Learning to Animate Images from A Few Videos to Portray Delicate Human Actions

Appendix

The Appendix is structured as follows:

- Section A1 includes supplementary examples comparing videos generated by advanced AI video generators.
- Section A2 elaborates on the details of FLASH.
- Section A3 describes detailed experimental setups.
- Section A4 provides more experimental results.
- Section A5 discusses the limitations of FLASH.
- Section A6 presents the Ethical Statement.

A1. Comparison of videos generated by advanced AI video generators

In Figure A1, we present four examples of animated human action videos from Dream Machine[1], KLING AI[2], Wan[3], and FLASH. The videos are also available on the webpage[4]. Dream Machine, KLING AI and Wan struggle to animate these actions accurately. In the balance beam jump action, Dream Machine and Wan produce unrealistic, physics-defying movements, while KLING AI generates a jump but fails to depict standard jumps on the balance beam. For the soccer shooting action, all three models fail to generate a correct shooting motion, with the person never kicking the ball. In the shoot dance action, Dream Machine and KLING AI generate unnatural, physically implausible movements, whereas Wan produces dance movements but does not capture the shoot dance correctly. In the Ice Bucket Challenge action, none of the three models accurately portray the motion of pouring ice water from the bucket onto the body. In contrast, FLASH generates these actions with higher fidelity to the real actions.

A2. FLASH

A2.1. Components in Latent Video Diffusion Models

Temporal Attention Layers. To capture temporal dynamics in videos, [12, 25, 26, 34] add temporal attention layers after each spatial attention layer of U-Net. In each temporal attention layer, we first reshape the input features $\boldsymbol{F}_{in} \in \mathbb{R}^{N\times h'\times w'\times c'}$ to $\tilde{\boldsymbol{F}}_{in} \in \mathbb{R}^{B\times N\times c'}$, where $B = h' \times w'$. Here, we treat the features at different spatial locations as independent samples. Then, we add temporal position encoding to $\tilde{\boldsymbol{F}}_{in}$ and employ a self-attention layer to transform $\tilde{\boldsymbol{F}}_{in}$ into $\tilde{\boldsymbol{F}}_{out} \in \mathbb{R}^{B\times N\times c'}$. Finally, we reshape $\tilde{\boldsymbol{F}}_{out}$ to $\boldsymbol{F}_{out} \in \mathbb{R}^{N\times h'\times w'\times c'}$ as the output features. The temporal attention layer integrates information from different frames for each spatial location, enabling the learning of temporal changes.

Cross-frame Attention Layers. To enhance temporal consistency across generated frames, [44, 98] replace spatial self-attention layers with cross-frame attention layers. While self-attention layers use features from the current frame as key and value, cross-frame attention layers restrict key and value to the features from the reference frame (typically the first frame in image animation tasks). These layers carry over the appearance features from the reference frame to subsequent frames, improving temporal consistency in the generated videos.

Noise-Free Frame Conditioning. To further preserve the appearance of the reference image in image animation tasks, [70, 98] keep the latent reference image noise-free in the noised latent video. Specifically, at the noising step t, the latent video $\boldsymbol{Z}_t = \langle \boldsymbol{z}_t^i \rangle_{i=1}^N$ is modified to $\check{\boldsymbol{Z}}_t = \langle \boldsymbol{z}_0^1, \boldsymbol{z}_t^2, \cdots, \boldsymbol{z}_t^N \rangle$, where $\boldsymbol{z}_t^1$ is replaced by $\boldsymbol{z}_0^1$, which is noise-free. During inference, a sample $\boldsymbol{Z}_T$ is drawn from $\mathcal{N}(\mathbf{0}, I)$, and $\boldsymbol{z}_T^1$ is substituted with $\boldsymbol{z}_0^1 = \mathcal{E}(I)$, where I is the user-provided reference image. The modified latent video $\check{\boldsymbol{Z}}_T = \langle \boldsymbol{z}_0^1, \boldsymbol{z}_T^2, \cdots, \boldsymbol{z}_T^N \rangle$ is then used for denoising. This technique effectively maintain the features from the first frame in subsequent frames.

FLASH adopts these components in its base video diffusion model, and designs the Motion Alignment Module and the Detail Enhancement Decoder on top of it.

A2.2. Strongly Augmented Videos

To create a strongly augmented version of an original video, we sequentially apply Gaussian blur and random color adjustments to the original video. This process is designed to preserve the original motion while altering the appearance uniformly across all frames.

