ROBUST MOTION GENERATION USING PART-LEVEL RELIABLE DATA FROM VIDEOS

Boyuan Li[1], *Sipeng Zheng*[4], *Bin Cao*[3], *Ruihua Song*[1*], *Zongqing Lu*[2,4]

[1]RUC [2]PKU [3]CASIA [4]BeingBeyond

ABSTRACT

Extracting human motion from large-scale web videos offers a scalable solution to the data scarcity issue in character animation. However, some human parts in many video frames cannot be seen due to off-screen captures or occlusions. It brings a dilemma: discarding the data missing any part limits scale and diversity, while retaining it compromises data quality and model performance. To address this problem, we propose leveraging credible part-level data extracted from videos to enhance motion generation via a robust part-aware masked autoregression model. First, we decompose a human body into five parts and detect the parts clearly seen in a video frame as "credible". Second, the credible parts are encoded into latent tokens by our proposed part-aware variational autoencoder. Third, we propose a robust part-level masked generation model to predict masked credible parts, while ignoring those noisy parts. In addition, we contribute K700-M, a challenging new benchmark comprising approximately 200k real-world motion sequences, for evaluation. Experimental results indicate that our method successfully outperforms baselines on both clean and noisy datasets in terms of motion quality, semantic consistency and diversity.

Index Terms— Character Animation, Robust Motion Synthesis

1. INTRODUCTION

The creation of lifelike human motion is critical for applications, e.g., spanning film, gaming, and virtual reality, fueling significant interest in data-driven motion generation [1]. However, the field is constrained by a scarcity of high-quality data, as motion capture systems are costly and typically limited to controlled laboratory settings [2]. To overcome this bottleneck, recent works [3, 4, 5] have turned to extracting motion from large-scale web videos. Though existing motion reconstruction methods [6, 7, 8] excel with fully visible human body, they suffer significant performance degradation under real-world conditions with partial occlusion or incomplete views [9]. Taking the online video dataset Kinetics-700 [10] as an example, the frames containing full bodies account for only about 24% of the total frames. Consequently, these approaches introduce low-quality data with inherent noise, such as foot sliding and occlusion, when extracting the 3D motion from real-world videos.

Prevailing text-to-motion (T2M) models [11, 12, 13, 14, 15, 16] often overlook the pervasive issue of part-level noise in web-sourced training data. This oversight compromises dataset integrity and teaches models an erroneous distribution, raising a critical question: *How can we effectively harness noisy web data for robust T2M generation?*

To address this problem, we propose a novel framework designed to selectively use part-level reliable reconstructed motion data to generate the full-body motion from text. Specifically, we first divide the human body into five parts based on kinematic structure and use a pose estimation model [17] to detect whether a part has actually appeared in a video frame, which is labeled as "credible" and whose 3D motion data is reliable for training usage. Conversely, the data of body parts not present in the video are considered "noisy". Second, we propose using a part-aware Variational Autoencoder (P-VAE) tolearn part-level continuous encoding through variational loss on reliable part-level data. Third, we train a robust part-aware mask generation model based on the credible parts while ignoring the noisy parts to take a full use of data and avoid negative impact as well.

