Human Motion Aware Text-to-Video Generation with Explicit Camera Control

Anonymous WACV Algorithms Track submission

Paper ID 919

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 two-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 in which 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 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 codes are available in `https://anonymous.4open.science/r/HMTV-26BB`.

1. Introduction

Nowadays, Text-to-Image model (T2I) that generates images using 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. Along with the growth of T2I, Text-to-Video model (T2V), which generates the corresponding video with given text prompt is also developed.

Seminal research on T2V has gained momentum with the advent of models such as Dreamix [3], VDM [4], ImagenVideo [5], and Make-A-Video [6] based on the diffusion models have achieved outstanding results in T2I. However, they are not without problems. First, it's very awkward when they create complex movements like human motions. To solve this problem, Skeleton-Guided Text-to-Video Generation [7, 8] which conditioned on human pose skeleton, when creating a video enables pose control of the subject in the video. However, it is difficult to use for various applications because not only text prompt information but also the human pose is required. Second, even with the given human pose, undesired results are generated. For example, with certain human pose condition the generated outputs have limited viewing direction 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 scale of a generated human is an issue: too big or too small subject is generated.

On the other hand, as various generation models have been actively studied recently, Human Motion Generation is also attracting a lot of attention. The early methods of Human Motion Generation use human motion prediction [9–11] 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] which is recently released, is expected to be highly applicable as it can generate complex movements with long sentences.

In this paper, we propose a new video generation algorithm that naturally generates human movements by combining Text-driven Human Motion Generation and Skeleton-Guided Text-to-Video Generation. In particular, it creates high-quality human motions that guide video generation through Text-driven Human Motion Generation for input text prompts. Next, it projects the generated 3D human motion sequence to a 2D skeleton format. At this time, when an additional camera prompt is given, move it to a specific camera position using a pre-defined camera extrinsic parameter. Lastly, the first input text prompt with a hu-

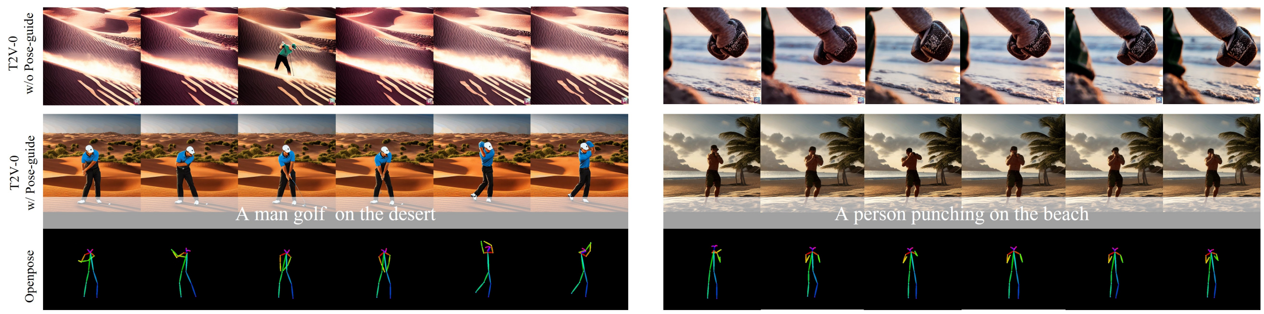

Figure 1. This figure shows the difference of output video between with using pose guidance and without using it to T2V network. Text prompts are applied to both. Network without pose guidance show problems with size of human and has inconsistency between consecutive frames, but using pose guidance these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

man motion sequence projected in 2d are used to generate videos where a person's movements are naturally created with using text prompt only.

The most noteworthy point to focus on is that we not only improve the generation quality of videos containing 'human motion', but also control the viewing angle and the movement of the camera. In addition, the generated SMPL [17] meshes can be converted to skeletons in the form of OpenPose [18] or MSCOCO [19] so that it can be applied to various Skeleton-Guided Text-to-Video Generation. As far as we know, our framework is the first one which can control the camera composition and the scale of the subject as desired. Moreover, 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.

It is also meaningful that our model is highly likely to be used as a framework connecting Text-driven Human Motion Generation and Text-to-Video. Seminal research has be conducted in the two fields we have employed, which are currently experiencing a surge of active study and hold unlimited potential for development. With better models in each fields and with combinations of them within our framework the better Text-to-Video model we can have. In this perspective, our proposed framework has great potential for development.

