

# Supplementary Material for HumanDreamer: Generating Controllable Human-Motion Videos via Decoupled Generation

Boyuan Wang<sup>1, 2, 3\*</sup>, Xiaofeng Wang<sup>1, 2, 3\*</sup>, Chaojun Ni<sup>4, 5</sup>, Guosheng Zhao<sup>1, 2, 3</sup>, Zhiqin Yang<sup>6</sup>, Zheng Zhu<sup>4†</sup>,  
Muyang Zhang<sup>1, 2</sup>, Yukun Zhou<sup>4</sup>, Xinze Chen<sup>4</sup>, Guan Huang<sup>4</sup>, Lihong Liu<sup>1</sup>, Xingang Wang<sup>1, 3†</sup>

<sup>1</sup>Institute of Automation, Chinese Academy of Sciences, China

<sup>2</sup>School of Artificial Intelligence, University of Chinese Academy of Sciences, China

<sup>3</sup>Luoyang Institute for Robot and Intelligent Equipment, China

<sup>4</sup>GigaAI, China <sup>5</sup>Peking University, China <sup>6</sup>The Chinese University of Hong Kong

In the supplementary material, we begin by elaborating on the implementation details of the filter and the model used in *HumanDreamer*, then provide a detailed description of our proposed *MotionVid* dataset, and finally present further quantitative comparison results.

## 1. Implementation Details

In this section, we detail the specific calculation methods and filtering criteria employed in the *Video Quality Filter* and *Human Quality Filter* within our work. Subsequently, we provide an in-depth elaboration on the implementation of PoseVAE, the pipeline of *Pose-to-Video*, and the compositional specifics of the MotionVid dataset.

### 1.1. Details in Video Quality Filter

Below, we introduce the specific calculation methods and corresponding thresholds for the four filtering criteria used in the *Video Quality Filter*.

**Movement Intensity.** To assess the dynamic nature of the videos, we utilize the GMFlow method [26] for estimating optical flow. The purpose is to filter out videos with insufficient movement, which may not be engaging or informative. The movement intensity is defined as follows:

$$S_{\text{Move}} = \frac{1}{T-1} \sum_{t=1}^{T-1} \|\mathcal{M}(\mathbf{I}_t, \mathbf{I}_{t+1})\|_{\text{avg}}, \quad (1)$$

where  $T$  is the total number of frames,  $\{\mathbf{I}_t\}_{t=1}^T$  denotes the sequence of input images over time,  $\mathcal{M}(\cdot, \cdot)$  represents the model-based optical flow prediction function, and  $\|\cdot\|_{\text{avg}}$  indicates the average magnitude of the optical flow across all pixels. The movement intensity is computed as the average of the optical flow magnitudes over consecutive frames, providing a quantitative measure of the motion within the video. Videos satisfied  $S_{\text{Move}} \leq 0.5$  are discarded to ensure

that the dataset consists of content with sufficient dynamic activity.

**Text Coverage.** To ensure the quality and readability of video content, we adopt the methodology outlined in [1] for detecting text regions within frames. Following this detection, we calculate the area of each text bounding box, denoted as  $S_{\text{text}}$ , and compare it against the total area of the frame, represented as  $S_{\text{frame}}$ . Videos are excluded from further processing if the condition  $S_{\text{text}} > 0.07 \times S_{\text{frame}}$  is met.

**Aesthetic Score.** To evaluate the aesthetic quality of the videos, we employ LAION-AI’s aesthetic predictor [19] to compute aesthetic scores. Videos with an aesthetic score  $S_{\text{Aes}}$  that does not satisfy  $S_{\text{Aes}} \geq 4$  are eliminated from the dataset.

**Blur Intensity.** To evaluate the sharpness of the videos, we apply the Laplacian operator [2] to measure the blur intensity. The objective is to discard videos that exhibit excessive blurring, as such videos can detract from the visual quality and clarity. The blur intensity is defined as:

$$S_{\text{Blur}} = \frac{1}{T} \sum_{i=1}^T \text{Var}(\mathcal{L}(\text{Gray}(\mathbf{I}_i))), \quad (2)$$

where  $\text{Gray}(\cdot)$  denotes the conversion of an RGB image to a grayscale image,  $\mathcal{L}(\cdot)$  represents the computation of the Laplacian transform, and  $\text{Var}(\cdot)$  indicates the calculation of the variance. The blur intensity  $S_{\text{Blur}}$  is computed as the average variance of the Laplacian-transformed grayscale images across all frames. Videos with a blur intensity  $S_{\text{Blur}} \leq 20$  are discarded to ensure that the dataset contains only high-quality, clear visuals.