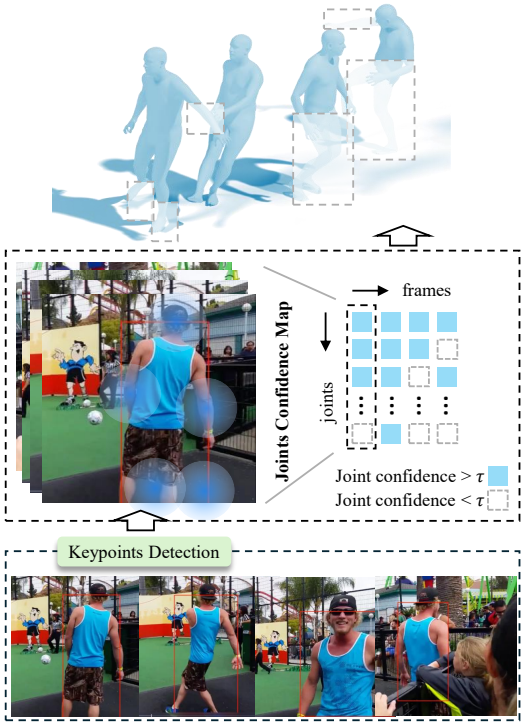

Fig. 1. We determine whether a joint is actually visible in the frames of web videos by joint confidence detected by ViTPose [17]. The invisible joints will be given a lower confidence, whose 3D information is not reliable.

By enhancing robustness to noise, our approach significantly expands the usable scope of web-mined motion data. For validation, we introduce K700-M based on Kinetics-700 [10], a challenging real-world dataset comprising approximately 200K noisy motion sequences curated from web videos. Extensive experiments demonstrate that our model achieves significant improvements over strong baselines in quality, semantic consistency, and diversity.

In summary, our main contributions are as follows: **(1) Robust Part-aware Text-to-Motion Generation Framework.** We present a novel framework designed for noise robustness. By decomposing

*Corresponding author.

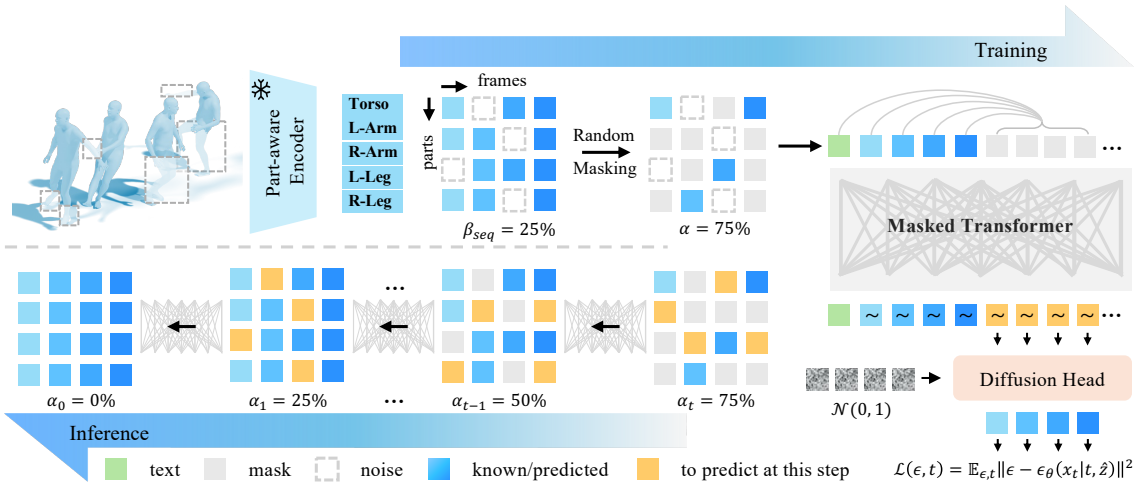

Fig. 2. Overview of our proposed Robust Part-aware Generation Model(RoPAR). The model first performs part-wise motion decomposition, followed by a masked autoregressive generation on the credible latent tokens. Finally, a diffusion head refines the latent quality.

motion data and employing a selective encoding mechanism , our model becomes the first to effectively learn from training data containing part-level motion noise, focusing on high-fidelity segments while discarding artifacts. **(2) The K700-M Dataset.** We construct a large-scale, real-world dataset comprising around 200K motion sequences extracted from web videos. This dataset addresses the need for a benchmark that reflects the challenges of part-level motion data. **(3) Advanced T2M Performance using Part-level Data.** We empirically validate that our framework sets a new state-of-the-art, outperforming existing methods in motion quality, semantic alignment with text, and output diversity. This success demonstrates the significant potential of leveraging noisy dataset on the Internet to enhance model generalization and capability.

