Text-Based Video Generation With Human Motion and Controllable Camera

Anonymous ICCV submission

Paper ID 32

Abstract

As expectations for generative models have risen recently, Text-to-Video models have been actively studied. Existing Text-to-Video models have limitations in that it is difficult to generate complex movements such as human motions. Then often generate unintended human motions and the scale of the subject. In order to improve the quality of videos that include human motion, we propose a three-stage framework. In the first stage, Text-driven Human Motion Generation network generates 3D human motion from input text prompt. In the second stage, 3D human motion sequence is projected to a 2D skeleton format. In the third stage, and then Skeleton-Guided Text-to-Video Generation module generates a video where the motion of subject is well represented. In addition, we can manipulate the camera view point and angle to generate a video we want, since the human motion generated in the first stage is 3D, not, 2D. We demonstrated that the proposed framework outperforms the existing Text-to-Video models in quantitative and qualitative manners. To the best of our knowledge, the our framework is the first methods using Text-driven Human Motion Generation networks to improve video with human motions. Our Project pages are available in `https://anonymous.4open.science/w/HMTV_docs-1801/`.

1. Introduction

Nowadays, Text-to-Image model (T2I) that generates images using a given text prompt is being actively studied. In particular, models such as Stable Diffusion [1] and DALL-E2 [2] are attracting more attention for their outstanding performance. Alongside the growth of T2I, Text-to-Video model (T2V), which generates the corresponding video with a given text prompt is also developed.

Seminal research on T2V has gained momentum with the diffusion based models such as Dreamix [3], VDM [4], ImagenVideo [5], and Make-A-Video [6]. However, they are facing some difficulties. First, complex movements such as human motions are generated with a degree of awkwardness. To solve this problem, Skeleton-Guided Text-to-Video Generation [7, 8], which conditioned on human skeleton, enables pose control of the subject. However, it is difficult to use for various applications because not only a text condition but also a human skeleton is required. Second, even with the given human skeleton, undesired results are generated. For example in Fig. 1, with certain human skeleton condition the generated outputs have limited spectrum of views such as only the back side of a person. Since the model does not know which direction they are looking at, it is trivial to get these results. If we use models that do not use a skeleton as a guidance, the quality of a generated human is an issue. In the result without skeleton guidance in the top row of fig. 1, the video of the person playing golf has very poor frame consistency, and the video of the person punching is very large due to the inability to scale the subject.

On the other hand, as various generation models have been actively studied recently, Human Motion Generation is also attracting a lot of attention. Early methods of Human Motion Generation use human motion prediction [9, 10, 11] that predict the next actions based on previous actions and generating in-between motion [12, 13]. Recently, Text-driven Human Motion Generation, which generates 3D human motion sequences from text prompts, has been studied, opening the possibility of countless expansion of Human Motion Generation. For example, MDM [14], MotionDiffuse [15] and T2M-GPT [16] are one of those. In particular, T2M-GPT [16] is expected to be highly applicable as it can generate complex movements with long sentences.

In this paper, we propose a novel video generation algorithm that naturally generates human movements by combining Text-driven Human Motion Generation and Skeleton-Guided Text-to-Video Generation. In particular, a text prompt as an input results a high-quality 3D human motion to guide video generation model. Next, the generated consecutive 3D humans motions are converted to 2D skeletons. In this step, an additional camera prompt is given which is paired with a predefined camera extrinsic parameter. Lastly, the input text and the 2D human skeleton sequence from generated 3D motion are used to generate a

