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

Taehoon Kim*[1] ChanHee Kang*[2] JaeHyuk Park*[1]
Daun Jeong*[1] ChangHee Yang*[2] Suk-Ju Kang†[2] Kyeongbo Kong†[3]
Pukyong National University[1], Sogang University[2], Pusan National University[3]
kimth52001@pukyong.ac.kr, jasperai@sogang.ac.kr, hyeok0831@naver.com
ekdnsdl15@naver.com, yangchanghee2251@gmail.com, sjkang@sogang.ac.kr, kbkong@pusan.ac.kr

Abstract

With the rise in expectations related to generative models, text-to-video (T2V) models are being actively studied. Existing text-to-video models have limitations such as in generating complex movements replicating human motions. These model often generate unintended human motions, and the scale of the subject is incorrect. To overcome these limitations and generate high-quality videos that depict human motion under plausible viewing angles, we propose a two stage framework in this study. In the first stage a text-driven human motion generation network generates three-dimensional (3D) human motion from input text prompts and then motion-to-skeleton projection module projects generated motions onto a two-dimensional (2D) skeleton. In the second stage, the projected skeletons are used to generate a video in which the movements of a subject are well-represented. We demonstrated that the proposed framework quantitatively and qualitatively outperforms the existing T2V models. Previously reported human motion generation models use texts only or texts and human skeletons. However, our framework only uses texts and outputs a video related to human motion. Moreover, our framework benefits from using skeleton as an additional condition in the text-to-human motion generation networks. To the best of our knowledge, our framework is the first of its kind that uses text-driven human motion generation networks to generate high-quality videos related to human motions. The corresponding codes are available at https://github.com/CSJasper/HMTV.

1. Introduction

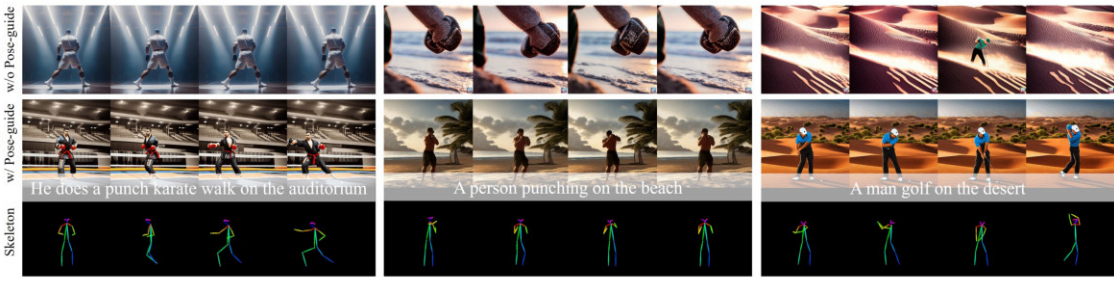
Present, text-to-image models (T2I) that generates images using a given text prompt are being actively studied. Particularly, models such as stable diffusion [1] and DALL-E2 [2] are attracting considerable attention owing to their outstanding performance. Alongside the growth of T2I models, text-to-video models (T2V), which generates a video based on a given text prompt, are also being developed.

Seminal research on T2V has gained momentum owing to diffusion-based models such as a Dreamix [3], video diffusion model (VDM) [4], ImagenVideo [5], and Make-A-Video [6]. However, these models face challenges when generating videos using prompts involving human motion. First, a prompt indicating the desired motion incorrectly produces the intended pose as shown in the first row in Fig.1 (a). Second, owing to the lack of information regarding the scale of the human body, a scaling problem emerges, as presented in the first row in Fig. 1 (b). Lastly, the lack of specific guidance causes inconsistencies across frames, as demonstrated in the first row in Fig. 1 (c), in which the prompt is meant to generate human motion, yet the human body disappear in some frames. To solve these problems, the skeleton-guided text-to-video generation [7, 8], which is conditioned on a human skeleton, has been introduced. However, this method creates or locates a pose that matches the given text, accurately obtaining the desired motion is difficult.

Early methods of human motion generation use human motion prediction [9–11] methods that predict subsequent actions based on previous actions and generate in-between motion [12, 13]. Recently, text-driven human motion generation, using models such as motion diffusion model (MDM) [14], MotionDiffuse [15], and text-to-motion generative pre-trained transformer (T2M-GPT) [16], which generates three-dimensional (3D) human motion sequences from text prompts, has been studied. In particular, T2M-GPT [16] is expected to be widely applicable as it can generate complex movements using sentence much longer than previous models.

In this study, we propose a novel video generation algorithm that generates natural human movements with text-to-skeleton module and a pose guided text-to-video module.