In summary, our contributions are:

- We propose a framework that combines the Text-driven Human Motion Generation module and the Skeleton-Guided Text-to-Video Generation module to generate videos that express complex human behavior well with only text prompt.
- Our framework has advantages of being able to control camera angles and locations and switch to various skeleton formats (MSCOCO [19], OpenPose [18]) because it extracts human poses in 3D rather than 2D.
- Our methods outperform previous Text-to-Video methods not only in quantitative metrics but also in qualitative results. Moreover actual film shooting techniques can be applied so that it is highly scalable in various applications.

2. Related Work

2.1. Text-driven Video Generation

Text-driven video generation is a task that combines natural language processing and computer vision to generate a video from a given prompt. Compared to T2I field, T2V field is still one of the difficult tasks due to the lack of text-video paired datasets and the complexity of modeling high-dimensional video data.

In early studies, the mainstream methods predicts next frame from the initial frame. [20], [21]. After that, Generative Adversarial Networks (GAN) [22]-based models were appeared which can generate unconditional videos without using the first frame or with only classes given as condition. DIGAN [23] proposed a model of implicit neural representations (INRs) by integrating temporal dynamics for video generation. Using DIGAN, it is possible to generate long videos. However, GAN-based models can have difficulties modeling large datasets. Since then, as language models and transformers have developed, video generation from text prompt has become possible. GODIVA [24] extended the Vector Quantized-Variational AutoEncoder (VQ-VAE) [25] to T2V generation by mapping text tokens to video tokens and generated more realistic scenes. NUWA [26] presents an auto-regressive framework that can be used for both T2I and T2V tasks, and is an extension of GODIVA [24]. Diffusion [27] is a technique that adds noise to the input image and then removes the noise in several steps to produce a realistic image. Video Diffusion Models (VDM) [4] uses a space-time decomposition U-Net [28] to directly perform the diffusion process at the pixel. Make-A-Video [6] uses T2I to learn the relationship between text and video, and learns motions with unsupervised learning on unlabeled video data. Unlike the above motion which generate video directly, our method conditionally generates video using text-driven human motion.

2.2. Text-driven Human Motion Generation

Recently, various methods such as VAE (Variational AutoEncoder) [29], GAN [22], Diffusion [27] have been studied for human motion generation. First of all, GAN [22] aims to learn a distribution similar to the dataset through competition between generator and discriminator. Harvey *et al.* [30] attempted to solve the problem generating blurry motion. They constructed two discriminator networks for short-term and long-term critics.

Furthermore, models based on VAE [29] are also often used in human motion generation. Yan *et al.* [31] and Aliakbarian*et al.* [32] generate human motion by predicting the next human motion given a human motion sequence. These two methods use VAE [29], which encodes a pair of current and future sequences and then reconstructs the future sequence. ACTOR [33] proposed a transformer-based VAE. They proposed human motion generation in a non-autoregressive way in an action class. TEMOS [34] is an extended model of ACTOR [33], which brings an additional text encoder and generates various motion sequences. T2M-GPT maps motion to discrete values with VQ-VAE [25] and use motion-GPT, a GPT [35] like network [16] to predict the next discrete values or indices which correspond to motions. Then decode these indices to the motions that corresponds to the given text.

Most recently, diffusion based methods, such as Motion Diffuse [15] and Motion Diffusion Model (MDM) [14], have been widely studied. Motion Diffuse [15] is the first model that uses diffusion model [27] for text to motion generation. MDM [14] is diffusion-based generative model without classifier for the human motion domain. Most recently, diffusion based methods have been widely studied. Motion Diffusion Model (MDM) [14] is a diffusion-based generative model without classifier for the human motion domain.

Among various text-driven human motion generation models, we tested on T2M-GPT [16] and diffusion-based MDM [14] in this paper. The outputs of this model are used as a guidance for T2V model.

3. Proposed Method

In this section, we overview our proposed framework which is shown in Fig. 2. Our method aims for enhancing a quality and diversity of generated videos with human motions inside and consists of three stages. (1) Text-to-Motion Generation stage which generates motion from given texts. Various Text-to-Motion Generation models can be applied in this stage. (2) 3D to 2D joint projection is the stage that projects generated 3D human motions to 2D space. In this stage, we use camera projection module (CPM) which projects human motion with text driven camera matrix. In this module, texts which describe the viewing direction can be given. With given texts, this module controls the output projection style by adjusting camera angles and distances. Therefore, we can create diverse scenes with different camera angles and movements. Finally, (3) a Skeleton-Guided Text-to-Video Generation stage generates video using texts from the first stage and projected human joints together. Similar to the first stage, we can use various Text-to-Video networks in our last stage.