2. METHOD

2.1. Overview

Given a text prompt T, our goal is to generate a motion sequence $\{m_i\}_{i=1}^N$ by leveraging noisy motion data reconstructed from online videos, most of which contain unreliable information of parts. To address the problem, we propose a novel framework composed of three main modules. First, as shown in Fig. 1, we reconstruct 3D motion data from video frames and output the detected joints with the confidence on how reliable the information is. Thus we can identify the credible body parts, of which the average confidence of joints is lower than a threshold τ. Second, we propose a part-aware vairantional autoencoder, as shown in Fig. 2, to encode each credible part into a latent token. Consequently, we leave the other noisy parts out and use only the credible part tokens to compose a part-aware motion matrix. Its row corresponds to a part and its column corresponds to a frame. Third, we propose a masked transformer with a diffusion head [18, 19, 20] that learns to predict masked part tokens and refine them before decoding. During training, the model randomly masks some credible part tokens and optimize the loss of predicting them. During inference, the model is able to predict all parts step by step. We give more details on each module as follows.

2.2. Identifying Credible Parts in Reconstructed Motion

The 3D motion data reconstructed from web videos inevitably contains a portion of body part data which has never appeared in the video. This portion introduces the common noise such as jitter and foot sliding. In order to distinguish this kind of data, we leverage a 2D pose estimation model called ViTPose [17] to obtain per-joint confidence scores, C_j, which indicate the visibility and detection accuracy of each joint. We utilize these scores to identify reliable motion data.

Specifically, we first partition the human skeleton into five kinematic chains: the torso, left arm, right arm, left leg, and right leg. Then, as shown in Fig. 1, a body part p is classified as *credible* if its average confidence score C_p exceeds a threshold τ, where $C_p = \frac{1}{|J_p|}\sum_{j\in J_p} C_j$. J_p denotes the set of joints in part p A high C_p indicates clear visibility and accurate prediction, denoting reliable motion data. Conversely, parts falling below τ are marked as *noisy*, perhaps due to occlusion or poor detection. In summary, this mechanism allows our model to differentiate between credible and incredible motion at a granular level.

2.3. Part-aware Variational Autoencoder (P-VAE)

To compress the partially observed motion sequences into a compact and semantically rich latent space, we propose P-VAE to learn robust, noise-free latent part-level representations using only the reliable (non-noisy) body parts in the motion sequences. This serves as a critical foundation for our subsequent generative model.

Specifically, the motion data for a single credible part p within a sequence of frames is denoted as m^p, Following [15, 21], each frame m_i^p is represented by the tuple $\{r^x, r^z, r^a, j^p, j^v, j^r\}$, encompassing root linear velocities, root angular velocity, and the joint positions, velocities, and rotations for the joints in part p, respectively. To ensure global spatial consistency, the root and spine joints are included in every part's joint set. The training objective for the P-VAE is defined as:

$$\mathcal{L}_{\text{P-VAE}} = \sum_{p\in P_{\text{credible}}} (\mathcal{L}_{\text{recon}}(m^p) + \lambda \cdot \mathcal{L}_{\text{KL}}(m^p)) \quad (1)$$

where $P_{\text{credible}} = \{m^p | C_p > \tau\}$ is the set of all credible part motions. Here, $\mathcal{L}_{\text{recon}}$ denotes the reconstruction loss, and $\mathcal{L}_{\text{KL}}$ is the KL

Table 1. Comparison with existing T2M generation methods on the HumanML3D and K700-M dataset.