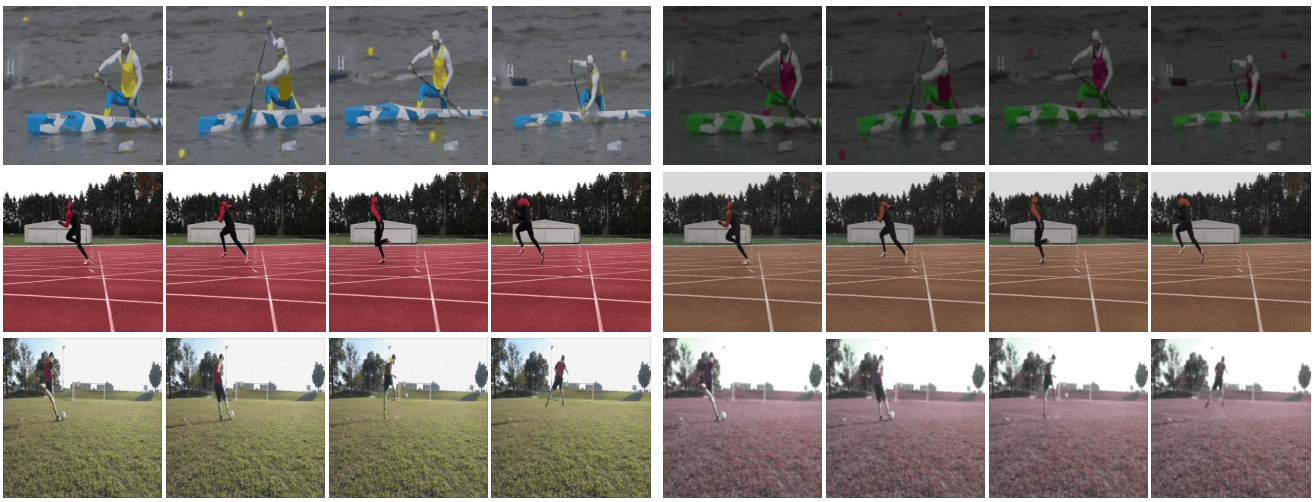
- Gaussian blur: A kernel size is randomly selected from a predefined range, as specified in Sec. A3.2. This kernel size is used to apply Gaussian blur to every frame of the original video, ensuring a uniform level of blur throughout.
- Random color adjustments: After applying Gaussian blur, we randomly adjust the brightness, contrast, saturation, and hue of the video. For each property, an adjustment factor is randomly chosen from its respective predefined range, detailed in Sec. A3.2. The adjustment with the

[1] https://lumalabs.ai/dream-machine
[2] https://www.klingai.com/
[3] https://tongyi.aliyun.com/wanxiang/
[4] https://lihaoxin05.github.io/human_action_animation/

Figure A1. Comparison of human action videos generated by Dream Machine, KLING AI, Wan and FLASH (our method). Human faces are anonymized for privacy protection.

chosen factor is applied uniformly across all frames to maintain consistent color alterations without introducing cross-frame inconsistencies. We implement it with the *ColorJitter* function in PyTorch.

By applying these augmentations with consistent parameters across all frames, the augmented video retains the motion in the original video while showing altered appearances. Figure A2 presents examples of strongly augmented videos. These augmented videos exhibit considerable differences from the original ones in aspects such as the back-



(a) The original videos (b) The augmented videos

Figure A2. Examples of three original videos alongside their corresponding strongly augmented videos.

ground and the actors' clothing. However, the motion from the original videos is preserved.

A2.3. Detail Enhancement Decoder

Multi-scale Detail Propagation. Before feeding the two features $\boldsymbol{g}_l^1$ and $\boldsymbol{h}_l^i$ to the two branches, we first interpolate $\boldsymbol{g}_l^1$ to match the spatial size of $\boldsymbol{h}_l^i$ and use a fully connected layer to adjust $\boldsymbol{g}_l^1$ to the same number of channels as $\boldsymbol{h}_l^i$, resulting $\tilde{\boldsymbol{g}}_k^1$ as the input of the two branches. In the *Patch Attention Branch*, before applying the cross-attention layer $\mathcal{A}$, we use a fully connected layer to transform each patch of the two features into a feature vector.

