Supplementary Material for *NewMove: Customizing text-to-video models with novel motions*

Joanna Materzynska[1,2], Josef Sivic[2,3], Eli Shechtman[2], Antonio Torralba[1], Richard Zhang[2], and Bryan Russell[2]

[1] Massachusetts Institute of Technology
[2] Adobe Research
[3] CIIRC CTU

We refer the reader to our website of qualitative results of our method, including video results shown in the main paper as well as comparison with the pre-trained text-to-video generations. Please allow video results to load on the webpage. This supplemental contains the following material:

- (Section 1) We provide additional implementation details of our approach and our quantitative experiments.
- (Section 2) We qualitatively demonstrate how our method can be extended to customizing both appearance and motion.
- (Section 3) We demonstrate results of our method applied in one-shot setting.
- (Section 4) We discuss the tradeoff between the motion accuracy, appearance alignment, and copying score across training iterations.
- (Section 5) We show qualitative examples along with their copying score.
- (Section 6) We show the original model's (*i.e.*, before customization) performance on the Jester dataset.
- (Section 7) Comparison with concurrent methods (MotionDirector,AnimateDiffv1)
- (Section 8) We qualitatively show effectiveness of our method across two text-to-video backbones, ZeroScope and OpenSora.

1 Implementation Details

In all our experiments we use the ZeroScope text-to-video diffusion model [1] as the pre-trained network. Our implementation is building on the public repository [2]. In our concept exemplar dataset D^m, we apply spatial and color augmentation on the video examples. We train the model with a mini-batch size 1 on a single A100 Nvidia GPU. We train the models up to 3000 training steps with learning rate 5×10^{-6}. The video's spatial resolution is 384×384, and temporal 16. At inference, we use 50 diffusion time steps, guidance scale 9, we do not use negative prompt. Training the model on a single A100 GPU instance, for 3000 steps with a batch size of 1, takes 90 minutes and uses 21.3 GB of GPU memory.

As described in Section 4.2 for our quantitative analysis on the Jester Dataset, we design a test set of 100 prompts detailing a person's appearance and fix three random seeds per prompt. We show the list of the prompts in Figure 5. We discuss the GPU cost analysis across different baselines in Table **??** shows compute and memory analysis on a single V100 GPU. Inference time, based on 50 denoising steps for 16-frame videos, shows our method offers the best trade-off between motion fidelity and efficiency.

Method	Train/Infer. (min/sec)	GPU Mem. (GB)
Textual Inversion	85/28	11.7
Custom Diffusion	145/28	12.5
Dreambooth	170/28	24.1
Tune-a-Video	10/23	9.0
Pix2Video	-/90	10.0
StableVideo	600/30	10.0
Diffusion Motion Transfer	-/1020-1500	11.7
Ours	170/28	21.3

Table 1. Compute and memory costs for baseline methods.

Spatial	Temporal	Train/Infer. (min/sec)	GPU Mem. (GB)
All	None	170/28	24.1
All	All	191/28	31.4
None	All	170/28	21.0
(Ours) k,v	All	170/28	21.3
LoRA	LoRA	130/28	11.8

Table 2. Compute and memory costs for ablations and our method.

