# Appendix for "GENMO: A GENeralist Model for Human MOtion"

Jiefeng Li Jinkun Cao Haotian Zhang Davis Rempe Jan Kautz Umar Iqbal Ye Yuan
NVIDIA
https://research.nvidia.com/labs/dair/genmo

## A. Implementation Details

**Model Architecture.** GENMO comprises 16 layers, each consisting of a ROPE-based Transformer block followed by a multi-text injection block. The ROPE-based Transformer block incorporates a LayerNorm, a ROPE attention layer with residual connections, and an MLP layer. Each attention unit features 8 attention heads to capture diverse motion patterns. The number of neurons in the MLP layer is $d_{\text{mlp}} = 1024$. The multi-text injection block maintains a similar architecture to the ROPE-based Transformer block, but replaces the standard attention with multi-text attention, which processes text embedding sequences to enrich and update the motion feature representations. The maximum self-attention window size is $W = 120$.

**Training Datasets.** GENMO is trained from scratch on a diverse set of mixed motion datasets, including motion estimation datasets AMASS [14], BEDLAM [1], Human3.6M [5], 3DPW [26], music-to-dance dataset AIST++[11], and text-to-motion datasets HumanML3D [3] and Motion-X [12]. Since motion data in HumanML3D are represented in their own format, we convert them to SMPL parameters with inverse kinematics [9] for training. For AMASS data lacking video, music, or text inputs, we follow [20, 21] to simulate static and dynamic camera trajectories and project 3D motions to 2D keypoints as input conditions. The simulated camera trajectories are also used as input conditions during training. Although AMASS and HumanML3D share some motion sequences, we treat them as independent datasets.

For Motion-X, we only utilize its 2D keypoints and text descriptions due to noisy 3D ground truth. When training with BEDLAM and Human3.6M datasets, we use video frames and 2D keypoints as conditioning inputs, with global 3D motions serving as target outputs. For the 3DPW dataset, video frames and 2D keypoints are used as conditions; however, since 3DPW provides only local 3D motions, we implement a strategy analogous to $\mathcal{L}_{\text{gen-2D}}$: we first generate pseudo-clean global human trajectories from the estimation mode, then utilize these to produce noisy motions for training the generation mode, with loss computation restricted to local poses. For AIST++, training incorporates video frames, 2D keypoints, and music as conditions. Regarding the camera condition, we utilize ground-truth camera trajectories as the input condition for datasets that either provide such trajectories or feature static cameras; for datasets lacking labeled camera trajectories, we employ DROID-SLAM [23] to generate camera trajectories as input conditions during training. We train a single unified model on this comprehensive collection of datasets, enabling evaluation across diverse motion-related tasks.

**Condition Processing.** For video conditions, we employ a frozen encoder from TRAM [27], whereas for AMASS data lacking video inputs, we utilize zero vectors as placeholders. The 2D keypoint conditions undergo normalization to the range $[-1, 1]$ based on their bounding boxes, which are further normalized by the focal length of their corresponding video conditions. For music processing, we extract features using the music encoder from EDGE [25], while camera parameters are formulated as the camera-to-world transformation and derived from input videos via DROID-SLAM [23]. Textual descriptions are encoded through the T5 encoder architecture [18].

**Training Details.** During training, we employ data augmentation techniques on the 2D keypoints, including random masking and Gaussian noise perturbation to enhance model robustness. To further improve model robustness, we implement random masking of input conditions throughout the training process. We configure the sequence length to $N = 120$ for training, while maintaining support for variable sequence lengths during inference. The model is trained from scratch for 500 epochs using the AdamW optimizer [13], with a mini-batch size of 128 per GPU distributed across 2 A100 GPUs.

## B. Evaluation Settings for Music-to-Dance Generation

We evaluate the music-to-dance generation capabilities of GENMO on the AIST++ [11] dataset. The same one-in-all checkpoint is employed for evaluation as used in all other tasks. Following established protocols [11, 25], our evaluation encompasses four key aspects: motion quality, generation diversity, physical plausibility, and motion-music correlation.

For motion quality and generation diversity assessment, we compute the Fréchet Inception Distance (FID) [4] and the average feature distance of generated motions using both kinetic features [17] (denoted as "k") and geometric features [16] (denoted as "g") in accordance with Li *et al*. [11].

To evaluate physical plausibility, we employ two metrics: Mean Per Joint Position Error (MPJPE) and Procrustes-aligned MPJPE (PA-MPJPE). Additionally, we calculate the Physical Foot Contact score (PFC) as proposed by Tseng *et al*. [25].

For quantifying motion-music correlation, we utilize the Beat Alignment Score (BAS) following the methodology of Li *et al*. [22]. This metric effectively measures the synchronization between musical beats and motion transitions by calculating the average temporal distance between each kinematic beat and its nearest musical beat.

