attention mask to enable prefix-bidirectional attention, in contrast to the conventional causal attention. This approach allows bidirectional attention over the text tokens while performing autoregressive prediction over the motion tokens. Additionally, RMSNorm is applied within the transformer blocks to improve training stability.

The final optimized loss can be written as:

$$\mathcal{L} = - \sum_{t=1}^n \log p(\hat{m}_t | m_{<t}, T),$$

where T represents the text token series and m denotes the motion token.

V. EXPERIMENT

A. Experiment setup

All our experiments are conducted on the 29-DoF version of the Unitree G1. In the first stage, the tokenizer employs convolutional residual blocks for both the encoder and the

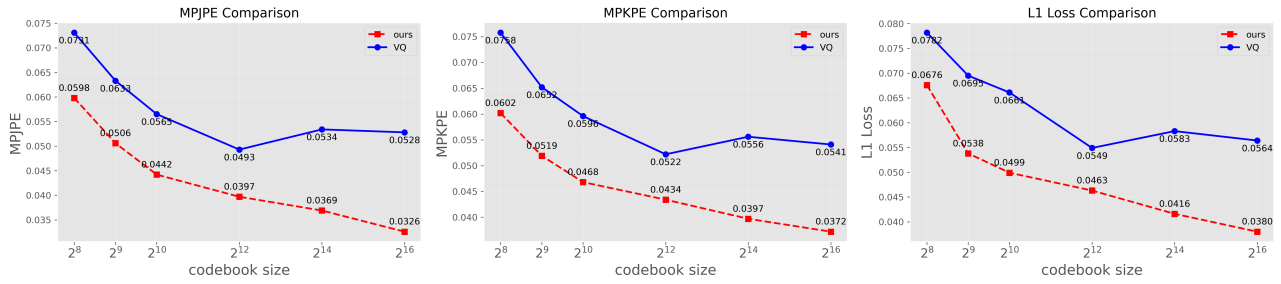


Fig. 5. Comparison of SCHUR’s tokenization stage with conventional VQ-VAE under different codebook sizes, using metrics such as MPJPE, MPKPE, and L1 loss. With the introduction of FSQ, SCHUR significantly outperforms conventional VQ-VAE in terms of reconstruction quality. Furthermore, it demonstrates strong scaling properties, with performance continuously improving as the codebook size increases. In contrast, VQ-VAE experiences instability during training with larger codebook sizes and eventually suffers from collapse.

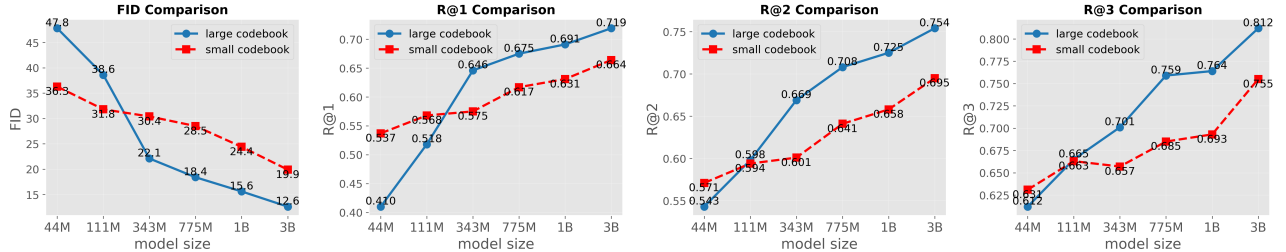


Fig. 6. In the text-based generation stage, the FID, R@1, R@2, and R@3 metrics are compared under small (256) and large (65536) codebook sizes, with the model’s parameter size increasing. As the model size grows, SCHUR is able to better leverage the data, improving its alignment with the text modality across all metrics. Additionally, the larger codebook size, due to its superior tokenization performance, generally leads to better generation results. However, when the model parameters are small, the larger codebook size struggles to learn effectively, resulting in poorer performance.

decoder. In the second stage, for text-based autoregressive generation, we adopt an architecture that is nearly identical to LLaMA, with model parameter scales aligned accordingly. Our tracker is fully trained on the AMASS dataset with IsaacGym [14] as simulator, following data process similar to PHC [12]. The tracker is universal in design and does not undergo additional fine-tuning for specific motion. MuJoCo [24] is used for evaluation. Out of the 29 degrees of freedom, 6 are fixed, leaving 23 actively controlled in practice. Detailed configurations for each component are presented in their respective sections.

