Towards Consistent Long-Term Pose Generation

Yayuan Li[1] Filippos Bellos[1] Jason J. Corso[1,2]
[1]University of Michigan [2]Voxel51
Project Page

Abstract

Current approaches to pose generation rely heavily on intermediate representations, either through two-stage pipelines with quantization or autoregressive models that accumulate errors during inference. This fundamental limitation leads to degraded performance, particularly in long-term pose generation where maintaining temporal coherence is crucial. We propose a novel one-stage architecture that directly generates poses in continuous coordinate space from minimal context - a single RGB image and text description - while maintaining consistent distributions between training and inference. Our key innovation is eliminating the need for intermediate representations or token-based generation by operating directly on pose coordinates through a relative movement prediction mechanism that preserves spatial relationships, and a unified placeholder token approach that enables single-forward generation with identical behavior during training and inference. Through extensive experiments on Penn Action and First-Person Hand Action Benchmark (F-PHAB) datasets, we demonstrate that our approach significantly outperforms existing quantization-based and autoregressive methods, especially in long-term generation scenarios.

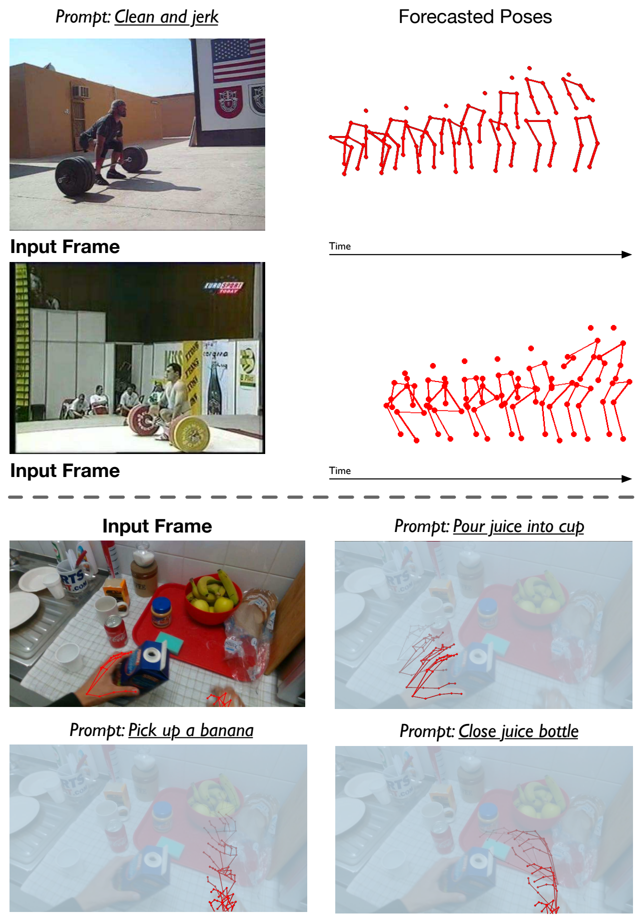

Figure 1. Examples of pose generation from a single RGB image and text description.

1. Introduction

Human pose generation has emerged as a fundamental problem in computer vision, with applications spanning animation synthesis, action understanding, and motion prediction [4, 10, 16]. Recent work has explored various approaches to control this generation process using different modalities: from textual descriptions [1, 11], to audio signals [9, 13], to scene context [3, 18].

Creating semantically meaningful and contextually appropriate poses remains challenging, particularly due to architectural limitations in existing approaches. These approaches typically fall into two restrictive paradigms. First, they rely on autoregressive models that generate poses frame-by-frame which injects a distribution shift between training and inference due to their nature [2]. This distribution shift then leads to degraded long-term performance due to accumulated performance [6], as we show later in this paper. Second, they are two-stage approaches that first convert continuous pose coordinates into discrete tokens, latent codes through VAEs [15, 16] or quantization before generation [11], introducing information loss and computational overhead.