## C. Evaluation Settings for Text-to-Motion Generation

For evaluating text-to-motion generation on HumanML3D [3], we utilize the pre-trained text and motion encoders from [3] after converting our motion representation to the HumanML3D format. This conversion process involves first recovering the SMPL parameters from our raw representation and subsequently deriving the HumanML3D-format representation as described in [3], employing the neutral gender SMPL model. Consistent with established evaluation protocols, we report the variance across five different inference trials on HumanML3D. The same one-in-all checkpoint is employed for evaluation as used in all other tasks.

It is important to note that the conversion from SMPL to HumanML3D format introduces some degradation in motion quality, as the predicted SMPL bone lengths do not precisely match the HumanML3D skeleton, resulting in artifacts such as foot skating. To address this limitation and provide a more comprehensive evaluation, we additionally report the Fréchet Inception Distance (FID) and Diversity metrics using both kinetic features [17] (denoted as "k") and geometric features [16] (denoted as "g") based on 24 keypoints. Since the SMPL model and the HumanML3D skeleton share an identical joint order, this approach enables direct comparison of GENMO's motion quality with state-of-the-art methods using keypoint-based metrics.

Table 1. **Benchmark of Human Motion Generation.** Motion quality is evaluated on the 3DPW-XOCC [10] dataset.

| **Methods** | MPJPE ↑ | PA-MPJPE ↓ | PVE ↓ | ACCEL → |
|---|---|---|---|---|
| HybrIK [9] | 148.3 | 98.7 | 164.5 | 108.6 |
| PARE [8] | 114.2 | 67.7 | 133.0 | 90.7 |
| PARE [8] + VIBE [7] | 97.3 | 60.2 | 114.9 | 18.3 |
| NIKI (frame-based) [10] | 110.7 | 60.5 | 128.6 | 74.4 |
| NIKI (temporal) [10] | 88.9 | 52.1 | 98.0 | 17.3 |
| Ours (Regression-only) | 89.0 | 50.2 | 103.8 | 21.1 |
| Ours | **76.2** | **48.4** | **94.2** | **17.1** |

For the evaluation on Motion-X [12], we implemented our own text and motion encoders, as the original encoders were provided by [12] and their implementation details were not disclosed in the literature. Unlike HumanML3D, Motion-X text prompts lack frame-based keywords, necessitating a different approach to text encoding. We employed a pre-trained CLIP language model with its corresponding tokenizer to process the raw text prompts, generating embeddings with a dimension of 512, consistent with the representation used in [3]. The same one-in-all checkpoint is employed for evaluation as used in all other tasks. For evaluation purposes on the Motion-X dataset, we utilized these trained encoders with frozen weights to ensure consistent and comparable feature extraction across all test samples.

## D. Evaluation Settings for Motion In-betweening

For motion in-betweening evaluation, we adopt the methodology established in prior diffusion-based approaches [24], wherein the noisy motion is overwritten with desired poses at specified keyframes prior to each denoising step. The same one-in-all checkpoint is employed for evaluation as used in all other tasks. Due to the constraints of our feature representation, which lacks global root information, we only overwrite the local body poses and global root orientation for the keyframes. We evaluate our approach on both the HumanML3D and Motion-X test sets under two experimental conditions: sampling either 2 or 5 keyframes from each test motion. For Motion-X, we utilize the reconstructed 3D motion as described in [12]. Additionally, we incorporate textual descriptions from these datasets as conditioning input. To account for the generative diversity of our model, we sample $N = 10$ different initial noise vectors for each test motion and execute the diffusion process with 50 denoising steps. For evaluation metrics, we report the minimum values among these diverse samples, which effectively captures the best performance achievable by our generative approach.