Methods	Humanml3d[1]					K700-M				
	FID ↓	R1↑	R2↑	R3↑	MM-D↓	FID ↓	R1↑	R2↑	R3↑	MM-D↓
ParCo[12]	19.929	0.416	0.490	0.596	19.472	44.321	0.430	0.544	0.618	20.392
MotionGPT[13]	14.175	0.436	0.598	0.668	17.890	30.954	0.525	0.683	0.793	17.259
Momask[11]	10.731	0.622	0.782	0.850	16.128	24.334	0.585	0.761	0.834	15.827
MotionStreamer[15]	**10.724**	0.631	0.784	0.851	16.639	25.231	0.591	0.758	0.836	15.459
Ours(RoPAR)	12.674	**0.701**	**0.836**	**0.895**	**15.382**	**19.682**	**0.711**	**0.864**	**0.891**	**13.612**

divergence that regularizes the latent posterior distribution towards a standard Gaussian prior. Besides, our P-VAE uses shared parameters across the part-specific encoders and decoders.This design prevents overfitting to specific body parts, and compels the model to learn a unified latent representation that can effectively describe the characteristics of body parts.

By training on credible data, our P-VAE avoids learning corrupted parts, ensuring the derived latent space are of high quality and devoid of noise. Consequently, the noisy motion sequence will be encoded to continuous latents Z by the P-VAE encoder for subsequent usage: $\{Z|z_i^p = E_{\text{p-vae}}(m_i^p)\}$ for $i \in \{1,\dots,N\}$ and $p \in \{1,\dots,P\}$, N and P denotes the frames and parts respectively.

2.4. Robust Part-aware Text-to-Motion Generation Model

To achieve the text to motion generation, where we can only harness the reliable parts in the data meanwhile avoid the impact of the part-level noise, we introduce a robust part-aware text-to-motion generation model(RoPAR). This model is trained to predict masked latent tokens conditioned on all unmasked tokens and the text prompt, with a specialized strategy to handle inherent noise, and a diffusion module to refine the predicted latents.

Part-aware Masking Strategy. Unlike [22, 23, 24], our strategy is explicitly guided by the identified part reliability. For a sequence with a noisy ratio β_{noisy}, the masking probability for a latent token z is defined as:

$$P(z=[M]) = \begin{cases} 1 & \text{if } z \in Z_{\text{noisy}} \\ \alpha - \beta_{\text{seq}} & \text{if } z \notin Z_{\text{noisy}} \end{cases} \quad (2)$$

Here, $[M]$ is a learnable mask token and $Z_{\text{noisy}} = \{z_i^p | C_i^p < \tau, z_i^p \in Z\}$ denotes the set of tokens corresponding to parts previously identified as noisy. Different from [22] that employ a uniform random masking strategy, we introduce a part-aware masking strategy adapt to corrupted motion data. Specifically, all latent tokens corresponding to noisy body parts are unconditionally masked, and will not be used to calculate gradients. For the remaining credible tokens, we apply a progressive random masking strategy as Eq.2. Consequently, the overall sequence masking ratio are unified to α regardless the initial noise ratio β_{seq} of sequence. This mechenism prevents the model from learning noise data from these incredible parts and is forced to learn the robust constraints between the credible tokens.

Diffusion Head. We employ a diffusion head to refine the output of the autoregressive model, combining the strengths of both paradigms for high-quality, diverse generation. Following [19, 20], a lightweight MLP denoiser [25] ϵ_θ is trained to predict the gaussian noise ϵ in the corrupted latent x_t at timestep t, given $\hat{z}$:

$$\mathcal{L}_{\text{Diffusion}} = \mathbb{E}_{\epsilon,t}\left\|\epsilon - \epsilon_\theta(x_t|t,\hat{z})\right\|^2 \quad (3)$$

This diffusion step acts as a powerful post-processor, enhancing the final motion's fidelity and variability by learning to correct and refine the initial autoregressive predictions.