*These authors contributed equally to this work
†Corresponding author


T2V w/o Pose-guide
T2V w/ Pose-guide
Skeleton
He does a punch karate walk on the auditorium
A person punching on the beach
A man golf on the desert

(a) Motion Ambiguity Problem (b) Scale Problem (c) Temporal Consistency Problem

Figure 1. This figure shows the difference of output video between the presence and absence of pose guidance applied to T2V models. Models without pose guidance shows the problem of ambiguity, improper scale and temporal consistency problems. In the presence of pose guidance these problem do not happen.

Specifically, a text prompt is used as an input to generate high quality 3D human motion. This process can yield accurate human motion. Next, motions are converted to two-dimensional (2D) skeletons using a text guided projection matrix. At this stage, we can obtain desired pose by projecting a 3D motion onto the 2D pose corresponding to the camera direction prompt. Lastly, the input text and the 2D-human-skeleton sequence obtained from the generated 3D human motion are used to create a video containing high-quality human movements.

Notably, this framework generates high-quality videos containing human motion and controls the viewing angle and movement of the camera. To the best of our knowledge, our algorithm is the first of its kind that can control the camera position and the scale of the subject as desired. We comfirmed the feasibility of the model, demonstrating the potential to implement camera techniques commonly used in acutal films.

Importantly, our model can potentially be used as a framework connecting text-driven human motion generation and text-to-video models. As these model improve owing to seminal research, their combination within our framework would yield better text-to-video outputs.

Our contribution are as follows:

- We proposed a framework that employs text to skeleton module and text to video generation modules to synthesize a high quality video expressing dynamic scenes with complex human behavior by a text based camera control.
- Our framework can control a viewing direction of generated video when generating projected 2D skeleton which enables the output more dynamic and plausible.
- Our text-to-video methods outperform both in quantitative and qualitative results than previous methods.

2. Related Work

2.1. Text-driven Video Generation

In early studies, mainstream human motion generation methods predicted subsequent frames based on the initial frame [17, 18]. Subsequently, generative adverarial network (GAN) [19]-based models [20] were introduced, which can generate videos unconditionally without using the initial frame or with classes as the only condition.

As language models and transformers have developed, video generation from text prompts has become possible. Godiva [21] extended the Vector Quantized-Variational AutoEncoder (VQ-VAE) [22] to T2V generation by mapping text tokens to video tokens and generated highly realistic scenes. NUWA [23] proposed an auto-regressive framework that can be used for both T2I and T2V tasks, and is an extension of the model proposed by Godiva [21].

Diffusion [24] is a technique that adds noise to the input image and then removes the noise in several steps to produce a realistic image. A VDM [4] uses a space-time decomposition U-Net [25] to directly implement the diffusion process at the pixel level. Make-A-Video [6] uses T2I to learn the relationship between text and video, and learns motions via unsupervised learning on unlabeled video data. In addition, ModelScope [26], Zero-Scope [27], Runway-AI [28] and CogVideo [29] are T2V models that generate plausible video. Even though, there have been several such efforts, there are problems, namely motion ambiguity problem, scale problem and temporal consistency problem as shown in Fig. 1. Our method is also in the field of Text-driven Video Generation and we addressed these problems in two stages, which consist of the text-to-skeleton module and the pose guided text-to-video module.

2.2. Text-driven Human Motion Generation

Our text-to-skeleton module consists of the text-to-motion generation module and the motion-to-skeleton projection module. Recently, various methods such as Variational AutoEncoder (VAE) [30], diffusion [24] have been studied for human motion generation. Models such as [31–34] which are based on VAE [30] are often used for human motion generation. The most representative method is T2M-GPT [16]. It maps motion to discrete values using VQ-VAE [22] and uses motion-GPT, a GPT [35]-like network [16], to predict the next discrete values or indices that correspond to the motions. Then, these indices are decoded to obtain the motions that correspond to the given text. Diffusion based methods, such as Motion Diffuse [15] and MDM [14], have been widely studied. Motion Diffuse [15] is the first model that uses a diffusion model [24] for text to motion generation. MDM [14] is a diffusion-based generative model without a classifier for the human motion domain. Text-driven human motion generation is a powerful technique that can output a desired motion for skeleton-guided T2V model. Among various text-driven models used for generating human motion, we tested T2M-GPT [16] and diffusion-based MDM [14] in this study. The outputs obtained from the models were then used as guidance for the T2V model in our framework.

2.3. Pose-guided Video Generation

Despite the efforts to implement text-driven video generation models, directly generating human motion in videos is difficult. Previously, generative models [36, 37] used skeleton guidance to generate human motion, however, there was a problem with generating random style video. Recently, based on these ideas, models [7, 8] have shown the possibility of solving this problem using both prompt and skeleton guidance, however, these models cannot be implemented immediately because of the problem of inputting skeleton videos and text. We solved this problem by combining text-driven human motion generation approach.