3.1. Text-to-Human Motion Generation

Text-to-Motion Generation stage uses predefined Text-to-Motion (T2M) network that generates sequential 3D human motion. Note that a 3D human motion is a sequence of a set of joints consisting human where joints represent relative locations to the root position. Formally given input text $\mathcal{P}$, T2M network $F(\mathcal{P};\theta)$ generates sets 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. T2M [36] uses a convolutional motion autoencoder to get motion snippet code which is latent sequences of motions. With given texts, they approximates the conditioned probability distribution with Text2Length Sampling. In motion generation stage, they generate 3D human motion conditioned on text and sampled motion length. MDM [14] and MotionDiffuse [15] use a transformer and diffusion based architectures to generate 3D human motion. We can use these models in our first stage. Similar to T2M [36], T2M-GPT [16] model uses VQ-VAE [37] to encode latent sequences. Then, it uses the motion-GPT to sequentially generate indices. In this model, the text-to-motion generation is formulated as auto-regressive fashion when predicting a next index. Formally with given $i-1$ indices $S_{<i}$ and with text c, they choose the next index which maximize the probability $p(S_i|c, S_{<i})$. Therefore, in this stage a pre-trained motion VQ-VAE is required.

3.2. 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 stage takes 3D mesh vertices from text-to motion network and uses joint regressor

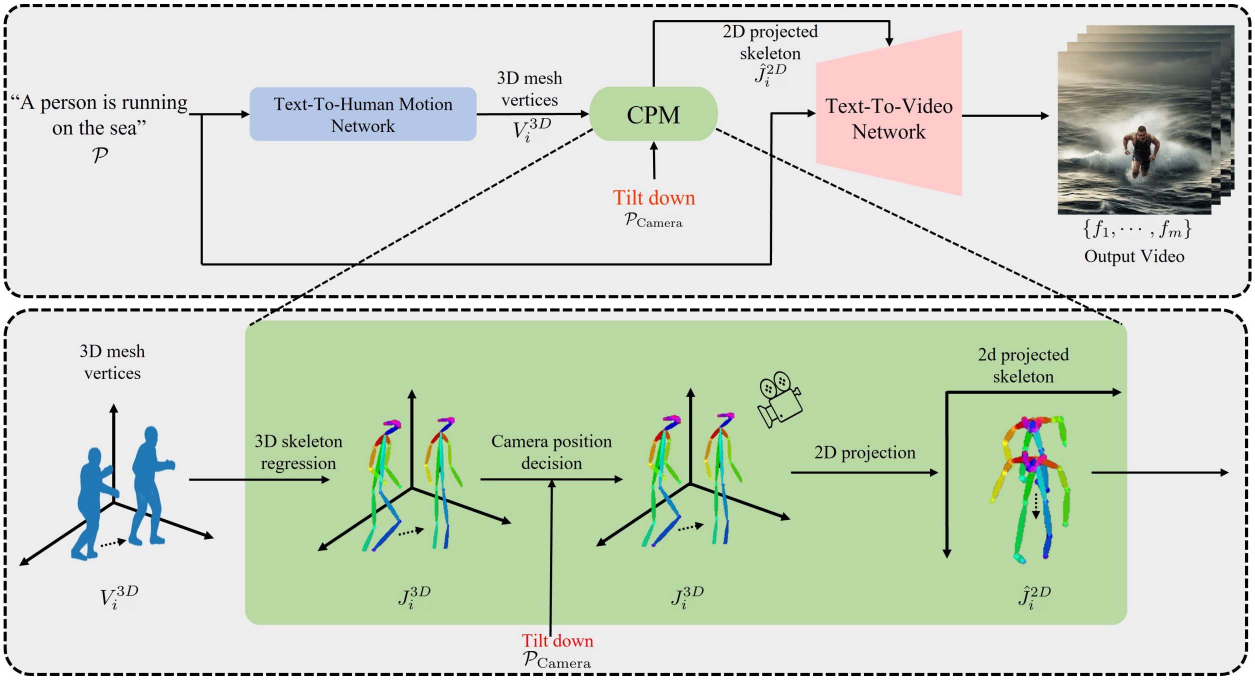

Figure 2. **Overall process of our proposed framework. Top**: Our framework consists of three stages: (1) Text-to-Motion Generation, (2) Camera Projection Module, (3) Skeleton Guided Text-to-Video Generation. A text prompt is passed to the Text-to-Human Motion Generation network to generate 3D mesh vertices 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 $\mathcal{P}_{\text{Camera}}$.