These approaches show significant degradation when

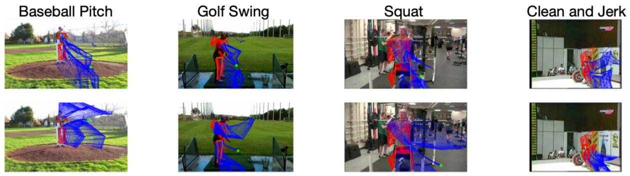

Figure 2. Long-term forecasting errors in existing methods: red indicates ground truth, blue indicates predictions. Errors accumulate due to autoregressive training. Top: LSTM; Bottom: Transformer.

generating longer sequences, as both quantization errors and distribution shifts compound over time (as demonstrated in Figure 2). This degradation affects many downstream applications (e.g., in task guidance where long-term semantic coherence is crucial [5, 17]). Additionally, most of these methods require complex inputs like 3D scene information [19, 20], assuming the availability of such detailed data, which limits their practicality in broad real-world applications.

To address these fundamental limitations in pose generation, we introduce two key novelties within our approach:

1. A unified prediction mechanism that ensures consistent distributions between training and inference, enabling reliable long-term generation.
2. A one-stage pose generation architecture that directly operates in continuous coordinate space from minimal input—a single RGB image and text description—preserving both spatial fidelity and semantic alignment, without relying on scarce 3D detailed scene information.

We also explore how language guidance can provide semantic control over the generated motions. Natural language offers an intuitive and flexible way to specify desired movements. We leverage short and concise natural language descriptions rather than the detailed movement specifications required by prior work [7, 12]. This enables effective control without requiring complex movement specifications or detailed scene understanding. This combination of robust long-term generation with language control facilitates applications from animation synthesis to motion planning and task-guidance.

We evaluate the effectiveness of our method on Penn Action [22] and First-Person Hand Action Benchmark (F-PHAB) [8] datasets across body and hand pose, viewpoints and domains. With four metrics measuring performance, we benchmark against five strong baselines. Our approach consistently outperforms baselines, achieving significant gains in both short-term and long-term pose generation. Notably, our method excels in challenging scenarios involving large motions and complex temporal dynamics. Ablation studies and qualitative results demonstrate the integration of visual and textual context, along with our architecture design choices, are crucial.

2. Approach

2.1. Problem Statement

Given a natural language prompt and a single RGB image $I \in \mathbb{Z}^{H\times W\times 3}$, our goal is to predict a sequence of k future poses $\mathbf{P} = \{P_i\}_{i=1}^{k}$ that aligns semantically with the prompt and visually with the scene. Each pose $P_i \in R^{2N}$ represents 2D coordinates of N keypoints. Unlike prior work requiring 3D scene data [20], we operate directly in the continuous coordinate space.

2.2. Method

Our one-stage architecture predicts future poses in continuous space from multimodal input. A vision-language encoder extracts features: the image I is processed by f_I to yield $F_I \in R^{N_I\times d_I}$; the prompt L is passed through f_M for fused features $F_M \in R^{N_M\times d_M}$. A Transformer decoder, conditioned on the initial pose P_0, forecasts future poses.

Training-Inference Alignment We avoid autoregressive drift by predicting all future poses jointly using non-masked self-attention and placeholder tokens [PRD]. Unlike next-token prediction (NTP) methods [11] prone to accumulating error, our decoder input aligns training and inference distributions:

$$X^{ours} = \begin{pmatrix} x_1^0 & y_1^0 & \cdots & x_N^0 & y_N^0 \\ [\text{PRD}]_1 & \cdots & \cdots & \cdots & [\text{PRD}]_{2N} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ [\text{PRD}]_1 & \cdots & \cdots & \cdots & [\text{PRD}]_{2N} \end{pmatrix} \quad (1)$$

The decoder maps (P_0, F_M) to $\hat{P} \in R^{T\times 2N}$ with a single forward pass:

$$\hat{P} = \text{Decoder}(P_0, F_M) \quad (2)$$

Relative Pose Forecasting Instead of predicting absolute coordinates, we forecast displacements from P_0, e.g., predicting $(\Delta x = -0.05, \Delta y = 0.1)$ from $(0.75, 0.8)$ to $(0.7, 0.9)$. This promotes spatial coherence and reduces global redundancy.