3. EXPERIMENTS

3.1. Experimental Setup

Datasets. To confirm whether our method can handle both clean and noisy data, we use two datasets, i.e., HumanML3D [1] and our newly introduced K700-M, to conduct experiments. HumanML3D dataset is widely-used noise-free dataset comprising 14,616 motion clips from AMASS [2], each paired with three text descriptions. To evaluate performance on noisy, in-the-wild data, we construct the K700-M dataset from Kinetics-700 [10], which consists of over 630K real-world YouTube videos exhibiting a wide variety of scenes, lighting conditions, and camera angles. We process these videos using a reconstruction method called WHAM [6] to extract 198,627 human motion sequences. The motions are subsequently smoothed with a low-pass filter to minimize jitter and abrupt changes in joint velocities. We annotate each video clip with Gemini [26]. We also retrain an evaluator on this dataset following [27].

Evalutation metrics. To comprehensively evaluate the quality of the generated motions, we adopt a series of widely used metrics from the human motion generation literature, including (1) Fréchet Inception Distance (FID) which measures the distributional similarity between generated and real motions; (2) R-Precision which evaluates the semantic match between a motion and its text description; (3) MM-Distance which measures the average distances of the generated motion and the text embeddings; and (4) Mean Per-Joint Position Error (MPJPE) which measures the reconstruction quality by average joint position error.

3.2. Comparisons to State-of-the-Art Baselines

We compare our method with the following baselines: ParCo [12] uses part-level tokens and needs accurate full-body data to train; MotionGPT [13], Momask [11] and MotionStreamer [15] are the recent state-of-the-art methods based on T5 [28], mask transformer, and decoder-only archtectures, but all of them are use the full-body data to train. Results are presented in Table 1. Results show that our model havecompetitive performance to the strong baselines on the

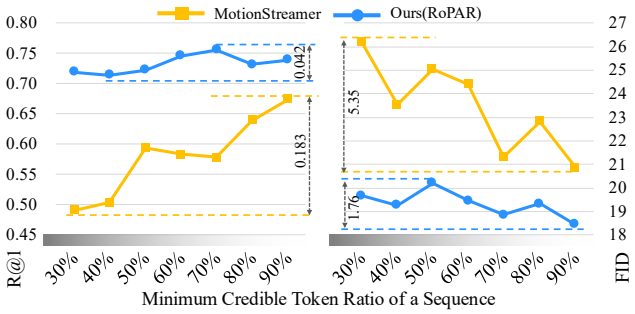

Fig. 3. Sensitivity analysis between RoPAR and baseline [15].

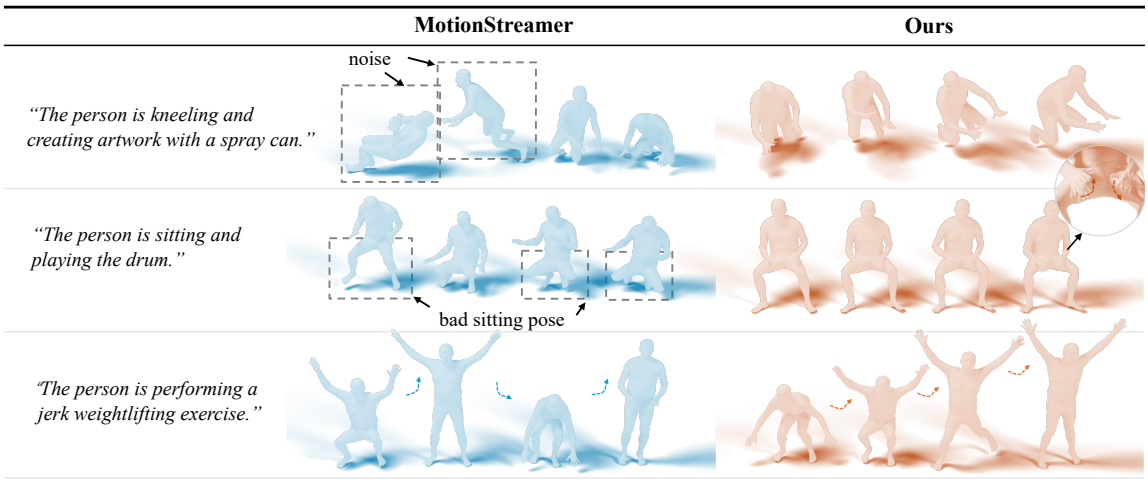

Fig. 4. Qualitative results between our method and baseline on K700-M. Ours show richer details and more natural transitions.