B. Model scaling of tokenizer

We conduct scaling experiments on the tokenizer under varying codebook sizes and compared its performance against a conventional VQ-VAE baseline, as shown in Fig.5. The evaluation metrics include MPJPE (Mean Per Joint Position Error), MPKPE (Mean Per Keypoints Position Error), and L1 loss. In our experiments, the Humanoid-Union dataset was partitioned into training, test, and validation sets with an 80%, 15%, and 5% split, respectively, and results are reported on the test set. From the results, we observe that our method, when incorporating FSQ, consistently outperforms the conventional VQ approach across all reconstruction metrics. This provides strong evidence that our tokenizer is more effective, achieving superior performance even under relatively small codebook sizes.

In addition, we observe that as the codebook size increases, conventional VQ exhibits an obvious collapse effect, failing to further improve reconstruction performance. In contrast, our method continues to achieve stable gains in

reconstruction quality with larger codebook sizes, demonstrating its superior scalability. To further validate the hypothesis proposed in our paper, we evaluated the codebook usage of different methods, defined as the proportion of the codebook utilized on the test set. Our method consistently achieves a usage rate above 99% across all settings, whereas conventional VQ shows a gradual decline to below 90% as the codebook size grows. Overall, the tokenization in SCHUR is shown to be highly effective and to possess strong scaling properties.

C. Ablation of tokenizer

We further conducted an ablation study on the tokenizer under the largest codebook size, comparing several settings: (i) directly training on human motion and subsequently applying retargeting to robot motion for evaluation, (ii) training with 20% of full dataset and (iii) training with representations that only include root and DoF information, without incorporating virtual keypoints, where metric computation is performed after the FK process. As shown in Tab.I direct training on human motion fails to achieve optimal performance due to distributional discrepancies and also complicates the practical deployment pipeline. Moreover, when the training data is limited, the performance degrades significantly, highlighting the importance of data scaling in the tokenization process. Finally, the inclusion of virtual keypoints in the representation substantially improves the reconstruction quality of the tokenizer.

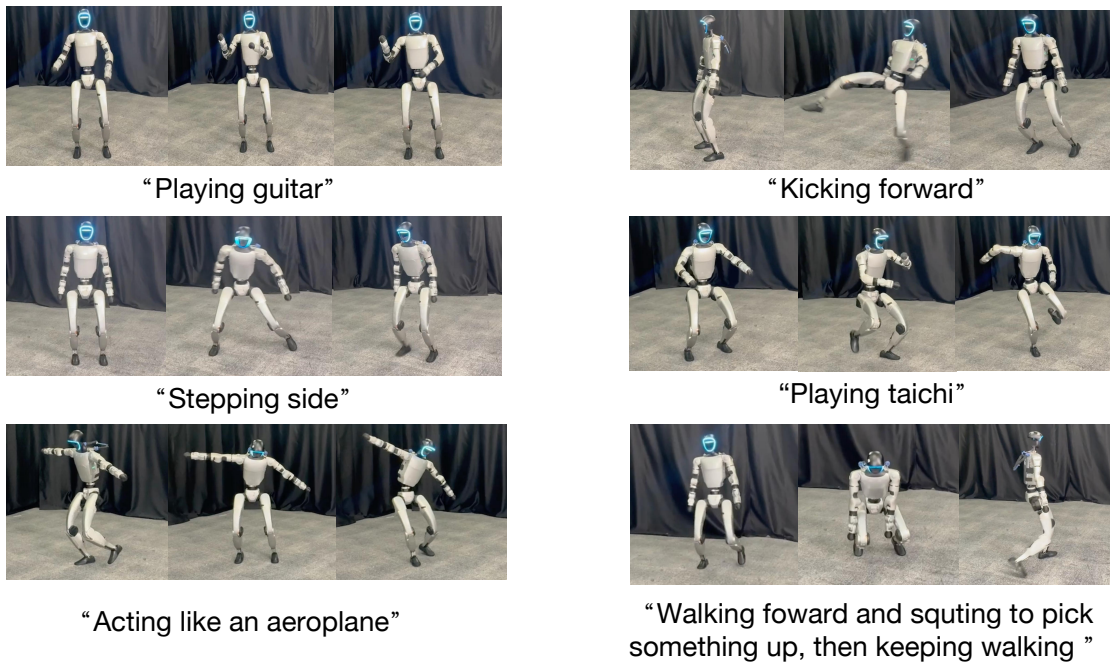


Fig. 7. The real-world deployment results across diverse textual inputs. SCHUR is capable of handling a wide range of requirements, including upper- and lower-limb movements, locomotion, abstract semantic control, and complex long-horizon control. It generates the corresponding whole-body information as control signals, which are then executed with high precision and stability through a universal whole body tracker, thereby achieving smooth and reliable real-world deployment performance. See more results in supplementary video.