3. Proposed Method

In this section, we present an overview of our proposed framework, which is shown in Fig. 2. Our framework aims to generate high quality and diverse videos depicting human motions and the method consists of two stages. The (1) text-to-skeleton generation stage generates motion from given texts and then use motion-to-skeleton projection module to obtain projected 2D skeletons. In the text-to-motion generation, various text-to-motion generation models can be applied in this stage. Motion-to-skeleton projection is a sub-stage that projects generated 3D human motions to 2D space. In this stage, we use motion-to-skeleton projection module which projects human motion with text driven camera matrix. In this step, texts which describe the viewing direction could be given. With given texts, this module controls the output projection style by adjusting camera angles and distances. Therefore, we can generate diverse scenes using different camera angles and movements inally, (2) the skeleton-guided text-to-video generation stage generates videos using the texts from the first stage along with the projected human joints. Similar to the first stage, we can use various text-to-video models in the final stage.

3.1. Text-to-Skeleton Motion Generation

This stage consists of two sub-stages: Text-to-motion Generation and Motion-to-skeleton Projection.

Text-to-motion Generation

Text-to-motion generation stage uses a predefined T2M network that generates sequential 3D human motion. Note that 3D human motion is generated by a set of joints that represent locations relative to a root position. Using input text $\mathcal{P}$, the T2M network $F(\mathcal{P};\theta)$ generates sets of vertices $\{V_i^{3D}\}_{i=1}^{K}$ which form meshes representing human motion as expressed below.

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

where the θ is a model parameter of T2M network and K is the number of vertices in the meshes.

T2M [38] uses a convolutional motion autoencoder to obtain motion snippet code, which contains latent sequences of motions. Using given texts, the network approximates the conditioned probability distribution using Text2Length Sampling. At the motion generation stage, 3D human motion, which is conditioned on the given text and sampled motion length, is generated. MDM [14] and MotionDiffuse [15] use transformer and diffusion based architectures to generate 3D human motion. We can use these models during the first stage. Similar to T2M [38], the T2M-GPT [16] model uses VQ-VAE [39] to encode latent sequences. Then, the model uses motion-GPT to sequentially generate indices.This model generates motion from text in an auto-regressive fashion when predicting the next index. Using the given $i-1$ indices, $S_{<i}$, and text c, the model chooses the next index which maximizes the probability $p(S_i|c, S_{<i})$. Therefore, at this stage a pre-trained motion VQ-VAE is required.

Motion-to-skeleton Projection

At this stage, we will introduce motion-to-skeleton projection module shown at the bottom of Fig. 2. This module accepts a preset text description of a camera direction, $\mathcal{P}_{\text{Camera}}$, as an input and outputs the corresponding projected 2D skeletons. This module consists of three parts. The first part involves 3D skeleton regression. This stage accepts 3D mesh vertices from the text-to-motion network and uses the joint regressor reported in [40] to regress joints from the mesh vertices. We can formulate this stage as shown below where $V_i^{3D} \in \mathbb{R}^3$ denotes the i^{th} vertex of the mesh, $J_i^{3D} \in \mathbb{R}^3$ denotes the i^{th} joint regressed from the mesh and J_{reg} is the joint regression matrix.

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

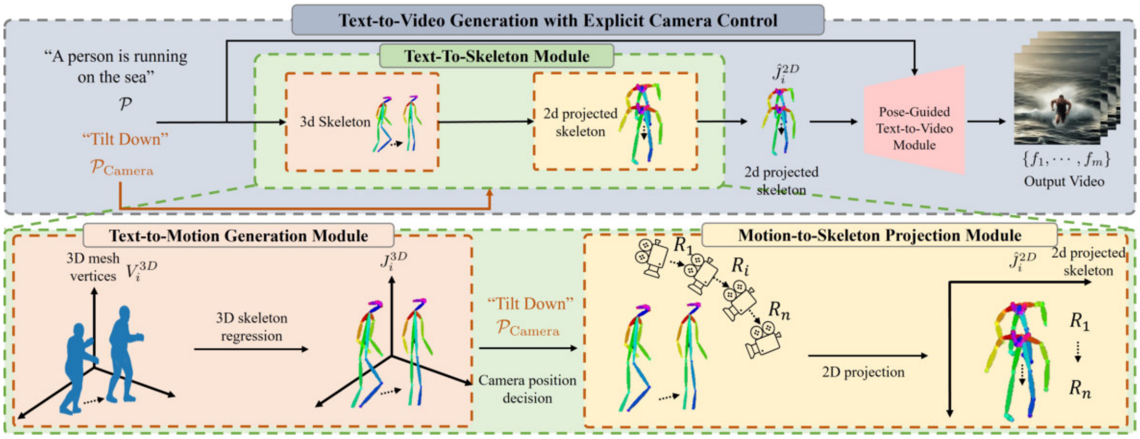


Figure 2. **Overall process of our proposed framework. Top**: Our framework consists of two stages: (1) Text-to-skeleton generation, (2) Skeleton-guided Text-to-Video Generation. A text prompt is passed to the text-to-human motion generation network to generate 3D mesh vertices of each frames of motion. Then, with camera direction description prompt, motion-to-skeleton projection module converts these vertices to the skeletons and project to 2D space corresponding to the camera direction prompt. The last stage, (2) Skeleton-guided text-to-video generation, we use text-to-video network with 2D projected skeletons from motion-to-skeleton projection module and generates the output video corresponds to input prompt $\mathcal{P}$. **Bottom**: Motion-to-skeleton projection module module in detail. Motion-to-skeleton projection module 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, motion-to-skeleton projection module project the 2D skeletons with the projection matrix determined by prompt $\mathcal{P}_{\text{Camera}}$.