Distorted Videos. The details of video distortions are as follows: The random Gaussian blur and random color adjustments follow the implementations described in Sec. A2.2. However, the random color adjustments here differ in that they are applied to only 80% of randomly selected regions rather than to all regions. This modification is intentional, as the goal is to create distorted videos with inconsistent color changes that simulate the distortions in latent videos, rather than to maintain consistent color changes as in Sec. A2.2. For random elastic transformations, displacement vectors are generated for all spatial positions based on random offsets sampled from a predefined range (detailed in Sec. A3.2) and are then used to transform each pixel accordingly. We implement it using the *ElasticTransform* function in PyTorch.

A3. Experiment Details

A3.1. Data

We conduct experiments on 16 actions selected from the HAA500 dataset [10], which contains 500 human-centric atomic actions capturing the precise movements of human, each consisting of 20 short videos. The selected actions include single-person actions (push-up, arm wave, shoot dance, running in place, sprint run, and backflip), human-object interactions (soccer shoot, drinking from a cup, balance beam jump, balance beam spin, canoeing sprint, chopping wood, ice bucket challenge, and basketball hook shot), and human-human interactions (hugging human, face slapping).

Training videos. For each selected action, we use 16 videos from the training split in HAA500 for training. We manually exclude videos that contain pauses or annotated symbols in the frames. Each action label is converted into a natural sentence as the action description; for example, the action label "soccer shoot" is converted to "a person is shooting a soccer ball."

Similarity between training videos in the same action class. Videos within the same action class do not need to share similar visual characteristics, such as scenes, viewing angles, actor positions, or shot types (*e.g.*, close-up or wide shot). Instead, they are different on these characteristics, as shown in the examples in Figure A3.

Testing images. For each selected action, we use the first frames from the four testing videos as testing images. Additionally, we search online for two human images depicting a person beginning the desired action as additional testing images.

Similarity between training videos and testing images. We provide the training videos and testing images (4 example actions on the webpage[5]). Most testing images differ

[5] https://lihaoxin05.github.io/human_action_animation/

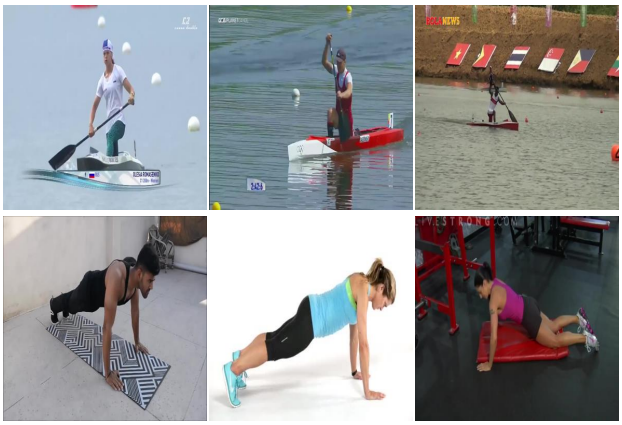


Figure A3. Similarity between training videos in the same action class. The first row presents three videos depicting the action canoeing sprint, and the second row showcases three videos containing the action push-up.

from training frames in viewpoint, subject size and appearance.

A3.2. Implementation Details

We use AnimteDiff [28] as the base video generative model. We initialize all parameters with pretrained weights of AnimteDiff. The spatial resolution of generated videos is set to 512×512. The video length is set to 16 frames, with a frame rate of 16 FPS.

Training of U-Net. For strong data augmentation, Gaussian blur is applied with a randomly sampled kernel size between 3 and 10. Random color adjustment modifies brightness, saturation, and contrast by random factors between 0.5 and 1.5, and modifies hue by a random factor between -0.25 and 0.25. Before applying strong augmentations to the original video, we first perform random horizontal flipping and random cropping on the original video. In the Motion Alignment Module, we set τ to 90 and apply motion feature alignment after each temporal attention layer in the U-Net. We combine features from the first and current frames as keys and values in the cross-frame attention layers. Inter-frame correspondence alignment is applied to 50% of the cross-frame attention layers, selected randomly. For simplicity, we replace $\boldsymbol{Q}$ and $\boldsymbol{K}$ of the augmented video with those of the original video when calculating S, instead of directly replacing S. We only train the temporal attention layers, and the key and value matrices of cross-frame attention layers. Following [39, 60], we redefine the sampling probability distribution to prioritize earlier denoising stages. The learning rate is set to 5.0×10^{-5}, with training conducted for 20,000 steps.