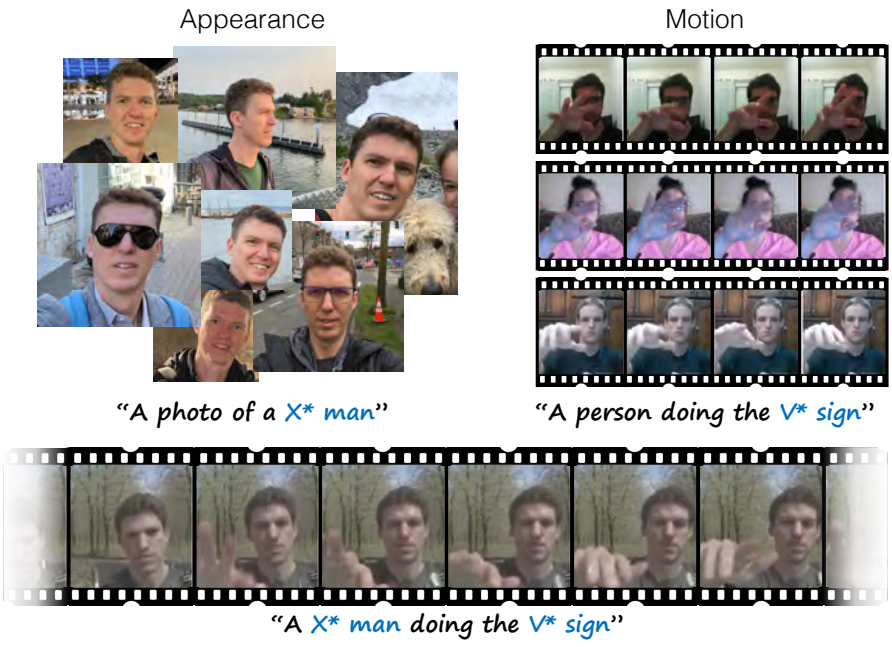

Fig. 1. Appearance and motion customization. Given few images of a person from the CustomConcept101 Dataset [5] and few videos of a custom motion (top), our method can generate a novel video depicting that person doing the novel motion (bottom).

2 Customizing Appearance and Motion

Our method can be easily extended to customizing both appearance and motion. The appearance of a new subject is represented through an exemplar dataset D^s of pairs of still images and captions describing the images. To learn the appearance of a person we follow [5] and update the spatial layers of the model, and optimize the text embedding associated with our subject. We customize the model first with the motion, then train both the appearance and the subject jointly, choosing an example from either the motion D^m or subject D^s datasets in the mini-batch.

We combine the appearance of a specific person with a gesture motion *Drumming Fingers* from the Jester Dataset. We use the publicly available images of a particular person from the CustomConcept101 Dataset [5], and annotate them with generic text captions {"A X* man", "A close-up of a X* man", "A X* man smiling."}. For the motion dataset, we choose the same 10 videos as described in Section 4.2 and a generic text description "a person doing a V* sign". Our qualitative results are shown in Fig. 1.

3 One-shot customization

We evaluated our method using a single training video input. Specifically, we trained the model on a video of the "Carlton dance," applying spatial and color video augmentations, for 1000 training steps. As shown in Fig 2, our approach generalizes the input motion to multiple individuals and to different subjects.

Fig. 2. One-shot customization Our method can be applied using a single-video demonstration of a novel motion. Given one video showing a single person performing the "Carlton dance", our method can produce novel videos with one (middle and bottom row; *A toddler dancing the V* dance*, *A fat man dancing the V* dance*) or multiple actors (top row; *Cowboys dancing the V* dance*) performing a novel motion.

4 Tradeoff between motion accuracy, appearance alignment, and copying score

When customizing a model with a new motion, we observe a natural bias-variance tradeoff. The longer we train the model, the better motion accuracy becomes, we memorize elements of the training examples, and the model becomes less diverse. We illustrate this in Figure 3. With more training iterations, the network memorizes the gesture as well as visual elements of the training examples. These findings are indicated by both the motion accuracy and copying score increasing. On the other hand, the CLIP appearance score, decreases. The CLIP appearance is defined as an average CLIP score between the part of the text prompt that describes a person's appearance and the video frames (i.e. in a prompt "A female firefighter doing the V* sign", the CLIP appearance score accounts for "A female firefighter"). This score is a proxy for visual diversity, since our test set contains visual description of a person, it measures how accurate the description is. As we increase the number of iterations, the videos become less diverse, hence the score decreases. This trend is persistent across the models from our ablation study. To find a balance between those factors, we perform early stopping for all models when the CLIP appearance score reaches 0.265 (i.e. when the image still reasonably well corresponds to the text description).