from [17] 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 changing of camera position using camera prompt. We can express rotation and translation with a camera extrinsic matrix using the homogeneous coordinate denote as below

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

Note that $R_{3\times 3}$ defines the rotation of a camera and $t_{3\times 1}$ defines the translation of the camera. With intrinsic matrix together we can define a projection matrix P_{proj} as below.

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

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

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

The final output of CPM is 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 identity matrix is used for camera extrinsic matrix.

3.3. Skeleton-Guided Text-to-Video Generation

Using the output of second stage, we use text-to-video network which uses 2D skeleton from the CPM as a guidance. Let G be a text-to-video network and γ is its parameter. With 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 a sequence of 2D projected motions represented as concatenated form. The formal definition of $\hat{\mathbf{J}}^{2D}$ and output of G are formulated as below.

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

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

4. Experiments

In this section, we conducted three main experiments and analyzed the results. First, compared Text-to-Video Generation results between with and without pose guidance from the first stage from our framework. Here, we exploited Modelscope [38], Text2Video-Zero (T2V-Zero) [39], Follow Your Pose (FYP) [7] as Text-to-Video networks. Second, we compared the generated videos using two different Text-to-Motion networks: T2M-GPT [16] and MDM [14]. Third, we experiment our framework using camera prompt. We used static shot (default), top view, lateral view, zoom in/out for camera position description prompts.

4.1. Prompt Set

Our framework needs text description $\mathcal{P}$ to generate video. Complex and diverse motion descriptions are required. Therefore, we took a test prompt set of HumanML3D [36] and randomly added location at the end of the prompts. Locations are randomly selected from these category: sea, forest, moon, beach, desert, and auditorium. We used these location specified prompts to Follow Your Pose [7] and in experiment with T2V-Zero [39], we simply modified the prompt into {person} – {verb} – {location} from the prompt that we use in FYP [7]. This is because if a complex prompt pass through T2V-Zero [39] then it outputs unrecognizable videos.

4.2. Evaluation Metrics

Action Classification (AC) accuracy The ratio of well classified video to whole generated video. It measures how well generated videos are matching with action in prompts. To evaluate how text prompts $\mathcal{P}$ are well aligned with video output, we use action classification model Text4Vis [40] to evaluate action classification accuracy on the classes (jump, run, climb, kick, punch, clap, golf, sit).

CLIPscore (CS) [41] This measures how well the generated videos are well aligned with text prompts.

Frame Consistency (FC) [42] 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.

Table 1. Quantitative comparison between two different Text-to-Motion networks on action classification (AC) accuracy, frame consistency (FC) [42], CLIPscore (CS) [41].

Without Pose Guidance			
	AC	FC	CS
ModelScope [38]	32.5%	88.9%	30.1
Text2Video-Zero [39]	44.1%	81.7%	28.4
With Pose Guidance			
Follow Your Pose [7]	48.9%	87.5%	30.4
Text2Video-Zero [39]	47.8%	92.2%	29.9

4.3. Experimental Results

Quantitative Results Table 1 shows the quantitative results on AC, FC [42] and CS [41] with and without pose guidance of text-to-motion network. 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 [42] 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 [42] and CS [41], T2V-Zero [39] has the best scores among others. This is because the background changes with pose changes in FYP [7], but not in T2V-Zero [39] where background images are fixed. We also compared the evaluation metrics using two different Text-to-Motion networks with using pose guidance which is shown in Table 3. Then, we compared our results using CPM. We give camera rotation and translation with two prompts each as shown in Table 2. This is the average results of every pre-defined classes. Even rotating and translating camera, the results outperform the results of not using a pose guidance. This implies modifying camera matrix with CPM dose not compensate the output video quality.