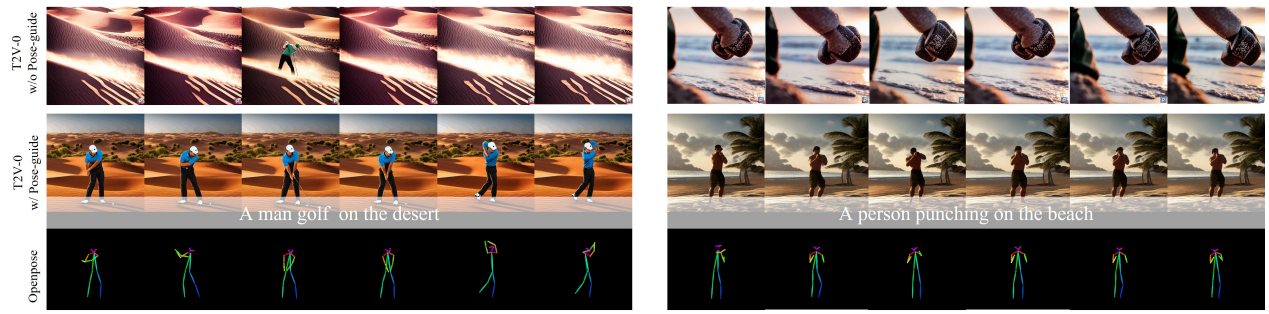

Figure 1. This figure shows the difference of output video between with using pose guidance and without using it to T2V network. Text prompts are applied to both. Network without pose guidance show problems with size of human and has inconsistency between consecutive frames, but using pose guidance these inconsistencies do not happen.

video with a high quality human movements.

Note that we not only improve the generation quality of video containing 'human motion', but also control the viewing angle and the movement of the camera. As far as we know, the first framework which can control the camera composition and the scale of the subject as desired. Furthermore, techniques used in actual film shooting such as Tilt up&down, zoom in&out, and dolly in&out can be applied to video generation, and the possibility of being used in various applications such as the contents industry is unlimited.

In summary, our contributions are:

- We propose a framework that combines the Text-driven Human Motion Generation and the Skeleton-Guided Text-to-Video Generation module to generate a high quality video expressing dynamic scenes with complex human behavior by a text based camera control.
- Our Text-to-Video methods outperform both quantitative and qualitative results than previous methods. Moreover actual film shooting techniques can be applied to video generation for various applications.

2. Proposed Method

In this section, we overview the proposed framework which is shown in Fig. 2. Our method aims to enhance a quality and diversity of generated videos with human motions inside and consists of three stages.

2.1. Text-to-Human Motion Generation

Text-to-Motion Generation stage uses predefined Text-to-Motion (T2M) network that generates sequential 3D human motion. Formally, given a input text $\mathcal{P}$, T2M network $F(\mathcal{P};\theta)$ generates a set of vertices $\{V_i^{3D}\}_{i=1}^{K}$ which form meshes of a human formulated as below

$$F(\mathcal{P};\theta) = \{V_1^{3D}, \cdots, V_K^{3D}\}, \tag{1}$$

where θ is a model parameter of T2M network and K is the number of vertices consisting meshes. In this stage, various kinds of T2M network can be applied. We use a T2M-GPT [16] that encodes motions using VQ-VAE [17]

2.2. Camera Projection Module

In this stage, we will introduce the Camera Projection Module (CPM) shown in the bottom side of Fig. 2. This module can takes a preset text description of a camera direction $\mathcal{P}_{\text{Camera}}$ as an input and output corresponding projected 2D skeletons. This module consists of three parts. First part is 3D skeleton regression. This step takes 3D mesh vertices from text-to motion network and uses joint regressor from [18] to regress joints from the mesh vertices. We can formulate this stage as below where $V_i^{3D} \in \mathbb{R}^3$ denotes the i^{th} vertex of mesh, $J_i^{3D} \in \mathbb{R}^3$ denotes the i^{th} joints regressed from the mesh and J_{reg} is the joint regression matrix.

$$J_i^{3D} = J_{\text{reg}} V_i^{3D} \tag{2}$$

Second part is a control of camera position using camera prompt. We can express a rotation and a translation with a camera extrinsic matrix using the homogeneous coordinate denote as below

$$\begin{pmatrix} R_{3\times3} & t_{3\times1} \\ 0_{1\times3} & 1_{1\times1} \end{pmatrix}. \tag{3}$$

Note that $R_{3\times3}$ defines the rotation of a camera and $t_{3\times1}$ defines the translation of the camera. With intrinsic matrix together we can define a projection matrix P_{proj} as below

$$P_{proj} = \begin{pmatrix} f_x & 0 & c_x & 0 \\ 0 & f_y & c_y & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} R_{3\times3} & t_{3\times1} \\ 0_{1\times3} & 1_{1\times1} \end{pmatrix}_{4\times4} \tag{4}$$

We pre-define the textual descriptions and corresponding directions, and CPM uses the lookup table to decide a position of camera. The final part is a 2D projection with the camera rotation and translation matrices. With determined P_{proj}, we can project 3D skeleton to 2D space using a homogeneous coordinate system:

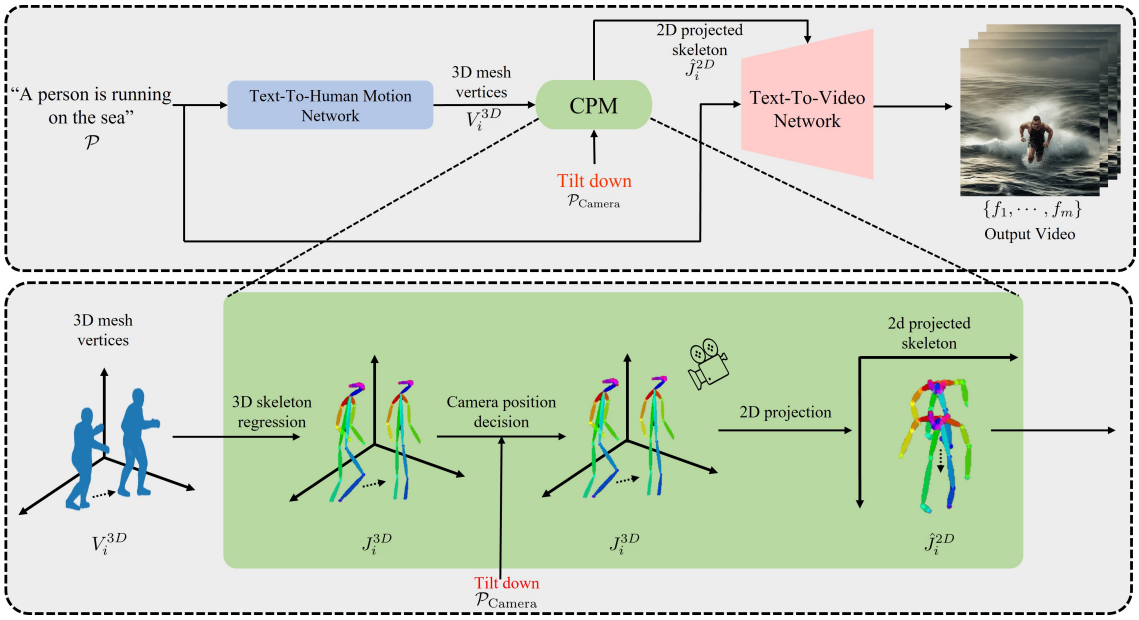

Figure 2. **Overall process of our proposed framework. Top**: Our framework consists of three stages: (1) Text-to-Motion Generation, (2) Camera Projection Module, (3) Skeleton Guided Text-to-Video Generation. A text prompt is passed to the Text-to-Human Motion Generation network to generate 3D mesh vertices of each frames of motion. Then, with camera direction description prompt, Camera Projection Module (CPM) convert these vertices to the skeletons and project to 2D space corresponding to the camera direction prompt. The last stage, (3) Skeleton Guided Text-to-Video Generation, we use Text-to-Video network with 2D projected skeletons from CPM and generates the output video corresponds to input prompt $\mathcal{P}$. **Bottom**: CPM module in detail. CPM takes 3D vertices of mesh and regress the 3D skeleton with joint regressor. And, decide camera position and direction with given textual description $\mathcal{P}_{\text{Camera}}$ about the camera. Then mapping pre-define parameter between prompt and camera direction and position, CPM project the 2D skeletons with the projection matrix determined by prompt $\mathcal{P}_{\text{Camera}}$.

$$\begin{pmatrix} X_I \\ Y_I \\ w \end{pmatrix} = P_{proj} \begin{pmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{pmatrix}. \tag{5}$$

The final output of CPM is a direction aware 2D projected skeleton $\hat{J}_i^{2D}$. Note that it is not necessary to use $\mathcal{P}_{\text{Camera}}$ to decide camera position. If there is no textual description on camera position, then a identity matrix is used for camera extrinsic matrix.

2.3. Skeleton-Guided Text-to-Video Generation

Using the output of the second stage, we use a text-to-video network which uses a 2D skeleton from the CPM as a guidance. Let G be a text-to-video network and γ is its parameter. Given 2D skeleton from the CPM $\hat{J}_i^{2D}$, we get the videos consists of m frames $\{f_1, \cdots, f_m\}$.