D. Model scaling of text-based generation

We further evaluate the scaling characteristic of the generation stage under different codebook sizes (256 and 65,536) in the first stage, as shown in Fig.6. The dataset partitioning strategy is kept consistent with that used for the tokenizer. For evaluation, we employ four common metrics for generation: FID, R@1, R@2, and R@3. Following a procedure from [26], we train an evaluator on our own dataset. The results indicate that as the model size increases, all metrics consistently improve, suggesting that larger models achieve better alignment between text and generated motion across all criteria, thereby validating the scaling capability of the generation stage. However, with smaller codebook sizes in the first stage, the relatively weaker reconstruction quality limits the ability to fully capture the data distribution, resulting in overall poorer generation quality. Interestingly, when the second-stage model size is very small, smaller codebook sizes sometimes yield relatively better results, as the limited model capacity prevents effective learning from large-scale data. Overall, these results demonstrate that SCHUR is able to effectively exploit large-scale data through scaling, achieving high-quality generative performance.

E. Comparison and ablation of text-based generation

We conduct comparisons under the largest model configuration between three settings: (i) generating human motion and retargeting it to robot motion, (ii) training with only 20% of the full data, and (iii) generation using MDM [23], a diffusion-based approach. As shown in Tab.II, consistent with tokenizer experiments, directly generating human motion and subsequently applying retargeting is constrained by

distributional distribution, which limit performance. Likewise, insufficient training data prevents the model from fully exploiting its capacity. Moreover, compared with MDM, SCHUR achieves superior generative performance, further demonstrating the effectiveness of our approach.

F. Tracker performance under different high-level control

To further verify the generation quality, we evaluate whether the universal tracker can reliably execute the generated motions, as shown in Tab.III. We adopt success rate, MPJPE, and MPKPE as evaluation metrics. As a reference, we first validate the tracker’s performance on full dataset. We then compare two alternative settings: (i) generating human motion followed by retargeting, and (ii) training SCHUR directly on raw robot motion data without our tracking-based filtering and post-processing.

The results show that in both alternative cases, the tracker’s success rate drops significantly. This is due to the presence of a large amount of infeasible or physically inconsistent motions in the training data, which contaminate the learning process. The increased execution difficulty also leads to degraded tracking performance. By contrast, our approach leverages higher-quality data, where unreasonable or overly difficult motions have been filtered out, resulting in a distribution more compatible with the tracker’s capabilities.

G. Real-world deployment

We further demonstrate the generated results on a real-world robot in Fig.7, with additional results provided in the supplementary video. In combination with our tracker for deployment, the generative model is capable of producing motions that can be reliably executed on the real-world robot.

These include, but are not limited to, directly controlled upper- and lower-limb actions, locomotion behaviors, as well as non-directly controlled motions with clearly abstract semantics. Moreover, SCHUR can compose multiple atomic actions into longer sequences, enabling versatile behavior generation. Our final deployment achieves highly precise and stable control, maintaining balance and motion stability even under whole-body tracking, while successfully executing a wide range of locomotion tasks.

H. Conclusion

We propose Humanoid-Union, a large-scale and diverse robot motion dataset with semantic information for modality alignment and control. This dataset provides high-quality robot motion, filtered through a universal whole-body tracker to remove unsuitable motions. To leverage this data, we introduce SCHUR, a scalable and effective method for high-quality, text-based robot motion generation. SCHUR combines FSQ-based tokenization and autoregressive generation with a text prefix, and adopts improved robot motion representation. Experiments demonstrate SCHUR's superior performance, robust scaling capabilities, and the value of large-scale data and direct robot motion usage. Real-world deployment on real-world robot further validates its generation and control capabilities. As a limitation, we explore the direct control by robot states and actions based on text without low-level tracker. But large models face challenges in real-time deployment, and the lack of balance priors makes locomotion control difficult, which is left as future work .