non-noisy HumanML3D dataset, particularly in R-precisions. This indicates that our part-aware encoding strategy effectively enhances text-motion alignment. More significantly, on the noisy K700-M dataset, our method substantially outperforms all baselines across all metrics. These results strongly demonstrate the effectiveness and robustness of our approach in efficiently leveraging large-scale, noisy motion data without discarding it or introducing disturbance. Overall, our part-aware masking strategy works the best on both the regular non-noisy and the partially noisy motion dataset, thanks to its unified training framework and advanced modeling capability.

Our method's robustness is further evidenced by the sensitivity analysis in Fig. 3. The x-axis represents the proportion of credible tokens in a training sequence (e.g., 30% means at least 30% of the tokens are noise-free). Our method maintains stable performance with minimal fluctuation in both R@1 and FID across all noise levels. In contrast, the baselines exhibit significant performance degradation. This consistency underscores our framework's superior robustness and its capability to handle real-world, imperfect motion data.

Table 2. Ablation results of P-VAE and RoPAR, where MPJPE and R1 is measured for reconstruction and semantic consistancy.

Reconstruction	FID ↓	MPJPE↓
P-VAE	**0.21**	**6.95**
w/o part-wise decomposition	1.86	19.83
w/o shared parameter	0.89	9.82
Generation	FID↓	R1↑
RoPAR	**19.68**	**0.71**
w/o part-wise decomposition	21.36	0.58
w/o diffusion head	71.92	0.41

3.3. Ablation Study

Table 2 presents ablation results on the K700-M dataset to validate the key components of our framework: the Part-aware VAE (P-VAE) and the Part-aware Masked Autoregressive Generation Model. To evaluate the P-VAE, we first replace it with a traditional full-body VAE that encodes motion sequences holistically without part-aware decomposition. This variant exhibits significantly higher FID and MPJPE, indicating a higher reconstruction error and a lower-quality latent space. This result confirms that our part-aware encoding is crucial for learning a precise latent representation, as it effectively isolates noisy body parts from interfering with credible ones. In addition, we validate the role of our shared parameter mechanism in the P-VAE's encoder and decoder. Removing such mechanism leads to a significant performance drop across all metrics, which underscores the importance of parameter sharing for maintaining spatial consistency among body parts and ensuring a high-quality latent space for motion reconstruction.

The ablation on RoPAR further highlights the importance of our design choices. First, training the autoregressive model without part-level decomposed sequences leads to clear performance degradation: FID increases from 19.68 to 21.36, and all R-Precision scores drop. This indicates that our part-aware decomposition enables more effectively learning from noisy incomplete data by processing credible and incredible parts separately. Second, removing the diffusion-based loss and denoising head causes a dramatic performance collapse, with FID soaring to 71.92 and R-Precision scores plummeting. The results emphasize that the diffusion head is essential for refining the coarse autoregressive output, allowing the model to learn the precise latent space distribution and generate high-fidelity motions.

3.4. Qualitative Results

We showcase the advantages of our method over the best baseline MotionStreamer by visualizing generation results from the same text prompt in Fig. 4. When generating motions which often filmed in upper-body close-ups (e.g., "sitting and playing the drum"), our RoPAR model generates motions with richer detail and more natural transitions due to its full utilization of partial credible data during training. In contrast, the baseline model, susceptible to dataset noise, often produces motions with unnatural stiffness and distortion.