In the second part, the camera position is changed using a camera prompt. We can express rotation and translation using a camera extrinsic matrix containing homogeneous coordinate, as expressed below.

$$\begin{pmatrix} R_{3\times3} & t_{3\times1} \\ 0_{1\times3} & 1_{1\times1} \end{pmatrix}_{4\times4} \tag{3}$$

Note that $R_{3\times3}$ defines the rotation of a camera and $t_{3\times1}$ defines the translation of the camera. An intrinsic matrix in combination with the extrinsic matrix is used to define a projection matrix P_{proj} as follows.

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

We pre-defined the textual descriptions and corresponding directions, and the motion-to-skeleton projection module uses the lookup table to decide camera position. The final step involves 2D projection based on camera rotation and translation matrices. Using the determined P_{proj}, we can project the 3D skeleton onto 2D space using a homogeneous coordinate system as follows:

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

The final output of the motion-to-skeleton projection module is a direction-aware 2D-projected skeleton $\hat{J}_i^{2D}$. Notably, $\mathcal{P}_{\text{Camera}}$ need not be used to decide the camera position. If no textual description related to camera position is available, then an identity matrix is used as the camera extrinsic matrix.

3.2. Skeleton-guided Text-to-video Generation

The output of the second stage is provided to a T2V network, which uses the 2D skeleton from the motion-to-skeleton projection module as a guide. Let G be a T2V network and γ be its parameter. Using the given 2D skeleton from the motion-to-skeleton projection module $\hat{J}_i^{2D}$, we obtain 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

4.1. Experimental Results

In this section, we describe the three main experiments conducted and analyze the results. First, we compare the

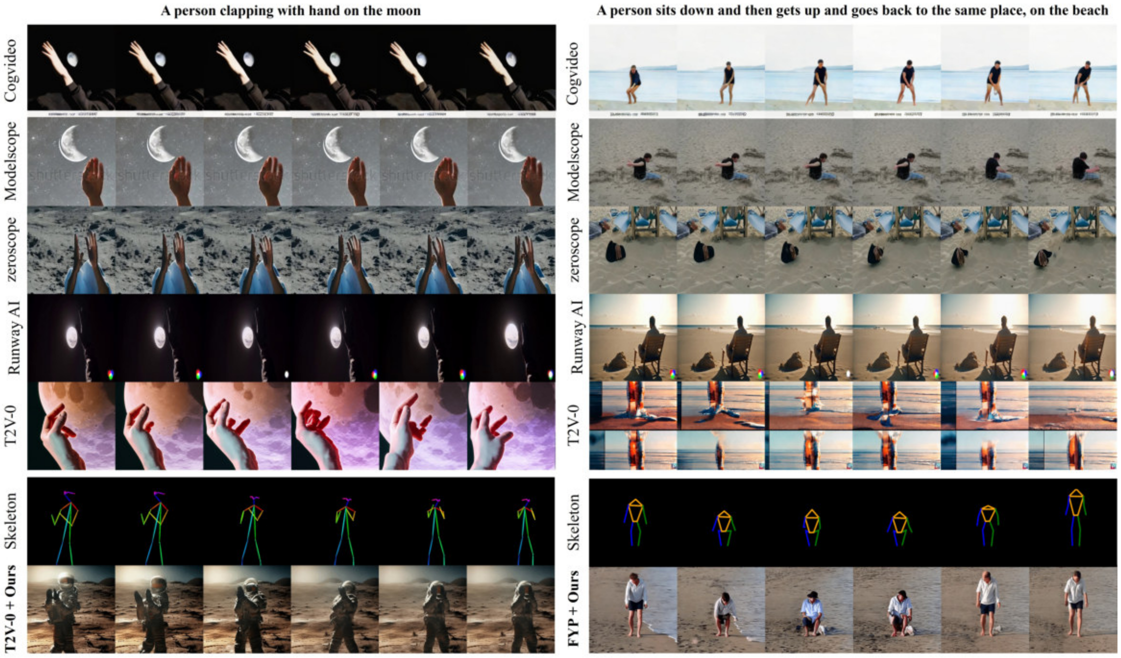

A person clapping with hand on the moon
A person sits down and then gets up and goes back to the same place, on the beach
Cogvideo
Modelscope
zeroscope
Runway AI
T2V-0
Skeleton
T2V-0 + Ours
Cogvideo
Modelscope
zeroscope
Runway AI
T2V-0
Skeleton
FYP + Ours