### 1.2. Details in Human Quality Filter.

Below, we introduce the specific calculation methods and corresponding thresholds for the four filtering criteria used in the *Human Quality Filter*.

**Motion Magnitude.** To filter out sequences with insufficient motion, we calculate the difference between the 2D poses of two consecutive frames. Specifically, we compute the average difference in body keypoints between adjacent frames. The motion magnitude is defined as:

$$Mag_{\text{mot}} = \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{1}{N} \sum_{i=1}^N \|\mathbf{k}_i^t - \mathbf{k}_i^{t+1}\|, \quad (3)$$

where  $N$  is the number of body keypoints, and  $\mathbf{k}_i^t$  represents the position of the  $i$ -th keypoint in the  $t$ -th frame. Videos satisfied  $Mag_{\text{mot}} \leq 10^{-3}$  are discarded to ensure that the dataset contains sequences with sufficient dynamic movement.

**Human Coverage.** To ensure that videos contain a significant presence of human subjects, we compute the ratio of the human detection bounding box area to the entire frame area, similar to the method used for text coverage. Videos with a human coverage ratio less than  $1/3$  are removed from the dataset.

**Human Count.** To ensure that the videos focus on individual human subjects, we uniformly sample 5 frames from each video and count the number of detected humans in each frame. Videos are discarded if the number of humans detected in any of the sampled frames exceeds 1.

**Face Visibility.** To ensure face visibility for training purposes, we uniformly sample 5 frames from each video. For each frame, we check the presence of 5 facial keypoints (eyes, ears, nose). If all 5 keypoints are detected in a frame, the face is considered visible. Videos are discarded if the face is not visible in any of the 5 sampled frames.

### 1.3. Details in CLoP.

CLoP consists of two versions: one trained on a subset to filter large-scale data in *Caption Quality Filter*, and another retrained on the fully filtered dataset for training and evaluating MotionDiT, similar to [6, 21, 28]. CLoP is not used during MotionDiT inference.

### 1.4. Details in PoseVAE.

The pose sequence  $\mathbf{p} \in \mathbb{R}^{f \times N \times 3}$ , consisting of coordinates and confidence scores, is input into a Variational Autoencoder (VAE) for reconstruction. The encoder of the VAE extracts spatial features through three layers of ResNet1D blocks and downsampling operations, which reduce spatial dimensions. This process yields a latent distribution parameterized by the mean  $\mu$  and variance  $\sigma^2$ . Using the reparameterization trick, a latent representation  $\mathbf{z} \in \mathbb{R}^{f \cdot N/8 \cdot 4}$  is sampled from this distribution. Here,  $N/8$  reflects three rounds of downsampling, each reducing the resolution by a factor of 2, while 4 denotes the number of channels in the latent space.

The decoder reconstructs the input sequence using three layers of ResNet1D[31] blocks that capture spatiotemporal features, combined with upsampling operations. This reconstruction process outputs  $\mathbf{p}_r \in \mathbb{R}^{f \times N \times 3}$ . The overall architecture draws inspiration from the VAE framework proposed by [3].

The VAE loss function,  $L_{\text{VAE}}$ , consists of a reconstruction loss  $L_R$  and a KL divergence term  $L_{\text{KL}}$ , formulated as follows:

$$L_{\text{VAE}} = L_R + \beta L_{\text{KL}}, \quad (4)$$

where  $\beta = 10^{-7}$ . The reconstruction loss is defined as

$$L_R = \|\mathbf{p} - \mathbf{p}_r\|_2^2, \quad (5)$$

and the KL divergence loss is expressed as

$$L_{\text{KL}} = \frac{1}{2} \sum_{i=1}^k (\sigma_i^2 + \mu_i^2 - \log(\sigma_i^2) - 1), \quad (6)$$

where  $k$  is the dimensionality of the latent space,  $\mu$  and  $\sigma^2$  denote the mean and variance of the latent variables' distribution, respectively. The KL divergence measures the difference between this distribution and a standard normal distribution  $\mathcal{N}(0, 1)$ , which serves as the prior. The specific architecture of the PoseVAE is illustrated in Fig. 1.