Qualitative Results As we mentioned before, the qualitative results shown in Fig. 3 implies that using pose guidance solves two problems: Undesirable scale problem and inconsistency problem between consecutive frames. Both FYP [7] and T2V-Zero [39] have undesirable scale problem which does not represent the whole body of human. Moreover, without using pose guidance in T2V-Zero [39] generated human disappear for some frames making the video inconsistent in time domain. However with using pose guidance consecutive frames are consistent relative to without using pose guidance. The image sequence on the top of Fig. 3 is the results from T2V-Zero [39] with using only text prompts. The sequence below is the results from the same network using pose guidance with text prompt. This means that with pose guidance, we can avoid undesirable scale in generated video. Moreover generated videos only guided with text prompts have a inconsistency problems. However

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

Table 3. Quantitative results comparison between two different Text-to-Motion networks using pose guidance.

Text-To-Motion Model			
	AC	FC	CS
T2M-GPT [16]	47.8%	92.2%	29.9
MDM [14]	46.0%	92.1%	30.2

using our method generated human does not disappear between consecutive frames which shows the consistent video on human motion.

User Study We applied our network to show evaluators the results of our main experiment, with and without pose guidance of Text-to-Motion network. They were presented with a total of three criteria for qualitative evaluation of video performance containing human behavior: semantic relevance of prompt and video, realism of human motion, overall quality and the results are as follows. 73% of the evaluators judged that the video applied to our network was better in terms of semantic relevance of prompt and video, 68% of the evaluators preferred our results in terms of realism of human motion, And 89% of respondents said that in terms of overall quality, the video that went through our network was better overall. Considering that the importance of all three criteria for evaluating videos is the same, 77% of all evaluators rated videos that passed through our network better.

4.4. Ablation Studies

In this section, we will show how a camera description prompt $\mathcal{P}_{\text{Camera}}$ affects the output video in quantitative and qualitative manners.

Variant T2M Model We choose state of the arts models T2M-GPT [16] and MDM [14] the quantitative results of T2M-GPT [16] are better than MDM [14] as in Table 3.

Viewing Direction 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 with using viewing direction prompt "side view" improved from 53.8% to 86.7% These implies certain viewing directions are more adequate to describe motions. Therefore, the control of viewing direction is important. This is why our proposed framework is meaningful. As shown in Fig. 5, we can scale the size of the human. The problem is that the generated background does not change. Note that this is the problem of Text-to-Video network. With better network this would not be a problem.

4.5. Application

As shown in Table 1 and Table 2, even adjusting camera matrix with CPM, the quantitative results tells the quality of output video still outperform ones without a pose guidance. In qualitative manner shown in Fig. 4 steering the direction of generated human These are important issues, since prompt aligned output and elaborate control are prerequisites for user level applications. In this sense, our research has provided one of the directions for future studies.

4.6. Limitations

Although our methods enhance the quality of output video with human motions, there are limitations of our method. First, our method cannot automatically regress adequate camera pose for viewing direction. We experiment an interpolation of camera matrix based on similarity of word embeddings in naive way. We left this for future works. We look forward the integration with advanced natural language processing fields. Second, as we mentioned the background does not change in the case of T2V-Zero [39] network. Even with an adjustment of camera position and direction, the size of the human change but not the background so that the output videos look like a human shrinking. Third, the output quality of our method depends on the performance of both Text-to-Motion and Text-to-Video networks. Shown in Fig. 6: **Top**, the generated motions from Text-to-Motion network does not align to the prompt resulting mis-aligned output video. Moreover, in Fig. 6: **Bottom**, even though Text-to-Motion network work well, the output video may be mis-aligned because of Text-to-Video network.

5. Conclusion

In this work, we addressed the problems on Text-to-Video. First, Text-to-Video models have weaknesses in generating text aligned output including human motions. Second, explicit controls of the angle and the view point are not possible in existing Text-to-Video models. Therefore, we proposed a method which generate videos guided by texts and 3D human motion conditioned on the same texts. Moreover, we show the explicit control of the angle and the view point of video by texts with CPM. To the best of our knowledge, we are the first framework which exploits Text-to-Motion networks to guide Text-to-Video models and explicit control of camera positions and directions by adjusting camera matrix directly. We hope that our research will have a positive impact on the subsequent studies and applications involving Text-to-Video tasks.