4. CONCLUSION

Our research provides a new perspective on how to achieve high-quality motion generation in the presence of large scale motion data collected from online videos with part-level noise. It also offers a more efficient and reliable solution for data pre-processing and model training in future related tasks. In this way, generative models can flexibly address the common challenge of missing data in the real world without relying on lengthy pre-processing steps.

5. REFERENCES

[1] 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*, 2022, pp. 5152–5161.

[2] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael 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.

[3] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang, "Motion-x: A large-scale 3d expressive whole-body human motion dataset," *Advances in Neural Information Processing Systems*, vol. 36, pp. 25268–25280, 2023.

[4] Ye Wang, Sipeng Zheng, Bin Cao, Qianshan Wei, Weishuai Zeng, Qin Jin, and Zongqing Lu, "Scaling large motion models with million-level human motions," *arXiv preprint arXiv:2410.03311*, 2024.

[5] 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," 2025.

[6] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black, "Wham: Reconstructing world-grounded humans with accurate 3d motion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2070–2080.

[7] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou, "World-grounded human motion recovery via gravity-view coordinates," in *SIGGRAPH Asia Conference Proceedings*, 2024.

[8] Wanqi Yin, Zhongang Cai, Ruisi Wang, Fanzhou Wang, Chen Wei, Haiyi Mei, Weiye Xiao, Zhitao Yang, Qingping Sun, Atsushi Yamashita, Lei Yang, and Ziwei Liu, "Whac: World-grounded humans and cameras," in *European Conference on Computer Vision*. Springer, 2024, pp. 20–37.

[9] Siwei Zhang, Bharat Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, Petr Kadlecek, Siyu Tang, and Federica Bogo, "Rohm: Robust human motion reconstruction via diffusion," 2024.

[10] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman, "A short note on the kinetics-700 human action dataset," *arXiv preprint arXiv:1907.06987*, 2019.

[11] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li 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.

[12] Qiran Zou, Shangyuan Yuan, Shian Du, Yu Wang, Chang Liu, Yi Xu, Jie Chen, and Xiangyang Ji, "Parco: Part-coordinating text-to-motion synthesis," in *European Conference on Computer Vision*. Springer, 2024, pp. 126–143.

[13] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen, "Motiongpt: Human motion as a foreign language," *Advances in Neural Information Processing Systems*, vol. 36, pp. 20067–20079, 2023.

[14] Chuhao Jin, Haosen Li, Bingzi Zhang, Che Liu, Xiting Wang, Ruihua Song, Wenbing Huang, Ying Qin, Fuzheng Zhang, and Di Zhang, "Planmogpt: Flow-enhanced progressive planning for text to motion synthesis," 2025.

[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.

[16] Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu, "Finemogen: Fine-grained spatio-temporal motion generation and editing," 2023.

[17] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao, "ViTPose: Simple vision transformer baselines for human pose estimation," in *Advances in Neural Information Processing Systems*, 2022.

[18] Prafulla Dhariwal and Alex Nichol, "Diffusion models beat gans on image synthesis," 2021.

[19] Alex Nichol and Prafulla Dhariwal, "Improved denoising diffusion probabilistic models," 2021.

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *arXiv preprint arxiv:2006.11239*, 2020.

[21] Bin Cao, Sipeng Zheng, Ye Wang, Lujie Xia, Qianshan Wei, Qin Jin, Jing Liu, and Zongqing Lu, "A real-time controllable vision-language-motion model," in *ICCV*, 2025.

[22] Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang, "Rethinking diffusion for text-driven human motion generation: Redundant representations, evaluation, and masked autoregression," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 27859–27871.

[23] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman, "Maskgit: Masked generative image transformer," 2022.

[24] Tianhong Li, Huiwen Chang, Shlok Kumar Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan, "Mage: Masked generative encoder to unify representation learning and image synthesis," 2023.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," 2015.

[26] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al., "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024.

[27] Mathis Petrovich, Michael J Black, and Gül Varol, "Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9488–9497.

[28] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2023.