Figure 3. This figure shows a comparison between the T2V models and our frameworks. **Left**: human face disappears in Zeroscope [27]. Scale problems occur for the rest of the models, creating a large hand. **Right**: Given a complex prompt, ModelScope [26], Zeroscope [27] and T2V-Zero [8] do not generate a proper person. In case of CogVideo [29] and Runway-Gen2 [28] the motions are not generated properly, but our framework is robust to the complex text prompt.

results obtained from the T2V stage in the presence and absence of pose guidance from the first stage. Here, we applied Modelscope [26], Text2Video-Zero (T2V-Zero) [8], Follow Your Pose (FYP) [7], Runway-Gen2 [28], Zeroscope [27], Cogvideo [29], as the T2V networks. In the case of Runway-Gen2 [28], only visualization results were included due to the limited number of accounts for using the model. Second, we compare the videos generated two different T2M networks, namely, T2M-GPT [16] and MDM [14]. Third, we tested the proposed framework using a camera prompt. We used static shot (default), top view, lateral view, and zoom in/out as prompts to describe the camera positions. The prompts we used in action classification (AC), CLIP score (CS), and frame consistency (FC) were conducted through the prompt set that we decided on separately, the details of which are included in the supplementary information.

4.2. Evaluation Metrics

Action Classification (AC) accuracy AC accuracy is the ratio of a well-classified video to the entire generated video. This ratio measures the extent to which the generated videos match with the action in the prompts. To evaluate the extent to which text prompt $\mathcal{P}$ is aligned with the video output, we used the action classification model Text4Vis [41] to determine the AC based on the classes ”jump”, ”run”, ”climb”, ”kick”, ”punch”, ”clap”, ”golf”, and ”sit”.

CLIPscore (CS) [42] CS measures the extent to which the generated videos are aligned with the text prompts. In precise manner, it is a metric that represents the extent to which a caption matches an image without relying on human annotations. Let I be an input image, C be a corresponding caption, and E_I, E_C be embeddings within the image and caption, respectively. Then, the CLIP score is defined as follows:

$$\text{CLIPScore}(I, C) = \max(100 * \cos(E_I, E_C), 0)$$

where the CLIP score is between $[0, 100]$. The closer the score is to 100, the better alignment.

Frame Consistency (FC) [43] FC is the average value of the cosine of the similarities between all consecutive pairs of CLIP image embeddings on all frames. This measures the extent to which naturally generated frames change. This metric is in the range of -1, 1, similar to the range of values of the cosine function. The closer the score is to 1, the better the result.

4.3. Quantitative Results

Table 1 shows the quantitative results from T2V models which obtained with and without pose guidance, based on AC, FC [43] and CS [42]. Note that the use of pose guidance is not necessary in T2V-Zero [8]. As shown in table 1,

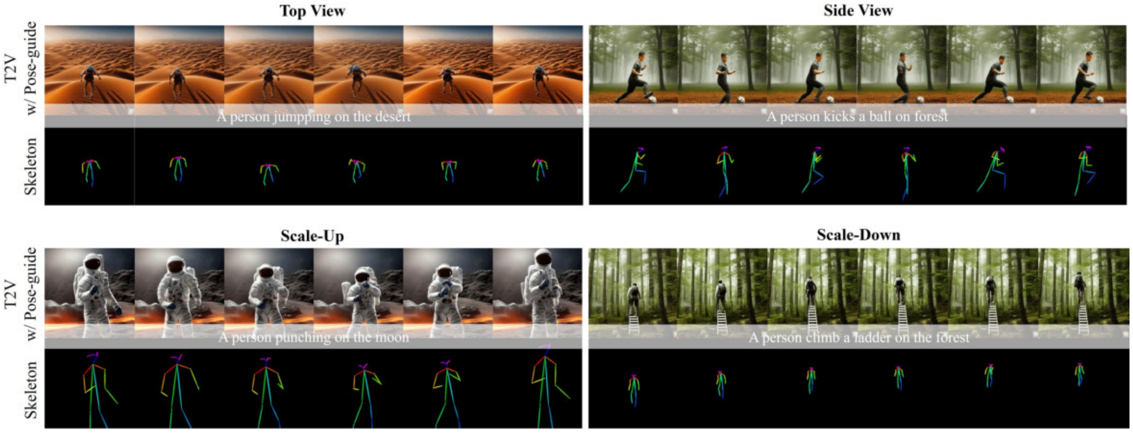

Top View
Side View
T2V w/ Pose-guide
Skeleton
A person jumping on the desert
A person kicks a ball on forest
Scale-Up
Scale-Down
T2V w/ Pose-guide
Skeleton
A person punching on the moon
A person climb a ladder on the forest

Figure 4. This figures shows that using motion-to-skeleton projection module, 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.