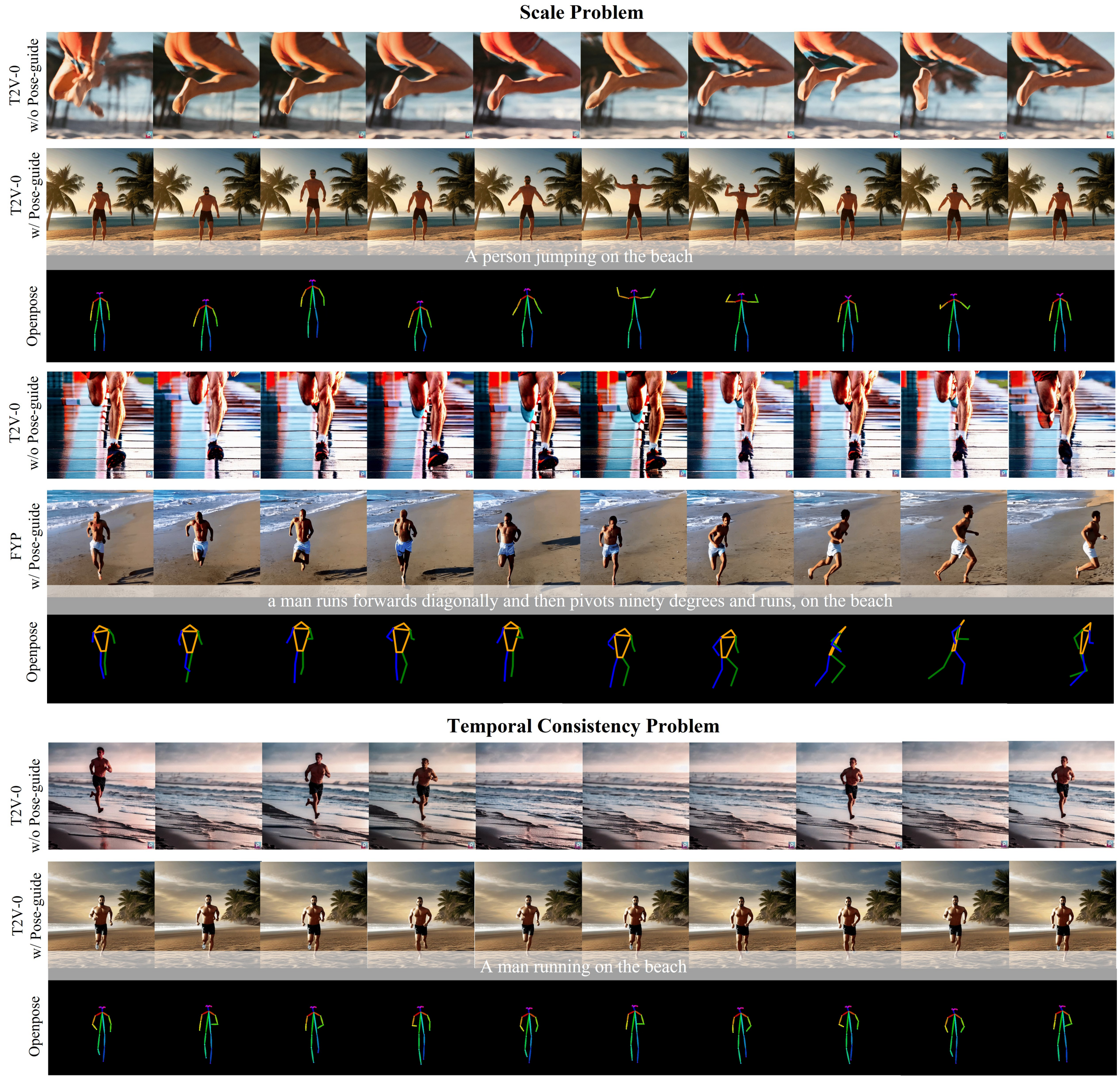

Figure 3. Models without pose guidance show problems with inadequate human size and inconsistencies between consecutive frames