`training_videos_and_testing_images`

Training of Detail Enhancement Decoder. The patch size in the Patch Attention Branch is set to 2. For video distortion, Gaussian blur is applied with a random kernel size between 3 and 10. Random color adjustment use random factors for brightness, saturation, and contrast between 0.7 and 1.3, and a random factor for hue between -0.2 and 0.2. For random elastic transformations, displacement strength is randomly sampled from 1 to 20. We only train the newly added layers, with a learning rate of 1.0×10^{-4} over 10,000 steps.

Hyper-parameter Tuning. We use the same hyper-parameters for all actions without action-specific tuning. We first select these hyper-parameters based on the model performance on 4 actions (see Sec. A4.3 for details). Following this selection, we apply the same hyper-parameter configuration to all actions.

Inference. During inference, we utilize the DDIM sampling process [82] with 25 denoising steps. Classifier-free guidance [31] is applied with a guidance scale set to 7.5. Following [98], we apply AdaIN [37] on latent videos for post-processing.

Computational Resources. Our experiments are conducted on a single GeForce RTX 3090 GPU using PyTorch, with a batch size of 1 on each GPU. We build upon the codebase of AnimateDiff [26]. Training takes approximately 36 hours per action.

A3.3. Evaluation Metrics

In line with previous works [29, 97, 98], we use three metrics based on CLIP [69] to assess text alignment, image alignment, and temporal consistency. (1) *Text Alignment*: We compute the similarity between the visual features of each frame and the textual features of the text prompt, and average the similarities across all frames. (2) *Image Alignment*: We compute the similarity between the visual features of each frame and the visual features of the provided reference image, and average the similarities across all frames. (3) *Temporal Consistency*: We calculate the average similarity between the visual features of consecutive frame pairs to obtain the temporal consistency score. We use ViT-L/14 from OpenAI [69] for feature extraction. In these three metrics, higher scores indicate better performance.

Following [100], we utilize Fréchet distance to compare generated and real videos. We use *CD-FVD* [16] to mitigate content bias in the widely used FVD [88]. We use VideoMAE [87], pretrained on SomethingSomethingV2 [21], for feature extraction and calculate distance between real and generated videos. In this metric, lower distances indicate better performance.

To evaluate the similarity between generated videos and ground-truth videos in the HAA dataset, we calculate the

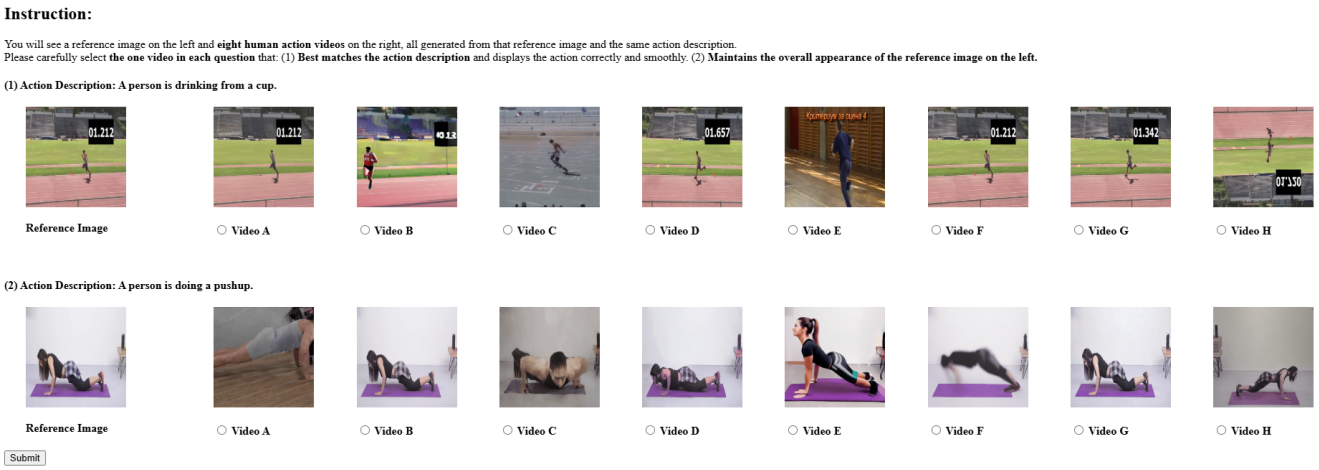


Figure A4. AMT user study interface.