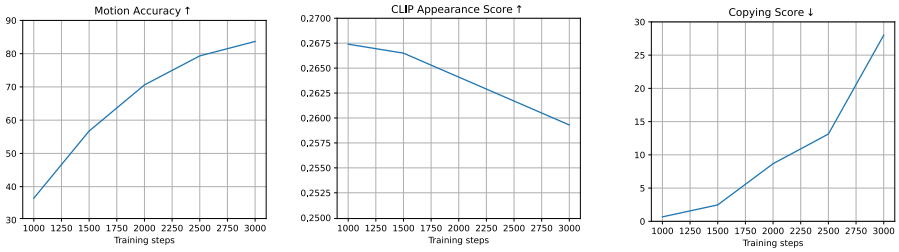

Fig. 3. Our model's performance in terms of motion accuracy (top), CLIP appearance score (middle), and copying score (bottom) across training steps (x-axis).

5 Copying score

Copying score, as defined in 4.2 measures how much of the training data appearance leaks into the generated data. The score corresponds to the percentage of examples in the test set that are above a specified threshold of the SSCD score with any of the training examples.

We find that the SSCD copy detection score [6] is a good proxy for determining the memorization of the training examples' appearance. We measure the SSCD detection as a maximum score between any frame of the training videos and any video frame from the test set. We empirically determine that pairs of videos with a score greater than 0.25 share significant visual similarities. As seen in Figure 4, the similarities can be either the background or the appearance of the person performing the gesture.

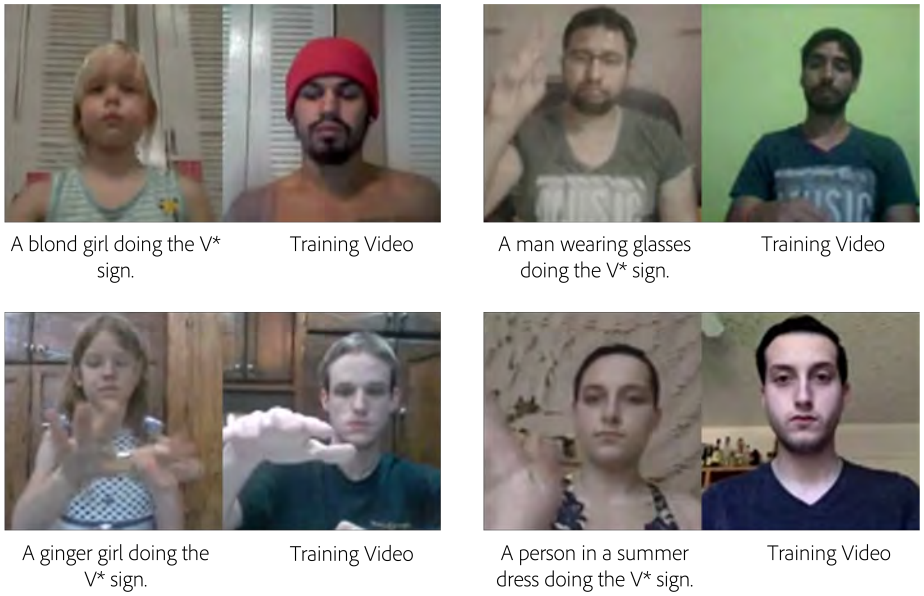

Fig. 4. Examples of pairs of generated frames and training frames with SSCD score 0.25. Notice the similarity in the background (top-left and bottom-left) and foreground (t-shirt logo in the top-right, person's face in the bottom-right).

6 Pre-trained model's performance on motions from the Jester dataset.

To test whether the gestures from the Jester dataset, are well-represented in the pre-trained text-to-video model, we evaluate the model with a similar experiment as our ablation experiment (Section 4.2, Table 2).

We query the pre-trained text-to-video model with the prompts from the test set (Section 1), replacing the "doing the sks gesture" with a natural language description of the class, eg, for the class "Drumming Fingers" - "drumming finger". We then classify the videos with the pre-trained network trained on the Jester dataset.

No videos were correctly classified because (i) the gestures are not well represented in the model and (ii) the generated videos do not adhere to the spatial assumption of the Jester dataset (where a single person is performing a gesture in front of a camera).

7 Comparison with concurrent methods (MotionDirector, AnimateDiff v1)

We qualitatively compare against MotionDirector [9] and AnimateDiff v1 [4] on selected examples in Fig 6.