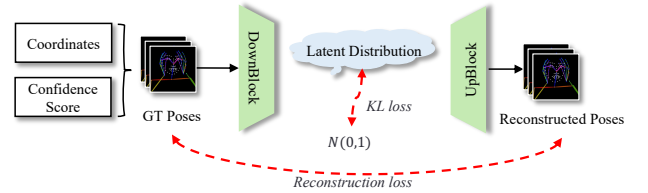


Figure 1. Structure of Pose VAE.

### 1.5. Details in Pose-to-Video.

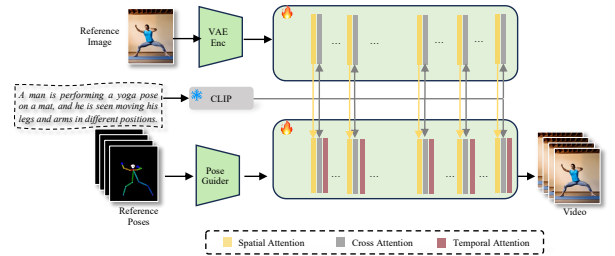


Figure 2. Pipeline of Pose-to-Video.

The structure of the *Pose-to-Video* model is shown in Fig. 2. The architecture is inspired by the work in [11] and utilizes the backbone from [27], which consists of stacked spatial and temporal attention layers. Textual inputs are

processed through CLIP[18] to obtain text features, while reference poses are provided in image form to guide the generation process. The initial frame of the person can be generated from a prompt or manually specified. In our approach, we utilize SD1.5 [20] combined with ControlNet [29]. More advanced text-to-image models could potentially enhance alignment further. The VAE is used to encode the input conditions into a latent representation, which is then integrated into the model via cross-attention mechanisms inspired by [29].

This design ensures that the generated videos are coherent and aligned with both the pose and textual inputs, leveraging advanced attention mechanisms to capture spatial and temporal dependencies effectively.

### 1.6. Details in Evaluation Metrics.

**FID.** In evaluating the overall quality of generated samples, the Fréchet Inception Distance (FID) [10] is widely used. It measures the similarity between the feature distributions of real and generated data. Specifically,  $\mu_{gt}$  and  $\mu_{pred}$  represent the means of the feature vectors for the ground truth and predicted data, respectively,  $\Sigma$  denotes the covariance matrix, and  $Tr(\cdot)$  stands for the trace of a matrix. Then, FID is calculated as follows:

$$FID = \|\mu_{gt} - \mu_{pred}\|^2 - \text{Tr}(\Sigma_{gt} + \Sigma_{pred} - 2(\Sigma_{gt}\Sigma_{pred})^{\frac{1}{2}}) \quad (7)$$

**R-precision.** R-precision is a metric used to evaluate the accuracy of matching between text descriptions and generated motions. It calculates the proportion of relevant items (motions) retrieved in the top-k results relative to the total number of relevant items. Specifically, it measures how many of the top-k motions correctly match their corresponding texts.

**Diversity.** Diversity assesses the variation in motion sequences throughout the dataset. In our experiments, we randomly sample  $S_{dis}$  pairs of motions, setting  $S_{dis}$  to 300 in our experiments. Each pair's feature vectors are denoted as  $f_{pred,i}$  and  $f'_{pred,i}$ . Diversity is then calculated by:

$$\text{Diversity} = \frac{1}{S_{dis}} \sum_{i=1}^{S_{dis}} \|f_{pred,i} - f'_{pred,i}\| \quad (8)$$

**MultiModality.** MM assesses the diversity of human motions generated based on the same text description. More precisely, for the  $i$ -th text description, 32 motion samples are generated, and a total of 100 text descriptions are used. The features of each motion sample are extracted using CLoP. The feature vectors of the  $j$ -th pair derived from the  $i$ -th text description are represented as  $(f_{pred,i,j}, f'_{pred,i,j})$ . The definition of MM is given as follows:

$$MM = \frac{1}{32N} \sum_{i=1}^N \sum_{j=1}^{32} \|f_{pred,i,j} - f'_{pred,i,j}\| \quad (9)$$

**MultiModality Distance.** MM Dist measures the feature-level distance between the text embedding and the generated motion feature. The features of the  $i$ -th text-motion pair are  $f_{pred,i}$  and  $f_{text,i}$ . Then, MM-Dist is defined as follow:

$$MM \text{ Dist} = \frac{1}{N} \sum_{i=1}^N \|f_{pred,i} - f_{text,i}\| \quad (10)$$

## 2. Dataset Details

MotionVid comprises 1.27M text-pose-video pairs, with 66.5% originating from public datasets and 33.5% sourced from the internet, as detailed in Tab. 1. This diverse composition reflects a wide variety of styles, encompassing general, action-specific, and domain-focused clips (e.g., facial and hand actions). Notably, datasets like Panda-70M[5] and Kinetics-700[4] contribute significantly to the collection, ensuring robust coverage of both general and specialized motion types. Such diversity enhances the dataset's utility for training models capable of handling heterogeneous real-world scenarios. Additionally, the inclusion of curated internet data complements the public datasets, providing more nuanced and potentially underrepresented motion patterns.