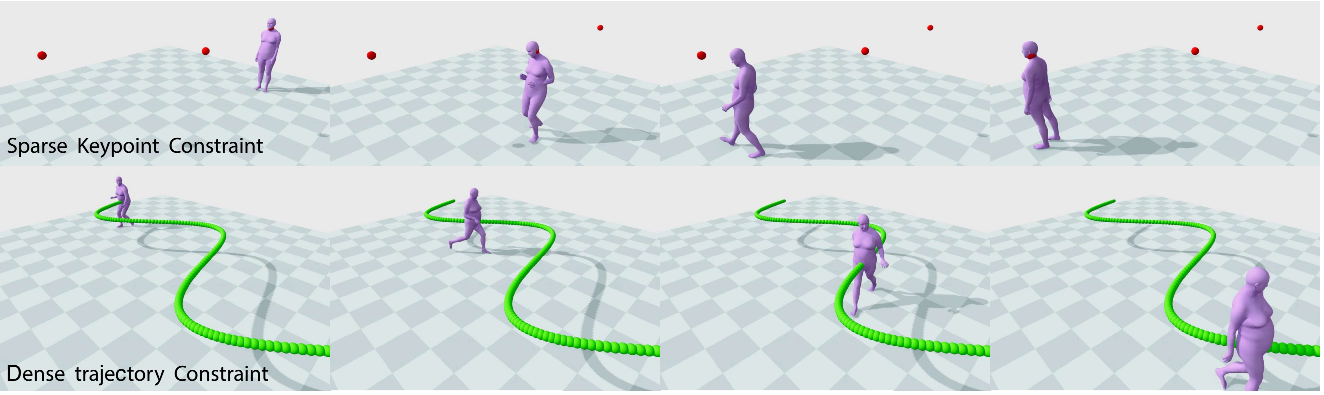

Figure 1. **Motion generation with spatial constraints.** Top: sparse head position control. Bottom: dense pelvis trajectory control.

## E. Motion Generation with Spatial Constraints

Similar to other state-of-the-art motion generation approaches, GENMO is capable of synthesizing human motions under explicit spatial constraints. Leveraging the flexibility of the diffusion-based framework, we employ classifier-free guidance to effectively control the generated motions without necessitating additional retraining. Representative examples illustrating spatially constrained motion generation are presented in Figure 1.

## F. Evaluation on Occlusion-Specific Benchmark

To evaluate the efficacy of generative priors in enhancing motion estimation robustness, we conducted comprehensive experiments on the 3DPW-XOCC benchmark [10]. This benchmark specifically evaluates 3D human pose estimation under challenging conditions of extreme occlusion and truncation, simulated through strategic placement of random occlusion patches and frame truncations. As evidenced in Table 1, GENMO demonstrates superior performance compared to state-of-the-art human motion estimation methods, including those explicitly designed to handle occlusions. Notably, our ablation study reveals that a variant of our model trained without generative tasks exhibits worse performance compared to the complete GENMO model. These findings substantiate that the generative priors incorporated within GENMO significantly enhance the plausibility and accuracy of estimated human motions under visually challenging scenarios, thereby underscoring the practical utility of our approach in real-world applications where occlusions frequently occur.

Table 2. **Ablation Study.** Camera-space motion estimation on the 3DPW [26] dataset.

| **Methods** | PA-MPJPE ↑ | MPJPE ↓ | PVE ↓ | ACCEL → |
|---|---|---|---|---|
| Cross-attention Fusion | 50.2 | 80.8 | 98.2 | 8.9 |
| MLP Fusion | **34.6** | **53.9** | **65.8** | **5.2** |

## G. Additional Ablation Study

### G.1. MLP Fusion vs. Cross-Attention Fusion

Cross-attention fusion is a widely adopted technique for integrating text and motion features. While it is straightforward to extend cross-attention modules to fuse video-based conditions, we observe that MLP-based fusion offers superior temporal alignment between modalities. To empirically validate this, we compare the camera-space motion estimation performance of MLP fusion and cross-attention fusion in Table 2. The results demonstrate that effective temporal alignment is critical for achieving accurate motion estimation. Furthermore, we observe that models employing cross-attention fusion require more training iterations to converge and exhibit less stable training loss compared to their MLP fusion counterparts.

## H. Additional Related Work

### H.1. Generative Priors for Estimation

Recent advances in computer vision have demonstrated the efficacy of leveraging generative priors from large-scale image models, such as StableDiffusion [19], for various estimation tasks. These approaches fine-tune diffusion-based generative models to predict geometric and semantic properties, including depth maps, surface normals, and semantic segmentation [2, 6, 15]. By repurposing the rich latent representations encoded in pre-trained generative models, these methods achieve substantial improvements in estimation accuracy across diverse visual understanding tasks.