Table 1. Quantitative comparison between various text-to-video networks on action classification (AC) accuracy, frame consistency (FC) [43], CLIPscore (CS) [42]. If we proceed with our framework, we can see that the performance is better than other T2V models. In addition, in the case of Text2Video-Zero [8], it can be seen that adding our framework increases all performance.

T2V model	Metric		
	AC↑	FC↑	CS↑
ModelScope [26]	32.5%	88.9%	30.1
Cogvideo [29]	35.2%	90.8%	-
Zeroscope [27]	34.6%	92.2%	30.3
Text2Video-Zero [8]	44.1%	81.7%	28.4
Text2Video-Zero + Ours [8]	47.8%	**92.2%**	29.9
Follow Your Pose + Ours [7]	**48.9%**	87.5%	**30.4**

Table 2. Quantitative results comparison between two different text-to-motion networks using pose guidance. It can be seen that even using various T2M models shows better performance than the existing T2V models.

T2M model	Metric		
	AC↑	FC↑	CS↑
T2M-GPT [16]	47.8%	92.2%	29.9
MDM [14]	46.0%	92.1%	30.2

Table 3. Quantitative results applying camera rotation and skeleton scaling on motion-to-skeleton projection module with pose guidance.

Technique	Metric		
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	AC↑	FC↑	CS↑
Default	51.9%	92.8%	30.3
Zoom in	48.0%	92.7%	29.6
Zoom out	45.2%	92.7%	29.3

based on T2V-Zero [8] the presence of our framework outperform the results on AC accuracy, FC [43] , and CS [42] than T2V-Zero [8] without ours. Moreover, using our framework with FYP [7], we obtained the state-of-the-art performance in T2V tasks. This demonstrates that our framework has a significant role in T2V task. Table 2 shows the effect of T2M networks on our framework, experimented by fixing the FYP [7] of the second stage of our framework. Even if MDM [14] is used, which has lower performance than T2M-GPT [16], the performance of T2V network using our framework is better. Table 3 shows metrics that various camera prompts applied to motion-to-skeleton projection module. In case of camera rotation, top view shows the best performance in overall metrics. In skeleton scale it rather shows worse results than not using scaling, since T2V models are not tend to generate small person. Moreover, in case of too large subjects they do not generate videos too.

4.4. Qualitative Results

Fig. 3 shows the comparison between various T2V models and our frameworks. In the images on the left of Fig. 3, a face disappears in the results using Zeroscope [27], and a scale problem occurs for the rest of the models such as creating a giant hand. Given a complex prompt, ModelScope [26], Zeroscope [27] and T2V-Zero [8] do not generate a person properly. And in case of CogVideo [29] and Runway-Gen2 [28] do not generate human motion aligned with the text prompt. However, using our framework, we generate human motion which aligned well with the given text prompt. We control camera direction using motion-to-skeleton projection module and the results are shown in Fig 4. A motion like “jump” is shown to be represented well using top view. Moreover, a motion like “kick” tends to have better representation at the lateral view. These implies with adequate viewing direction is provided, the better quality of outputs. Therefore, controlling the viewing direction is important work to be studied. Figure 6. shows the problems that we mentioned before. In case of motion ambi-

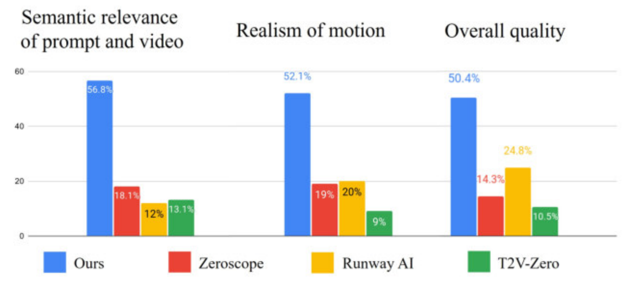


Figure 5. The above figure shows the user study results for Zeroscope, Runway AI, T2V-Zero, and Our. We evaluated four models based on a total of three, and obtained 56.8% and 52.1% and 50.4% results for Semantic reference of Prompt and video, Realism of motion, and Overall quality, respectively. For the rest of the models, we have up to 24.5% performance for quality, but we can see that our results are overwhelming.

guity, the figure shows that our results shows more plausible motion since adequate projections are applied. In case of scale problem, this problem does not occur since we can project skeleton to the scale what we want. Finally in case of temporal consistency problem, since the skeletons consisting human motions are generated in continuous and smooth manner, we can have the better quality than before.

User Study We show the results of four T2V models applying the the same prompts to 100 users and let them evaluate the video according to three criteria: semantic relevance of prompt and video, realism of human motion, overall quality. Fig 5 visualizes the results. 56.8% of the evaluators judged that the video applied to our network was better in terms of semantic relevance of prompt and video, 52.1% of the evaluators preferred our results in terms of realism of human motion, And 50.4% of respondents said that in terms of overall quality, the video that went through our framework was better overall. The video applying our algorithm's motion-to-skeleton projection module was also shown and evaluated by the user. The results of applying motion-to-skeleton projection module were shown to the user and the effect was predicted, and a total of 49.8% of users guessed correctly the motion-to-skeleton projection module applied to the picture. The details appear in the supplementary materials.