REFERENCES

- [1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [2] P. Ding, J. Ma, X. Tong, B. Zou, X. Luo, Y. Fan, T. Wang, H. Lu, P. Mo, J. Liu, *et al.*, "Humanoid-vla: Towards universal humanoid control with visual integration," *arXiv preprint arXiv:2502.14795*, 2025.
- [3] C. Guo, Y. Mu, M. G. Javed, S. Wang, and L. Cheng, "Momask: Generative masked modeling of 3d human motions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1900–1910.
- [4] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng, "Generating diverse and natural 3d human motions from text," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5152–5161.
- [5] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. Kitani, C. Liu, and G. Shi, "Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning," *arXiv preprint arXiv:2406.08858*, 2024.
- [6] M. Ji, X. Peng, F. Liu, J. Li, G. Yang, X. Cheng, and X. Wang, "Exbody2: Advanced expressive humanoid whole-body control," *arXiv preprint arXiv:2412.13196*, 2024.
- [7] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.
- [8] Y. Liao, Y. Fu, Z. Cheng, and J. Wang, "Animationgpt: an aigc tool for generating game combat motion assets," 2024.
- [9] F. Lin, Y. Hu, P. Sheng, C. Wen, J. You, and Y. Gao, "Data scaling laws in imitation learning for robotic manipulation," *arXiv preprint arXiv:2410.18647*, 2024.
- [10] J. Lin, A. Zeng, S. Lu, Y. Cai, R. Zhang, H. Wang, and L. Zhang, "Motion-x: A large-scale 3d expressive whole-body human motion dataset," *Advances in Neural Information Processing Systems*, vol. 36, pp. 25 268–25 280, 2023.
- [11] S. Lu, J. Wang, Z. Lu, L.-H. Chen, W. Dai, J. Dong, Z. Dou, B. Dai, and R. Zhang, "Scamo: Exploring the scaling law in autoregressive motion generation model," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 27 872–27 882.
- [12] Z. Luo, J. Cao, K. Kitani, W. Xu, *et al.*, "Perpetual humanoid control for real-time simulated avatars," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 895–10 904.
- [13] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "Amass: Archive of motion capture as surface shapes," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5442–5451.
- [14] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, *et al.*, "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv preprint arXiv:2108.10470*, 2021.
- [15] J. Mao, S. Zhao, S. Song, T. Shi, J. Ye, M. Zhang, H. Geng, J. Malik, V. Guizilini, and Y. Wang, "Learning from massive human videos for universal humanoid pose control," *arXiv preprint arXiv:2412.14172*, 2024.
- [16] F. Mentzer, D. Minnen, E. Agustsson, and M. Tschanen, "Finite scalar quantization: Vq-vae made simple," *arXiv preprint arXiv:2309.15505*, 2023.
- [17] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Zionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 975–10 985.
- [18] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.
- [19] A. Razavi, A. Van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with vq-vae-2," *Advances in neural information processing systems*, vol. 32, 2019.
- [20] Y. Shafir, G. Tevet, R. Kapon, and A. H. Bermano, "Human motion diffusion as a generative prior," *arXiv preprint arXiv:2303.01418*, 2023.
- [21] Y. Shao, X. Huang, B. Zhang, Q. Liao, Y. Gao, Y. Chi, Z. Li, S. Shao, and K. Sreenath, "Langwbc: Language-directed humanoid whole-body control via end-to-end learning," *arXiv preprint arXiv:2504.21738*, 2025.
- [22] J. Shi, X. Liu, D. Wang, O. Lu, S. Schwertfeger, F. Sun, C. Bai, and X. Li, "Adversarial locomotion and motion imitation for humanoid policy learning," *arXiv preprint arXiv:2504.14305*, 2025.
- [23] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano, "Human motion diffusion model," *arXiv preprint arXiv:2209.14916*, 2022.
- [24] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 5026–5033.
- [25] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.
- [26] L. Xiao, S. Lu, H. Pi, K. Fan, L. Pan, Y. Zhou, Z. Feng, X. Zhou, S. Peng, and J. Wang, "Motionstreamer: Streaming motion generation via diffusion-based autoregressive model in causal latent space," *arXiv preprint arXiv:2503.15451*, 2025.
- [27] J. Yue, Z. Wang, Y. Wang, W. Zeng, J. Wang, X. Xu, Y. Zhang, S. Zheng, Z. Ding, and Z. Lu, "RI from physical feedback: Aligning large motion models with humanoid control," *arXiv preprint arXiv:2506.12769*, 2025.
- [28] Y. Ze, Z. Chen, J. P. Arañjo, Z.-a. Cao, X. B. Peng, J. Wu, and C. K. Liu, "Twist: Teleoperated whole-body imitation system," *arXiv preprint arXiv:2505.02833*, 2025.
- [29] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," *arXiv preprint arXiv:2203.03605*, 2022.
- [30] J. Zhang, Y. Zhang, X. Cun, Y. Zhang, H. Zhao, H. Lu, X. Shen, and Y. Shan, "Generating human motion from textual descriptions with discrete representations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 14 730–14 740.