MotionDirector is a recent unpublished, concurrent work [9] that has the most similar objective to ours and a publicly available implementation. AnimateDiff v1 [4] is another concurrent, unpublished work that enhances personalized text-to-image models with a motion module. For MotionDirector, we use the same training set as in our experiments and refer to the "Carlton dance" with V^* as used in our training. For AnimateDiff v1, we use *realisticVisionV20_v20.safetensors* as a personalized text-to-image model and *mm_sd_v15.ckpt* as a motion module from their publicly available demo. For AnimateDiff v1 and the ZeroScope base model, we use verbatim "the Carlton dance" in the text prompt.

First, we illustrate (top rows, Fig 6) that the original pre-trained text-to-video model ZeroScope, used in both our method and MotionDirector, has no concept of "the Carlton dance".

When directly using MotionDirector with the default parameters (1000 training steps), we can see that the method does not perform as well (second rows, Fig 6). For a fair comparison we train MotionDirector for more iterations to match our setup (3000 steps) and see that it starts to pick up the "Carlton dance" characteristics (third rows, Fig 6). We can see that AnimateDiff v1 also does not know the concept of the Carlton dance (fourth rows, Fig 6). Our method can successfully render the novel motion (bottom rows, Fig 6).

Secondly, we quantitatively compare to MotionDirector on the Jester task as described in Section 4.2 (main paper) (Table 3).

We use the same training videos and train the models with the default parameter settings as described in [9]. Our method performs significantly better in terms of both motion accuracy and text alignment.

	Motion Accuracy ↑	Text Alignment ↑
Textual Inversion [3]	0.3	0.2733
Custom Diffusion [5]	10.5	0.2788
Dreambooth [7]	28.4	0.2796
Tune-a-Video [8]	18.9	**0.2818**
MotionDirector [9]	24.7	0.2786
Ours	**70.6**	**0.2818**

Table 3. Quantitative comparison with baseline methods applied to the motion customization task. Our approach achieves the highest scores for motion accuracy and text alignment compared to the baselines.

8 Comparison of our method across different pretrained backbones

To show the effectiveness of our method on a different text-to-video backbone, we adapt the OpenSora model and show the qualitative results in Fig 7. Using the open-source code and model weights from *https://github.com/hpcaitech/Open-Sora/*, we customized the model to include the "Shaking Hand" gesture from Jester dataset. The results illustrate that our method generalizes well across different pre-trained models.

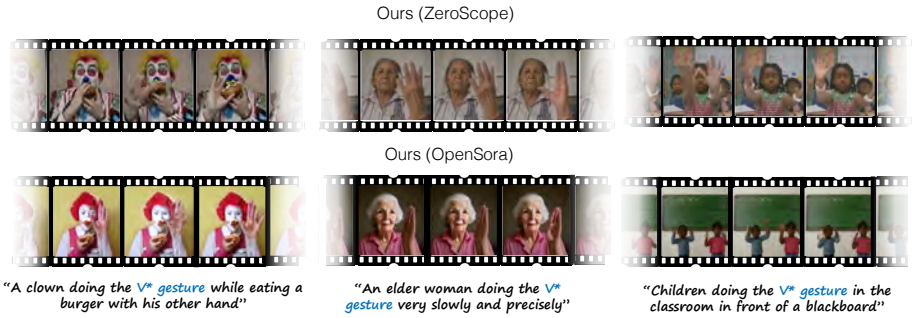

Fig. 7. Qualitative comparison of our method across different pre-trained models shows that the customized motion can generalize at the test time to multiple actors, motion variations, and different actions, regardless of the backbone model used.