This stage is formulated as below where $\hat{\mathbf{J}}^{2D}$ is the sequence of 2D projected motions represented as a concatenated form. The formal definition of $\hat{\mathbf{J}}^{2D}$ and output of G are formulated as below.

$$\hat{\mathbf{J}}^{2D} = \text{concat}(\hat{J_1}^{2D}, \cdots, \hat{J_m}^{2D}), \tag{6}$$

$$\{f_1, \cdots f_m\} = G(\hat{\mathbf{J}}^{2D}, \mathcal{P}; \gamma). \tag{7}$$

3. Experiments

In this section, we conducted three main experiments and analyzed the results. First, we compared Text-to-Video Generation results considering the presence or absence of pose generated from the T2M-GPT [16] guidance which are provided to the first stage from our framework. Second, we compared the generated videos using two different Text-to-Motion networks. Third, we experiment our framework using a camera prompt.

3.1. Evaluation Metrics

Action Classification (AC) accuracy The ratio of well classified videos to whole generated videos. It measures how generated videos are matching with actions in prompts. To evaluate how text prompts $\mathcal{P}$ are well aligned with video output, we use an action classification model Text4Vis [19] to evaluate action classification accuracy on the classes (jump, run, climb, kick, punch, clap, golf, sit).

CLIPscore (CS) [20] This measures how well the generated videos are well aligned with text prompts.

Frame Consistency (FC) [21] This is an average of cosine similarity between all consecutive pairs of CLIP image embeddings on all frames. This measures how naturally generated frames change.

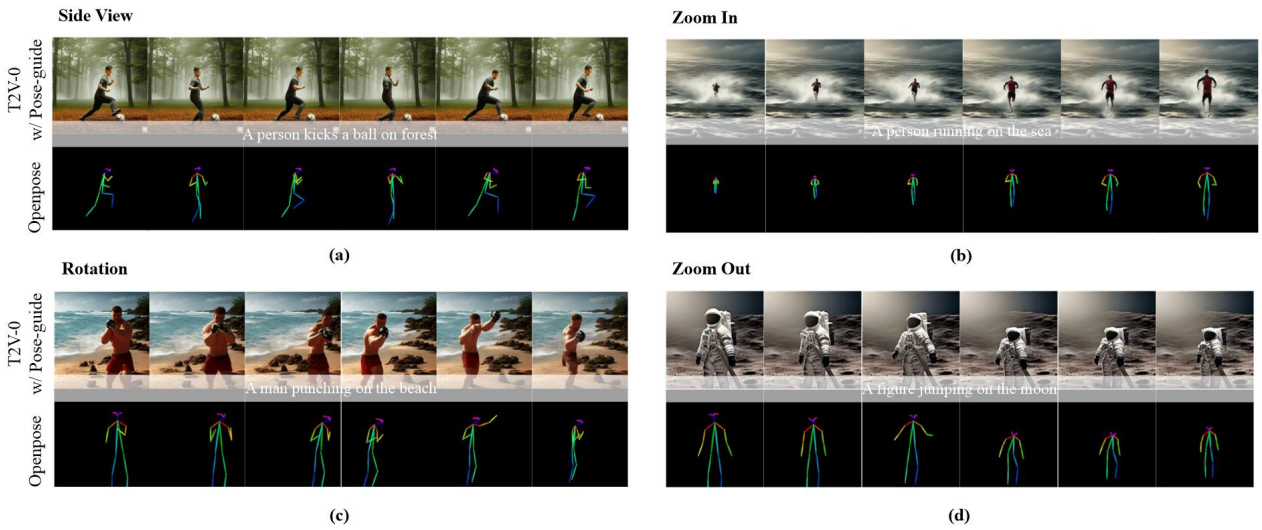

Figure 3. This figure shows the T2V results obtained through the CPM we suggested. (a),(b),(c), and (d) respectively show the outcomes of applying Side view, Zoom In, Rotation, and Zoom Out. When the CPM is applied to dynamic human motions such as kicking, running, punching and jumping, it can be observed that the motion is accurately represented, resulting in visually appealing T2V outcomes.

Table 1. Quantitative comparison between two different Text-to-Motion networks on action classification (AC) accuracy, frame consistency (FC) [21], CLIPscore (CS) [20]. Pose guidance was used as a T2M-GPT [16].