4.5. 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 [8] 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. 7: **Top**, the generated motions from text-to-motion network does not align to the prompt resulting mis-aligned output video. Moreover, in Fig. 7: **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 study, we addressed the problem associated with T2V models. Previous T2V models cannot generate suitable text-aligned outputs, including human motions. There are three main problem in previous T2V models. First, unintended or ambiguous motions are generated. Second, a video containing inadequate scale which poorly represents the text are generated. Third, a temporal consistency between frames in videos are not guaranteed. We solve this problem using our proposed framework. Our framework consists of two stages. The first stage is the text-to-skeleton module that generate projected skeletons from the T2M networks. This stage generate human motion from the prompt using T2M networks. Then, the motion-to-skeleton projection module projects generated skeletons with predefined viewing descriptions and camera parameters loop-up table. The second stage is pose-guided text-to-video generation which use pose guided T2V networks to exploit generated skeletons from the previous stage and the text prompt from the first stage to generate human motion included videos. Our frameworks outperforms previous T2V models in the quantitative manner. Moreover, even in qualitative results, the problems mentioned before do no appear. To the best of our knowledge, our framework is the first of its kind that uses text-driven-motion-generation networks to generate high-quality videos related to human motions. We hope that our research would have a positive impact on the subsequent studies and applications involving T2V tasks.

6. Acknowledgement

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2023-RS-2023-00260091) supervised by the IITP(Institute for Information Communications Technology Planning Evaluation), National RD Program through the National Research Foundation of Korea(NRF) funded by Ministry of Science and ICT(2021M3H2A1038042) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A3A01098940, 2021R1I1A1A01051225).