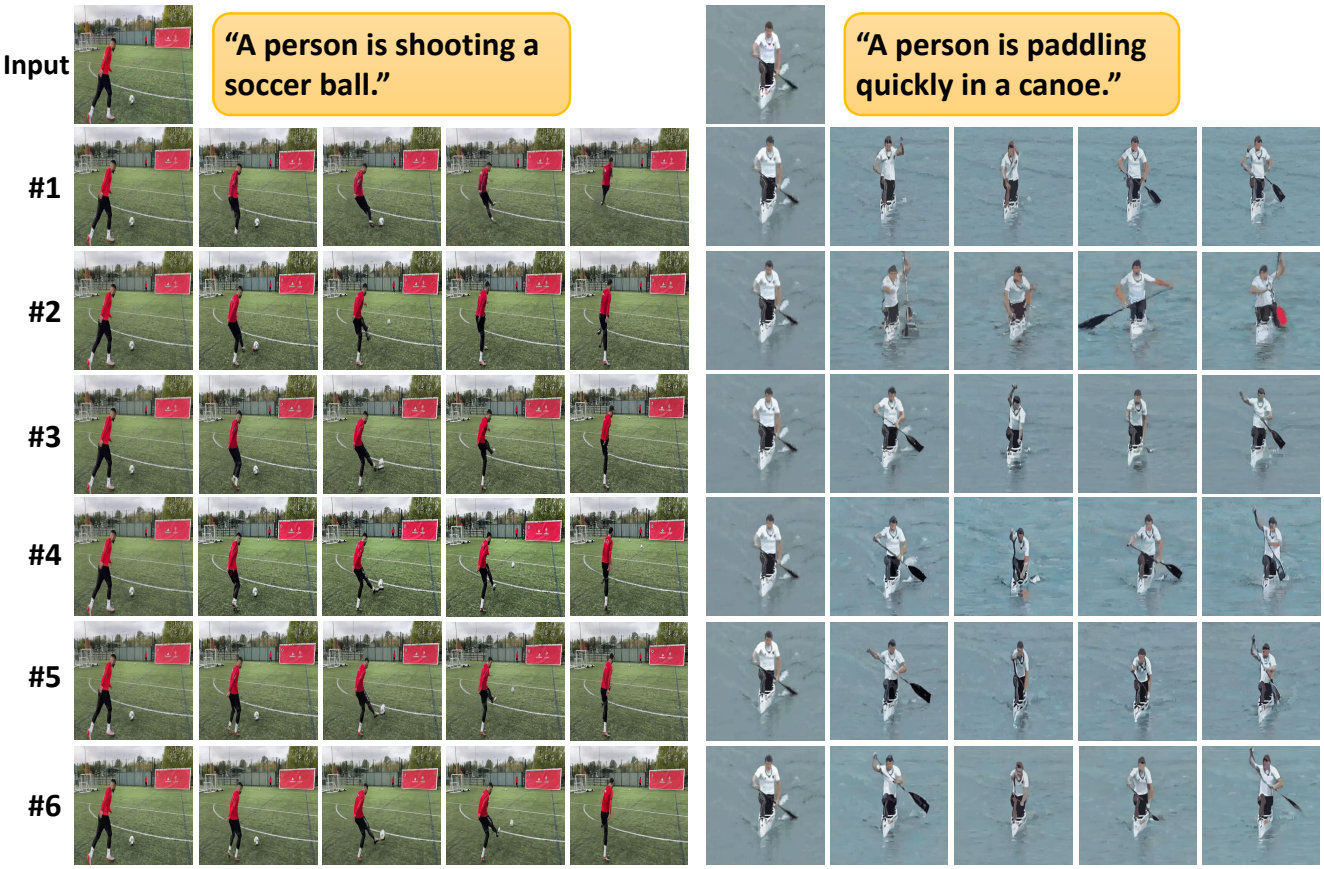


Figure A5. Qualitative ablation study on different components of FLASH. #1: baseline model, #2: model trained with strongly augmented videos, #3: with motion feature alignment, #4: with inter-frame correspondence alignment, #5: with both alignments, #6: full model with the Motion Alignment Module and the Detail Enhancement Decoder. See Table 2 for the details of each variant.

cosine similarity for each pair of the generated and ground-truth videos. (1) *Cosine RGB*: We extract video features using I3D [3], pretrained on RGB videos, for both the generated and ground truth videos, and calculate cosine similarity for the pair. (2) *Cosine Flow*: We extract optical flow using RAFT [85] and then use I3D [3], pretrained on optical

Table A1. The percentage of users that choose generated videos of each FLASH variant as the best in the user study on Amazon Mechanical Turk.