Vision-Language Encoding Compact prompts (e.g., *"swing golf"*) are encoded with BLIP [14]. f_I is the frozen image encoder; f_M is BLIP's image-grounded text encoder that fuses L and I.

2.3. Relative Pose Representation Loss

To model joint spatial structure, we define pairwise distance and direction matrices between joints.

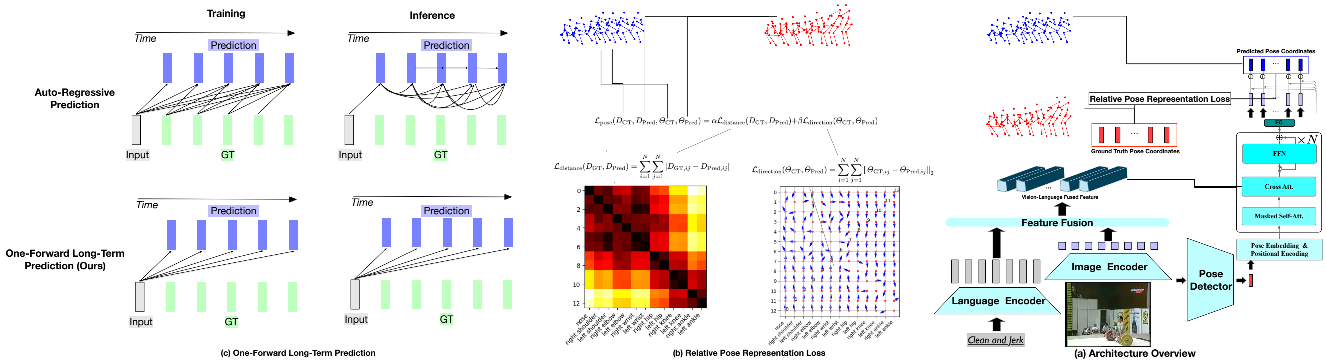

Figure 3. Overview of our proposed method. Given a single RGB image I and a natural language action description L, our model extracts vision-language fused features using a multimodal encoder. These features, along with the initial pose P_0, are fed into a Transformer decoder, which predicts a sequence of future poses $\hat{P}_{1\dots T}$. Our method employs cross-attention to capture the interaction between the visual and textual inputs, ensuring that the forecasted poses align with the provided context.

Distance Representation

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{3}$$

Direction Representation

$$\Theta_{ij} = \left(\frac{x_j - x_i}{D_{ij}}, \frac{y_j - y_i}{D_{ij}} \right) \tag{4}$$

Loss Formulation

$$\mathcal{L}_{\text{distance}} = \sum_{i,j} |D_{\text{GT},ij} - D_{\text{Pred},ij}| \tag{5}$$

$$\mathcal{L}_{\text{direction}} = \sum_{i,j} \|\Theta_{\text{GT},ij} - \Theta_{\text{Pred},ij}\|_2 \tag{6}$$

$$\mathcal{L}_{\text{pose}} = \alpha \mathcal{L}_{\text{distance}} + \beta \mathcal{L}_{\text{direction}} \tag{7}$$

$$\mathcal{L}_{\text{seq}} = \frac{1}{k} \sum_{i=1}^{k} \mathcal{L}_{\text{pose}} \tag{8}$$

$$\mathcal{L}_{\text{batch}} = \frac{1}{B} \sum_{i=1}^{B} \mathcal{L}_{\text{seq},i} \tag{9}$$

$$\mathcal{L} = \mathcal{L}_{\text{rel}}(\alpha, \beta) + \theta \mathcal{L}_{\text{batch},mse} \tag{10}$$

3. Experiments

We validate our model on two pose forecasting benchmarks and compare it against strong baselines using four standard metrics. This section details the datasets, evaluation metrics, implementation, baselines, and results, including ablations and comparisons with prior work.