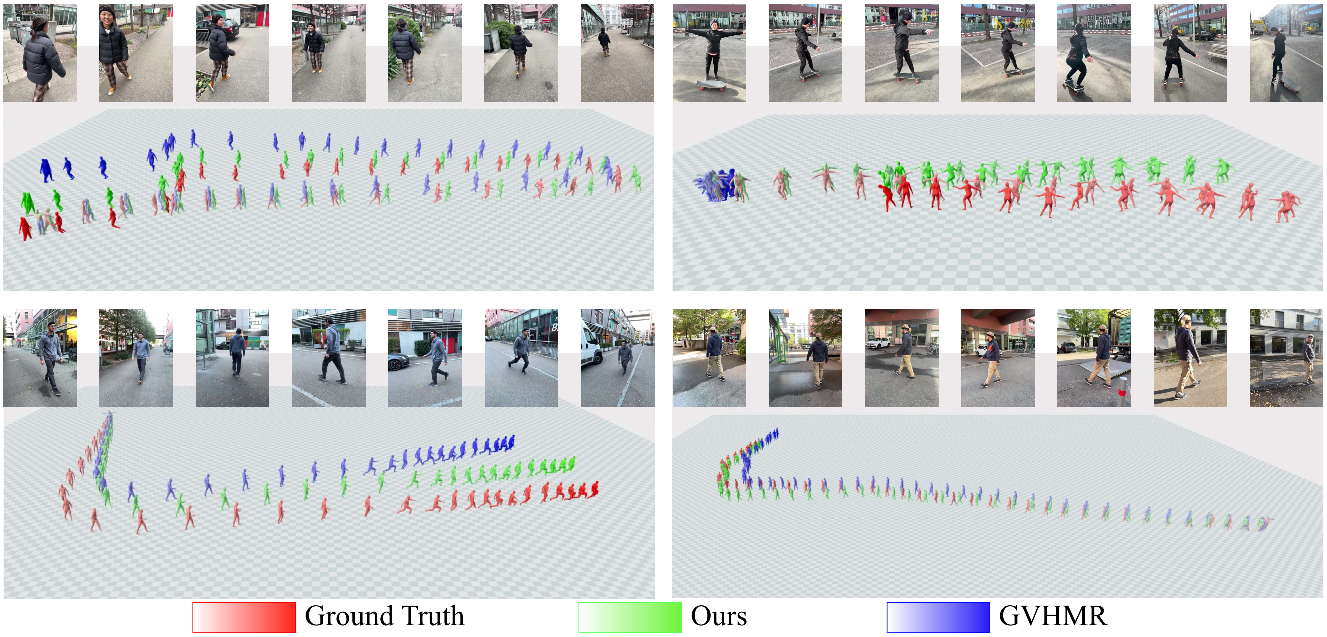

Figure 2. **Qualitative comparisons on motion estimation.**

Nevertheless, a significant limitation of these approaches is their tendency to sacrifice the inherent generative capabilities of the original models, as they predominantly focus on deterministic estimation outcomes rather than maintaining the ability to produce diverse outputs.

Our work fundamentally diverges from these approaches by introducing a unified framework that seamlessly integrates motion generation and estimation within a single coherent model. In contrast to previous methods that compromise generative capabilities during the fine-tuning process, our framework maintains both the stochastic diversity essential for high-quality generation and the deterministic precision required for accurate estimation. This dual capability represents a significant advancement in leveraging generative priors for human motion understanding.

## I. Limitations and Failure Cases

At present, GENMO depends on off-the-shelf SLAM algorithms to estimate camera parameters for video-conditioned motion generation. Consequently, failures in SLAM tracking cannot be rectified by GENMO, resulting in inconsistencies between camera-space and world-space motions. Furthermore, in scenarios lacking explicit foot-ground contact between the human subject and the environment, the fidelity of the generated motions becomes highly contingent on the reliability of SLAM tracking, which is prone to failure in such cases.

## References

[1] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8726–8737, 2023. 1

[2] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258. Springer, 2024. 3

[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 (CVPR)*, pages 5152–5161, 2022. 1, 2

[4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 2

[5] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2014. 1

[6] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024. 3

[7] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 2

[8] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *ICCV*, 2021. 2

[9] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*, 2021. 1, 2

[10] Jiefeng Li, Siyuan Bian, Qi Liu, Jiasheng Tang, Fan Wang, and Cewu Lu. Niki: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12933–12942, 2023. 2, 3

[11] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, 2021. 1, 2

[12] 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. In *NeurIPS*, 2023. 1, 2

[13] Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5, 2017. 1

[14] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 2019. 1

[15] Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan de Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. In *WACV*, 2025. 3

[16] Meinard Müller, Tido Röder, and Michael Clausen. Efficient content-based retrieval of motion capture data. In *ACM SIGGRAPH 2005 Papers*, pages 677–685. 2005. 2

[17] Kensuke Onuma, Christos Faloutsos, and Jessica K Hodgins. Fmdistance: A fast and effective distance function for motion capture data. *Eurographics (Short Papers)*, 7(10), 2008. 2

[18] 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. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 1

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[20] 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 2024 Conference Papers*, pages 1–11, 2024. 1

[21] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *CVPR*, 2024. 1

[22] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11050–11059, 2022. 2

[23] Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. *NeurIPs*, 2021. 1

[24] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. In *ICLR*, 2023. 2

[25] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 448–458, 2023. 1, 2

[26] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 1, 3

[27] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *European Conference on Computer Vision*, pages 467–487. Springer, 2024. 1