The evaluation dataset, extracted from *MotionVid* with 1000 samples, shows verb frequency in Fig.3 after removing common verbs, indicating diverse actions. Comparisons in Tab.2 show our R-precision is comparable to HumanML3D, ensuring a reasonable distribution, while *Diversity* is higher, reflecting a broader range of actions and poses. The evaluation dataset's distribution mirrors the whole dataset, which includes hundreds of action types from sources like ActivityNet200, Kinetics700, and internet data, enhancing diversity.

## 3. Experiment Results

Additional visualizations are presented to demonstrate the advancements in *Text-to-Pose* and *Pose-to-Video*, showcasing the improvements in the quality of generated videos.

### 3.1. Comparison of Text-to-Pose

We further used the poses generated by different *Text-to-Pose* methods to synthesize videos, comparing the quality of the resulting human-centric videos. The results of this comparison can be found in the folder *supplement/video\_in\_supplement/compare\_text\_to\_pose*, specifically in the files *Demo1.mp4* and *Demo2.mp4*.

Specifically, we employed four different models—T2M-GPT[28], PriorMDM[21], MLD[6], and *MotionDiT*—to generate pose sequences from textual input. Subsequently, these generated poses were utilized to produce video outputs. The results indicate that our proposed method is capable of generating more stable and semantically coherent

Table 1. The table presents the specific composition of *MotionVid*, including the sources from which it was collected, the names of the datasets, the number of clips after **video quality filter (VQF)**, the number of clips after **human quality filter (HQF)** and **caption filter (CF)**, and the data types. It shows that *MotionVid* includes a diverse range of data categories, including general, action, and actions specific to different body parts, indicating a high degree of diversity.

Source	Dataset	After VQF	After HQF+CF	Data Type
Public	Panda-70M[5](partial)	2,139,180	704,210	General
Public	Kinetics-700[4]	562,734	68,316	Action
Public	Kinetics-400[13]	298,337	30,855	Action
Public	Motion-X[17]	30,554	15,494	Action
Public	ActivityNet-200[8]	91,220	7,955	Action
Public	DFEW[12]	15,410	6,487	Facial Action
Public	CAER[15]	12,932	2,912	Facial Action
Public	UBody[16]	5,981	2,796	Action
Public	HAA500[7]	8,747	2,133	Action
Public	HMDB51[14]	4,678	1,604	Action
Public	Something-Something V2[9]	177,055	877	Hand Action
Public	Charades[22]	10,447	822	Action
Public	Charades-Ego[23]	8,845	478	Action
Internet	-	1,686,614	425,382	General
<b>Total</b>	-	<b>5,052,734</b>	<b>1,270,321</b>	-

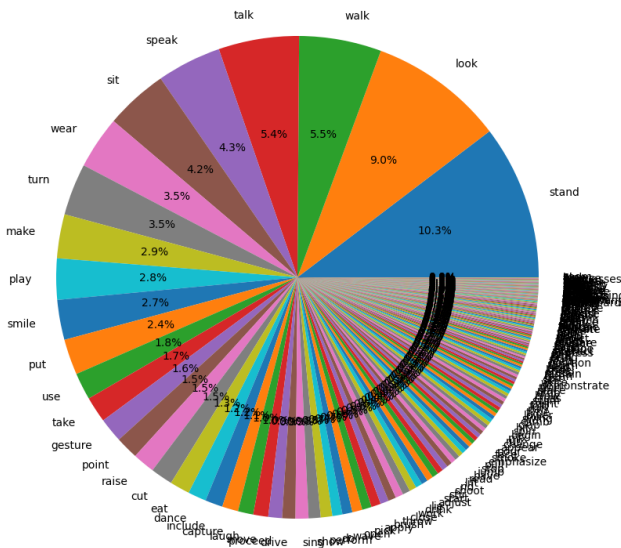


Figure 3. Distribution of verbs in MotionVid’s eval set.

Table 2. Statistics of MotionVid’s eval set and HumanML3D.