3.1. Datasets

We use Penn Action [22] for full-body pose and F-PHAB [8] for hand pose in egocentric views. Each dataset contains short natural language descriptions paired with videos. For missing annotations, we apply MediaPipe to extract pseudo-labels. Training uses 90% of the videos, and testing uses 10%. The forecasting horizon is 45 frames.

3.2. Metrics

We use:

- **ADE**: Average distance over predicted keypoints and frames.
- **FDE**: Distance at the last timestamp.
- **PCK**: Percentage of keypoints within a threshold (0.05 for body, 0.15 for hand).
- **RMSE**: Root mean squared error.

3.3. Implementation Details

BLIP is used for vision-language fusion (ViT-g/14 + BERT). We freeze BLIP and train the Transformer decoder using AdamW (lr 10^{-4}, batch 64) on one NVIDIA H100. The loss uses a mix of MSE and relative pose losses with weights α=1.0, β=1.0, and θ=0.1.

3.4. Baselines

We evaluate against:

- $\mathbf{NN_P}$: Nearest neighbor by input pose.
- $\mathbf{NN_{VL}}$: Nearest neighbor by fused features.
- **LSTM**: Autoregressive model with next-token prediction.
- **Transformer (NTP)**: Transformer decoder with causal masking.
- **Quant.+TF**: Two-stage approach with pose quantization and Transformer decoding.

Method	Penn Action				F-PHAB			
	ADE↓	FDE↓	PCK↑	RMSE↓	ADE↓	FDE↓	PCK↑	RMSE↓
NN_P	0.090	0.105	0.666	0.057	0.168	0.154	0.377	0.109
NN_{VL}	0.242	0.246	0.300	0.157	0.258	0.259	0.279	0.214
LSTM	0.164	0.262	0.382	0.106	0.194	0.194	0.302	0.136
Transformer (NTP)	0.173	0.230	0.344	0.111	0.192	0.203	0.300	0.146
Quant.+TF	0.255	0.248	0.180	0.166	0.243	0.239	0.208	0.160
Ours	**0.058**	**0.077**	**0.818**	**0.035**	**0.097**	**0.086**	**0.765**	**0.068**

Table 1. Comparison with baseline models on both datasets.

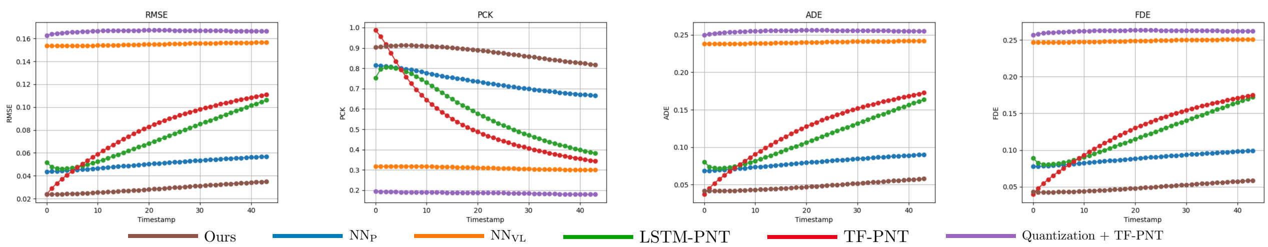

Figure 4. Performance across timestamps. Our model is robust to longer horizons.

	ADE↓	FDE↓	PCK↑	RMSE↓
NN_P	0.112	0.165	0.549	0.070
NN_{VL}	0.225	0.258	0.181	0.145
LSTM	0.174	0.289	0.316	0.112
Transformer	0.168	0.255	0.327	0.108
Quant.+TF	0.247	0.256	0.124	0.159
Ours	**0.092**	**0.157**	**0.682**	**0.057**

Table 2. Results on the hardest 10% test samples (Penn Action).