Without Pose Guidance	AC↑	FC↑	CS↑
ModelScope [22]	32.5%	88.9%	30.1
Text2Video-Zero [23]	44.1%	81.7%	28.4
With Pose Guidance			
Follow Your Pose [7] + Ours	**48.9%**	87.5%	**30.4**
Text2Video-Zero [23] + Ours	47.8%	**92.2%**	29.9

Table 2. Quantitative results applying camera rotation and skeleton scaling on CPM with pose guidance.

Camera Rotation	AC↑	FC↑	CS↑
Default	51.9%	**92.8%**	30.3
Top view	**57.7%**	**92.8%**	30.5
Lateral view	51.0%	92.7%	**30.7**
Skeleton Scale			
Default	**51.9%**	**92.8%**	**30.3**
Zoom in	48.0%	92.7%	29.6
Zoom out	45.2%	92.7%	29.3

3.2. Quantitative Results

Table 1 shows the quantitative results on AC, FC [21] and CS [20] with and without pose guidance of text-to-motion network. The pose guidance used is human motion generated from T2M model like T2M-GPT [16] with text prompts, not human motion Ground Truth. AC has improved using pose guidance than without using it in both text-to-motion networks. This shows that with pose guidance, the ambiguity of generated motions is reduced. Moreover increased FC [21] shows that using our frameworks, similarity of consecutive frames which means a sudden change on movements between frames decreases making more natural movements. Using FYP [7] as a Text-to-Video network has the best AC among three Text-to-Video networks. Moreover in FC [21] and CS [20], T2V-Zero [23] has the best scores among others. This is because the background changes with pose changes in FYP [7], but not in T2V-Zero [23] where background images are fixed. Then, we compared our results using CPM. We give camera rotation and translation with two prompts each as shown in Table 2. Even rotating and translating camera, the results outperform the results of not using a pose guidance. The AC for the class jump with using $\mathcal{P}_{\text{Camera}}$ to "bird's eye view" improved from 33.8% to 86.7%. Moreover, the class kick using viewing direction prompt "side view" improved from 53.8% to 86.7% These implies certain viewing directions are more adequate to describe motions.

3.3. Controllable Camera Results

There are motions where their representations are highly depend on camera directions. These problems occur when the difference of motions flows in 3D space and optical flows in projected 2D space is high. For example, as shown in (a) in Fig 3, an action kicking is well represented when a camera captures the person's lateral part. Imagine the same result when the camera is viewing straight to the person, the motion would be less plausible. This is because, in the case of the action "kicking", the motion flow in 3D space is substantial, while the optical flow in projected 2D space is not. Furthermore, more diverse representation of motions is possible as seen in Fig 3 (b), (c) and (d). So the control of camera is necessary.

References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[2] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[3] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023.

[4] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

[5] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[6] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[7] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv preprint arXiv:2304.01186*, 2023.

[8] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

[9] Judith Butepage, Michael J Black, Danica Kragic, and Hedvig Kjellstrom. Deep representation learning for human motion prediction and classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6158–6166, 2017.

[10] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 474–489. Springer, 2020.

[11] Arij Bouazizi, Adrian Holzbock, Ulrich Kressel, Klaus Dietmayer, and Vasileios Belagiannis. Motionmixer: mlp-based 3d human body pose forecasting. *arXiv preprint arXiv:2207.00499*, 2022.

[12] Xiangjun Tang, He Wang, Bo Hu, Xu Gong, Ruifan Yi, Qilong Kou, and Xiaogang Jin. Real-time controllable motion transition for characters. *ACM Transactions on Graphics (TOG)*, 41(4):1–10, 2022.

[13] Yinglin Duan, Tianyang Shi, Zhengxia Zou, Yenan Lin, Zhehui Qian, Bohan Zhang, and Yi Yuan. Single-shot motion completion with transformer. *arXiv preprint arXiv:2103.00776*, 2021.

[14] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023.

[15] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

[16] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xiaodong Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. *ArXiv*, abs/2301.06052, 2023.

[17] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[18] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[19] Wenhao Wu, Zhun Sun, and Wanli Ouyang. Revisiting classifier: Transferring vision-language models for video recognition. 2023.

[20] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[21] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models, 2023.

[22] ModelScope: bring the notion of model-as-a-service to life. `https://github.com/modelscope/modelscope`. Accessed: 2023-06-28.

[23] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023.