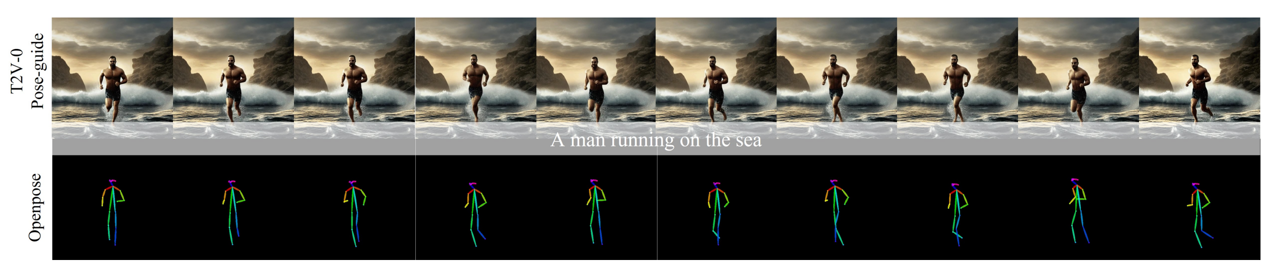

Figure 4. This figure shows that a translation of camera steers generated human motion to desired direction.

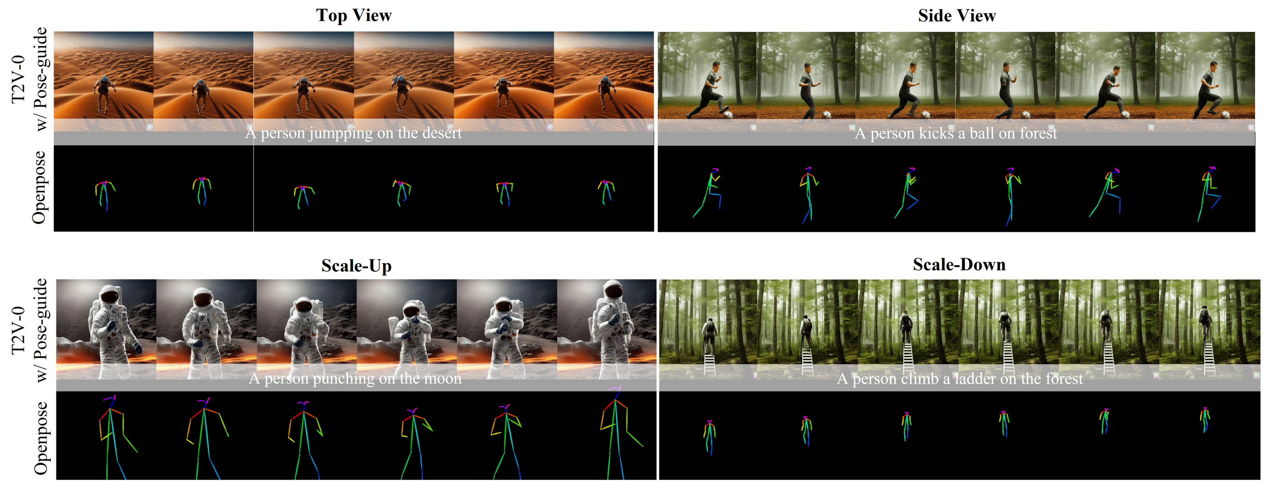

Figure 5. This figures shows that with CPM, we can generate human motion from diverse view points such as top view and side view. Moreover, scaling the size of the human is possible.