Dataset	Rp-top1 $\uparrow$	Rp-top2 $\uparrow$	Rp-top3 $\uparrow$	Diversity $\uparrow$
HumanML3D	0.424	0.649	0.779	11.08
MotionVid	0.450	0.639	0.744	70.11

poses, which are essential for the creation of high-quality human-centric videos.

### 3.2. Comparision of Pose-to-Video

We used the same reference image and pose sequences, but changed the models in the Pose-to-Video genera-

Table 3. Evaluation of *Pose-to-Video*.

Method	LPIPS ↓	FVD ↓
AnimateAnyone	0.285	171.90
MimicMotion	0.414	232.95
Animate-X	0.232	139.01
Our <i>Pose-to-Video</i>	<b>0.148</b>	<b>116.74</b>

tion process. Specifically, we compared the video generation results using our proposed method, as well as AnimateAnyone[11] and MusePose[25]. The visualization results of this comparison can be found in the folder *supplement/video\_in\_supplement/compare\_pose\_to\_video*, specifically in the files Demo1.mp4, Demo2.mp4, etc. The results show that our proposed model achieves the best visual outcomes in video generation. We provide quantitative comparisons of our *Pose-to-Video* with [11, 24, 30] under their experimental settings, with results summarized in Tab. 3. Our *Pose-to-Video* demonstrates strong performance in consistency and visual quality.

### 3.3. Comparision of Text-to-Video

Compared to CogVideoX, HumanDreameer excels in *Sensory Quality* and *Instruction Following* (CogVideoX’s metrics), as confirmed by the user study on the MotionVid evaluation set (Tab. 4). Additionally, the *Diversity* calculated from poses extracted from generated videos, shows our method outperforms CogVideoX.



Table 4. Evaluation between HumanDreamer and CogVideoX-5B.

Method	Sensory Quality $\uparrow$	Instruction Following $\uparrow$	Diversity $\uparrow$
CogVideoX-5B	0.531	0.688	25.285
HumanDreamer	<b>0.938</b>	<b>0.813</b>	<b>68.220</b>

## References

- [1] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *CVPR*, 2019. 1
- [2] Raghav Bansal, Gaurav Raj, and Tanupriya Choudhury. Blur image detection using laplacian operator and open-cv. In *SMART*, 2016. 1
- [3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2
- [4] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset, 2022. 3, 4
- [5] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers, 2024. 3, 4
- [6] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *CVPR*, 2023. 2, 3
- [7] Jihoon Chung, Cheng hsin Wu, Hsuan ru Yang, Yu-Wing Tai, and Chi-Keung Tang. Haa500: Human-centric atomic action dataset with curated videos. In *ICCV*, 2011. 4
- [8] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 4
- [9] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The” something something” video database for learning and evaluating visual common sense. In *ICCV*, 2017. 4
- [10] 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. *NeurIPS*, 2017. 3
- [11] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, 2024. 2, 4
- [12] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *ACM MM*, 2020. 4
- [13] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. 4
- [14] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *ICCV*, 2011. 4
- [15] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoonn Sohn. Context-aware emotion recognition networks. In *ICCV*, 2019. 4
- [16] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. *CVPR*, 2023. 4
- [17] 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. *NeurIPS*, 2024. 4
- [18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 3
- [19] Beaumont Romain and Schuhmann Christoph. Laion aesthetics predictor v1. 2022. 1
- [20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3
- [21] Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior. In *ICLR*, 2024. 2, 3
- [22] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding, 2016. 4
- [23] Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. In *ArXiv*, 2018. 4
- [24] Shuai Tan, Biao Gong, Xiang Wang, Shiwei Zhang, Dandan Zheng, Ruobing Zheng, Kecheng Zheng, Jingdong Chen, and Ming Yang. Animate-x: Universal character image animation with enhanced motion representation. *arXiv preprint arXiv:2410.10306*, 2024. 4
- [25] Zhengyan Tong, Chao Li, Zhaokang Chen, Bin Wu, and Wenjiang Zhou. Musepose: a pose-driven image-to-video framework for virtual human generation. *arxiv*, 2024. 4
- [26] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezaatofghi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8121–8130, 2022. 1
- [27] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2
- [28] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *CVPR*, 2023. 2, 3
- [29] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 3

- [30] Yuang Zhang, Jiayi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024. [4](#)
- [31] Fuyu Zhu, Hua Wang, and Yixuan Zhang. Gru deep residual network for time series classification. In *2023 IEEE 6th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, pages 1289–1293, 2023. [2](#)