A tribe man doing the sks sign.
A young woman doing the sks sign.
A korean girl doing the sks sign.
A korean boy doing the sks sign.
A korean man doing the sks sign.
A korean woman doing the sks sign.
A korean toddler doing the sks sign.
A japanese woman doing the sks sign.
A japanese boy doing the sks sign.
A japanese girl doing the sks sign.
A japanese toddler doing the sks sign.
A japanese man doing the sks sign.
A firefighter doing the sks sign.
A female firefighter doing the sks sign.
A clown doing the sks sign.
Mickey Mouse doing the sks sign.
A housemaid doing the sks sign.
A female nurse doing the sks sign.
A male nurse doing the sks sign.
A toddler doing the sks sign.
An african woman doing the sks sign.
An african man doing the sks sign.
An african boy doing the sks sign.
An african toddler doing the sks sign.
An elderly woman doing the sks sign.
An elderly man doing the sks sign.
A mexican man doing the sks sign.
A mexican woman doing the sks sign.
A mexican boy doing the sks sign.
A mexican girl doing the sks sign.
A mexican toddler doing the sks sign.
A slavic man doing the sks sign.
A slavic woman doing the sks sign.
A slavic boy doing the sks sign.
A slavic girl doing the sks sign.
A slavic child doing the sks sign.
A blond woman doing the sks sign.
A blond man doing the sks sign.
A blond boy doing the sks sign.
A blond girl doing the sks sign.
A ginger woman doing the sks sign.
A ginger man doing the sks sign.
A ginger boy doing the sks sign.
A ginger girl doing the sks sign.
A woman with glasses doing the sks sign.
A man with glasses doing the sks sign.
A girl with glasses doing the sks sign.
A boy with glasses doing the sks sign.
A child with glasses doing the sks sign.
A man with a beard doing the sks sign.
A teacher doing the sks sign.
A woman doing the sks sign.
A musician doing the sks sign.
A chef doing the sks sign.
A construction worker doing the sks sign.
A police officer doing the sks sign.
A student doing the sks sign.
A doctor doing the sks sign.
A scientist doing the sks sign.
A farmer doing the sks sign.
A dancer doing the sks sign.
A pilot doing the sks sign.
A yoga instructor doing the sks sign.
A surfer doing the sks sign.
A skateboarder doing the sks sign.
A hiker doing the sks sign.
A painter doing the sks sign.
A photographer doing the sks sign.
A writer doing the sks sign.
A woman with long hair doing the sks sign.
A man with a mustache doing the sks sign.
A woman with a ponytail doing the sks sign.
A man with a bald head doing the sks sign.
A teenager doing the sks sign.
A senior citizen doing the sks sign.
A person in a wheelchair doing the sks sign.
A person with a backpack doing the sks sign.
A person wearing a hat doing the sks sign.
A person in traditional clothing doing the sks sign.
A person in casual attire doing the sks sign.
A person in formal attire doing the sks sign.
A person with tattoos doing the sks sign.
A person with piercings doing the sks sign.
A person with a camera doing the sks sign.
A person holding a book doing the sks sign.
A person using a smartphone doing the sks sign.
A person with a pet doing the sks sign.
A person with a bicycle doing the sks sign.
A person in workout clothes doing the sks sign.
A person in a swimsuit doing the sks sign.
A person wearing a backpack doing the sks sign.
A person wearing a business suit doing the sks sign.
A person wearing a lab coat doing the sks sign.
A person wearing a uniform doing the sks sign.
A person in a winter coat doing the sks sign.
A person in a summer dress doing the sks sign.
A person in a cowboy hat doing the sks sign.
A person in a graduation gown doing the sks sign.
A person in a superhero costume doing the sks sign.
A person in a traditional robe doing the sks sign.

Fig. 5. Test set prompts for the quantitative evaluation of learning the Jester motions