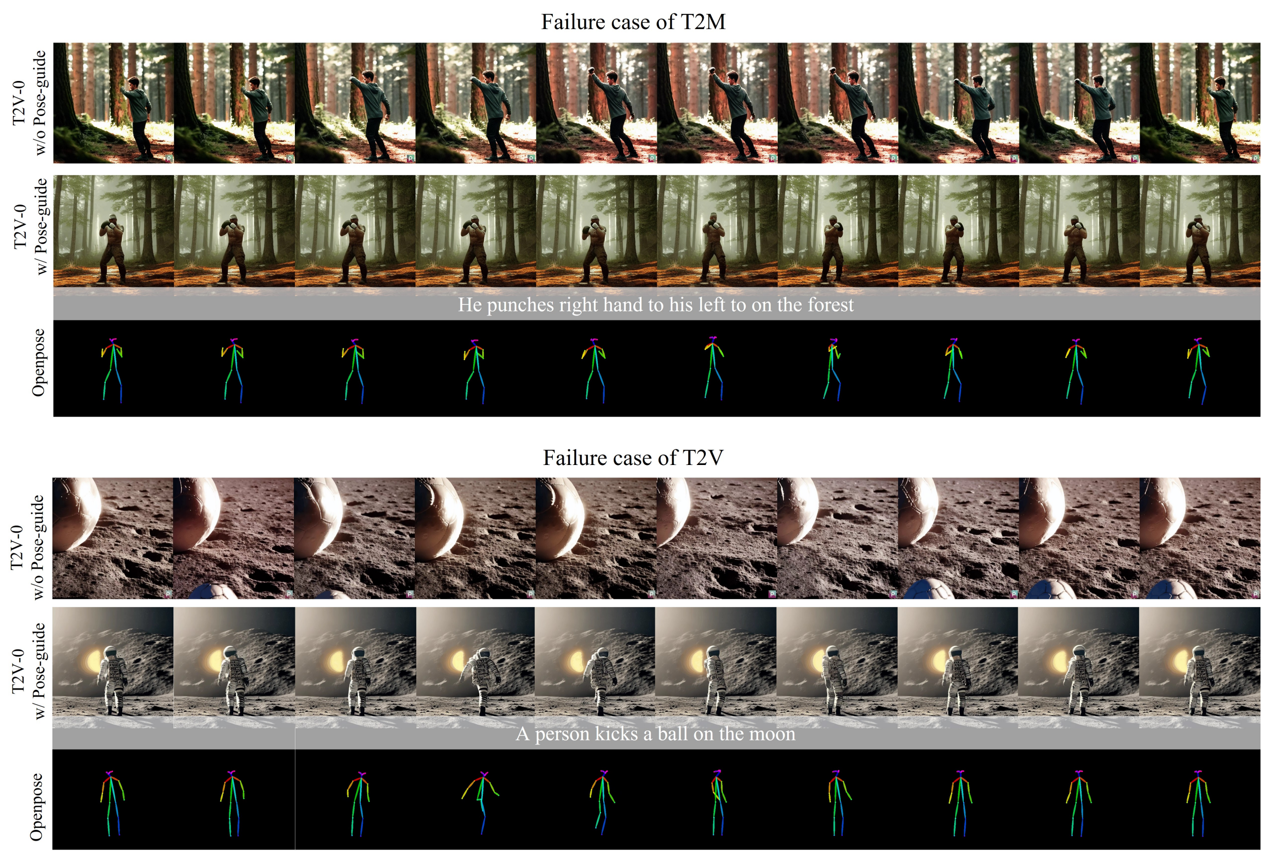

Figure 6. Failure cases depends on what networks are responsible for. **Top**: Output video does not align with text prompt, since T2M network generates motion with opposite side of hand. **Bottom**: Output video does not align with text prompt, since T2V network does not generate ball in the video.

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. 1

[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. 1

[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. 1

[4] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 1, 2

[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. 1

[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. 1, 2

[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. 1, 5

[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. 1

[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. 1

[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. 1

[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. 1

[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. 1

[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. 1

[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. 1, 3, 5, 6

[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. 1, 3

[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. 1, 3, 5, 6

[17] 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. 2, 4

[18] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 2

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[20] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. *Advances in neural information processing systems*, 29, 2016. 2

[21] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S Yu. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. *Advances in neural information processing systems*, 30, 2017. 2

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2, 3

[23] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. *arXiv preprint arXiv:2202.10571*, 2022. 2

[24] Jaemin Yoo, Lingxiao Zhao, and Leman Akoglu. End-to-end augmentation hyperparameter tuning for self-supervised anomaly detection. *arXiv preprint arXiv:2306.12033*, 2023. 2

[25] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2, 3

[26] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 720–736. Springer, 2022. 2

[27] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2, 3

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2

[29] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[30] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020. 3

[31] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *Proceedings of the European conference on computer vision (ECCV)*, pages 265–281, 2018. 3

[32] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5223–5232, 2020. 3

[33] 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. 3

[34] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 480–497. Springer, 2022. 3

[35] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 3

[36] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5142–5151, 2022. 3, 5

[37] 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. 3

[38] ModelScope: bring the notion of model-as-a-service to life. https://github.com/modelscope/modelscope. Accessed: 2023-06-28. 5

[39] 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. 5, 6

[40] Wenhao Wu, Zhun Sun, and Wanli Ouyang. Revisiting classifier: Transferring vision-language models for video recognition. 2023. 5

[41] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 5

[42] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models, 2023. 5