Variant	ADE↓	FDE↓	PCK↑	RMSE↓
TF (NTP)	0.173	0.230	0.344	0.111
+ pose det.	0.125	0.155	0.441	0.098
+ full attn.	0.069	0.069	0.774	0.043
+ causal mask	0.060	0.074	0.820	0.037
+ rel. loss (ours)	**0.058**	**0.077**	**0.818**	**0.035**

Table 3. Ablation study on Penn Action.

3.5. Results

Main (Tab. 1) Our model clearly outperforms all baselines in both datasets and across all metrics. Autoregressive models degrade due to error accumulation. Quant.+TF suffers from codebook limitations. Our model avoids both and delivers accurate predictions in one forward pass.

Timestamp and Hard Sample Analysis Figure 4 shows performance over time. Our accuracy remains stable while others degrade. In Tab. 2, we also evaluate on the hardest 10% samples (by keypoint motion variance):

Ablation Study Each design choice improves performance, particularly the transition to single-stage decoding and use of relative geometry loss.

Comparison with SOTA (Tab. 4) Despite using only one RGB frame and short text, our method outperforms both state-of-the-art text-only and vision-only pose generation models.

Method	ADE↓	FDE↓	PCK↑	RMSE↓
TM2T [11]	0.268	0.292	0.171	0.271
PHD [21]	–	–	0.772	–
Ours	**0.017**	**0.017**	**0.860**	**0.012**

Table 4. Comparison with SOTA single-modality methods on Penn Action.

4. Conclusion

We introduce a one-stage, vision-language-guided pose forecaster that operates in continuous coordinate space and, by aligning training and inference through relative-movement prediction, produces spatially faithful sequences. Extensive experiments on Penn Action and F-PHAB demonstrate state-of-the-art performance across multiple metrics, clearly surpassing strong baselines. Moreover, the model remains robust under large motions and long forecasting horizons.

References

[1] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. *2019 International Conference on 3D Vision (3DV)*, pages 719–728, 2019. 1

[2] Gregor Bachmann and Vaishnavh Nagarajan. The pitfalls of next-token prediction, 2024. 1

[3] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. *ArXiv*, abs/2007.03672, 2020. 1

[4] Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1423–1432. IEEE, 2019. 1

[5] Dima Damen, Michael Wray, Ivan Laptev, Josef Sivic, et al. Genhowto: Learning to generate actions and state transformations from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6561–6571, 2024. 2

[6] Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jian, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaïd Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality. *ArXiv*, abs/2305.18654, 2023. 1

[7] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J. Black. ChatPose: Chatting about 3d human pose. In *CVPR*, 2024. 2

[8] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3

[9] Kehong Gong, Dongze Lian, Heng Chang, Chuan Guo, Zihang Jiang, Xinxin Zuo, Michael Bi Mi, and Xinchao Wang. Tm2d: Bimodality driven 3d dance generation via music-text integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9942–9952, 2023. 1

[10] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, page 2021–2029, New York, NY, USA, 2020. Association for Computing Machinery. 1

[11] 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. 1, 2, 4

[12] Bolin Lai, Xiaoliang Dai, Lawrence Chen, Guan Pang, James M Rehg, and Miao Liu. Lego: Learning egocentric action frame generation via visual instruction tuning. *arXiv preprint arXiv:2312.03849*, 2023. 2

[13] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. *ArXiv*, abs/1911.02001, 2019. 1

[14] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2

[15] Sihan Ma, Qiong Cao, Jing Zhang, and Dacheng Tao. Contact-aware human motion generation from textual descriptions. *arXiv preprint arXiv:2403.15709*, 2024. 1

[16] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021. 1

[17] Jackson Spencer, Sanjiban Choudhury, Matthew Barnes, Christopher Dellin, et al. What matters in learning from offline human demonstrations for robot manipulation. In *Conference on Robot Learning*, 2022. 2

[18] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9401–9411, 2021. 1

[19] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[20] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say, interact as you can: Language-guided human motion generation with scene affordance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[21] Jason Y Zhang, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik. Predicting 3d human dynamics from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7114–7123, 2019. 4

[22] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE international conference on computer vision*, pages 2248–2255, 2013. 2, 3