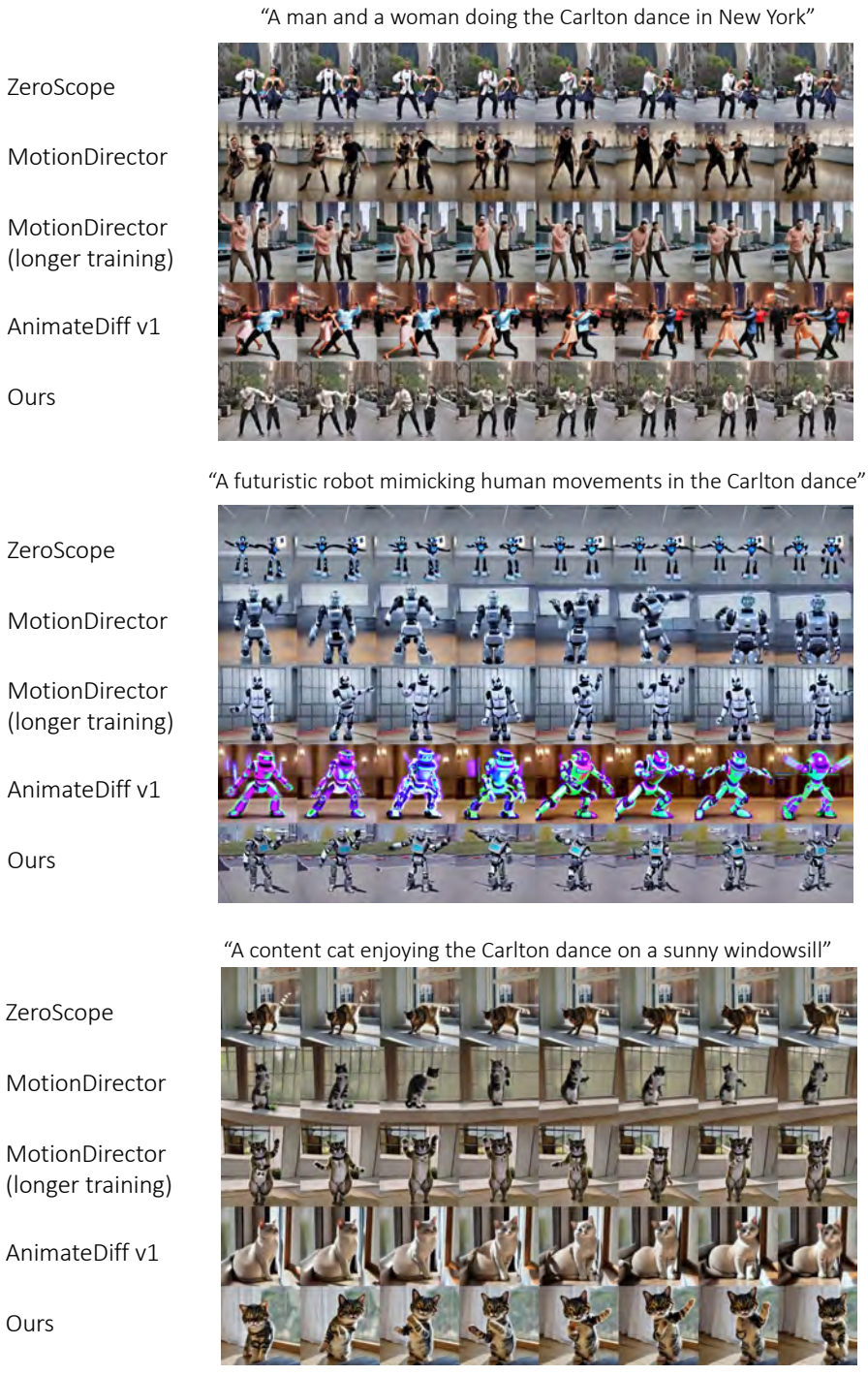

Fig. 6. Qualitative comparison of different motions across different methods.

References

1. cerspense: zeroscope_v2_xl. https://huggingface.co/cerspense/zeroscope_v2_XL/ (2023) 1
2. ExponentialML: Text-to-video-finetuning (2023), https://github.com/ExponentialML/Text-To-Video-Finetuning, text-To-Video-Finetuning 1
3. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: Proceedings of the International Conference on Learning Representations (ICLR) (2023) 6
4. Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning (v1). arXiv preprint arXiv:2307.04725 (2023) 6
5. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 2, 3, 6
6. Pizzi, E., Roy, S.D., Ravindra, S.N., Goyal, P., Douze, M.: A self-supervised descriptor for image copy detection. Proc. CVPR (2022) 4
7. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023) 6
8. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023) 6
9. Zhao, R., Gu, Y., Wu, J.Z., Zhang, D.J., Liu, J., Wu, W., Keppo, J., Shou, M.Z.: Motiondirector: Motion customization of text-to-video diffusion models. arXiv preprint arXiv:2310.08465 (2023) 6