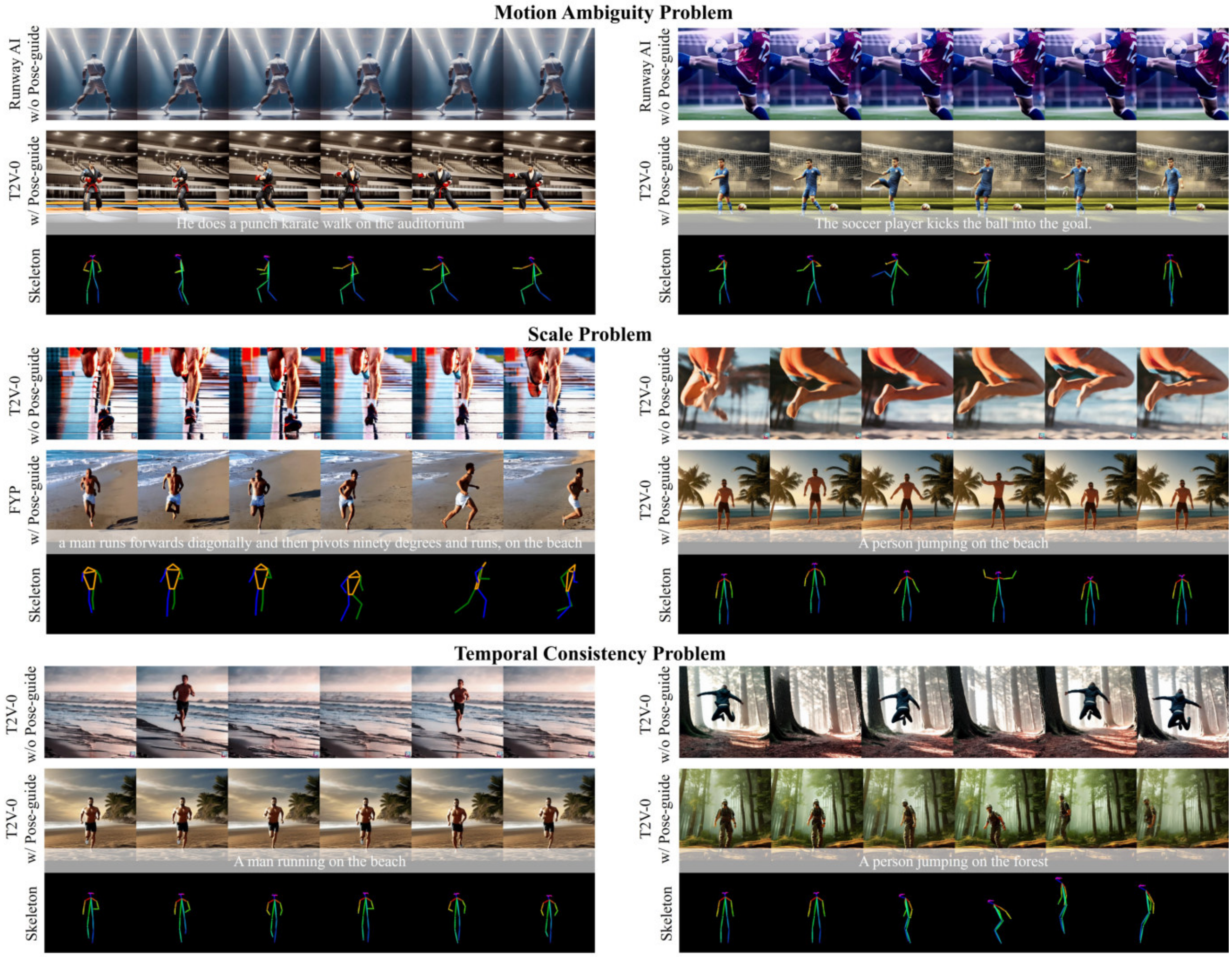

Motion Ambiguity Problem
Runway AI w/o Pose-guide
T2V-0 w/ Pose-guide
He does a punch karate walk on the auditorium
The soccer player kicks the ball into the goal.
Skeleton
Scale Problem
T2V-0 w/o Pose-guide
FYP w/ Pose-guide
a man runs forwards diagonally and then pivots ninety degrees and runs, on the beach
T2V-0 w/ Pose-guide
A person jumping on the beach
Skeleton
Temporal Consistency Problem
T2V-0 w/o Pose-guide
T2V-0 w/ Pose-guide
A man running on the beach
A person jumping on the forest
Skeleton

Figure 6. This figure shows problems of motion ambiguity, scale, and temporary consistency, which are the problems of existing T2V Model. At the top, despite given text prompts containing kick and punch, Runway-Gen2 [28] generate static human. At the middle, in case of text prompt containing words run and jump, the scale objects are too large which is a problem. At the bottom, a person is disappearing in the middle of the frames which is inconsistent. However, using our framework these problems are not occurred.

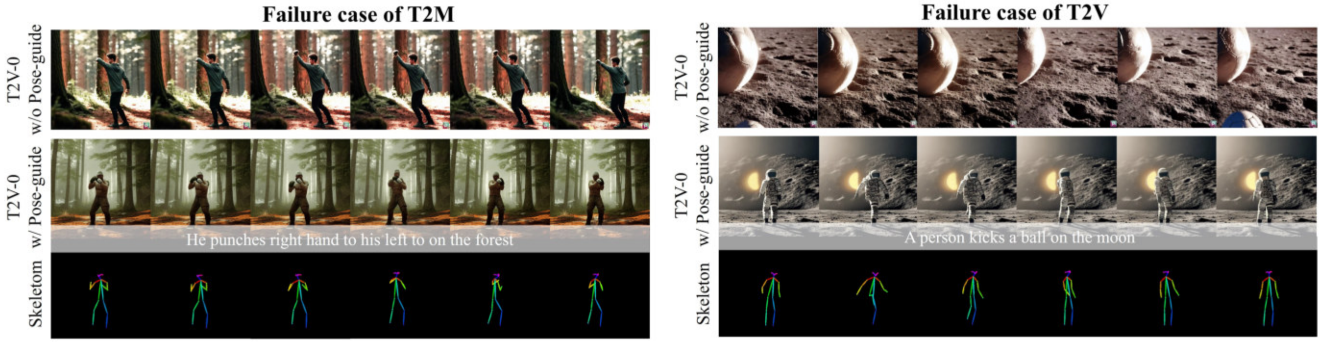


Figure 7. This figure shows the limitation for T2M and T2V. Look at the picture on the left, We can see that the motion of punching comes out only from the front. It is the limitation of the T2M that cannot control the camera to express the punch well, and the right is the limitation of the T2V, which explains that the background does not change according to the movement.

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, 3, 5, 6

[8] 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. 1, 3, 5, 6, 7

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

[18] 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

[19] 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

[20] 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

[21] 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

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

[23] 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

[24] 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

[25] 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

[26] ModelScope: bring the notion of model-as-a-service to life. https://github.com/modelscope/modelscope. Accessed: 2023-06-28. 2, 5, 6

[27] cerspense/zeroscope_v2_576w. https://huggingface.co/cerspense/zeroscope_v2_576w. Last Updated: 2023-07-01. 2, 5, 6

[28] Gen-2 by Runway. https://research.runwayml.com/gen2. 2, 5, 6, 8

[29] 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. 2, 5, 6

[30] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 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] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. Pose guided human video generation. In *European Conference on Computer Vision*, 2018. 3

[37] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. Towards image-to-video translation: A structure-aware approach via multi-stage generative adversarial networks. *International Journal of Computer Vision*, 128:2514 – 2533, 2020. 3

[38] 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

[39] 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

[40] 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. 3

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

[42] 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, 6

[43] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models, 2023. 5, 6