Variant	#1	#2	#3	#4	#5	#6
Preference Rate	10.99%	9.34%	15.38%	15.93%	20.33%	28.02%

flow data, to extract video features for cosine similarity calculation. In these two metrics, higher similarities indicate better performance.

A3.4. Baselines

We compare FLASH with several baselines: (1) TI2V-Zero [63], a training-free image animation model that injects the appearance of the reference image into a pretrained text-to-video model. We directly test image animation on its checkpoints. (2) SparseCtrl [25], an image animation model that encodes the reference image with a sparse condition encoder and integrates the features into a text-to-video model. It is trained on large-scale video datasets, and we directly test image animation on its checkpoints. (3) PIA [112], an image animation model that incorporates the reference image features into the noised latent video. It is trained on large-scale video datasets, and we directly test image animation on its checkpoints. (4) DynamiCrafter [100], an image animation model that injects the reference image features into generated videos via cross-attention layers and feature concatenation. It is trained on large-scale video datasets, and we directly test image animation on its checkpoints. (5) DreamVideo [96], a customized video generation model which learns target subject and motion using a limited set of samples. We train it to customize motion per action using the same training videos as FLASH. (6) MotionDirector [116], a customized video generation model which learns target appearance and motion with limited videos. We train its motion adapter per action with the same training videos as FLASH and train its appearance adapter per sample with the testing reference images. Therefore, MotionDirector has access to more data (*i.e.*, the testing reference images) than other methods. (7) LAMP [98], a few-shot image animation model which learns motion patterns from a few videos. We train it per action with the same training videos as FLASH.

A4. Results

A4.1. User Study

We conducted a user study on Amazon Mechanical Turk (AMT), where workers were asked to select the best generated video from a set of candidates. For each action, we randomly selected four different reference images and their corresponding generated videos for this user study. The AMT assessment interface, shown in Figure A4, presented workers with the following instructions: "You will see a reference image on the left and eight human action videos on the right, all generated from that reference image and an action description. Please carefully select the one video in each question that: (1) Best matches the action description and displays the action correctly and smoothly. (2) Maintains the overall appearance of the reference image on the left." The interface also displayed the reference image and action description.

To identify random clicking, each question was paired with a control question that has obvious correct answer. The control question includes a real video of a randomly selected action alongside clearly incorrect ones, such as a static video, a video with shuffled frames, and a video from the same action class that mismatches the reference image. The main question and the control question were randomly shuffled within each question pair, and each pair was evaluated by 10 different workers. Responses from workers who failed the control questions were marked as invalid.

Comparison with Baselines. Each question contains generated videos from the baselines and FLASH. In total, we collected 488 valid responses. The preference rates for different methods are shown in the pie chart in Figure 3 in the main paper. FLASH was preferred in 66% of the valid responses, significantly outperforming the next best choices, DynamiCrafter(13%) and LAMP (11%). We also provide 4 examples where humans prefer LAMP or DynamiCrafter on the webpage[6], which are mainly low-range motion.

Ablation Study. Each question contains generated videos from different variants of FLASH (see Table 2). In total, we collected 182 valid responses. The preference rates for these FLASH variants are presented in Table A1. We observe that variants #3, #4, and #5 achieve higher preference rates than variant #2, and variant #6 outperforms variant #5. These results validate the effectiveness of each designed component in FLASH.

A4.2. More Results

The definitions of the FLASH variants are consistent with those in the main paper (Table 2). Variant #1 serves as the baseline, excluding both the Motion Alignment Module and the Detail Enhancement Decoder. Variant #2 uses strongly augmented videos for training without any alignment technique. Variants #3, #4, and #5 progressively incorporate motion feature alignment, inter-frame correspondence alignment, and both, respectively, on top of variant #2. Lastly, variant #6 builds upon variant #5 by incorporating the Detail Enhancement Decoder.

Applicability with Fewer Training Videos. To assess the few-shot learning capability of the Motion Alignment Mod-

[6] https://lihaoxin05.github.io/human_action_animation/example_other_methods_win

Table A2. Analysis of training with fewer videos and joint training with multiple action classes.

Variant	# Videos Per Class	joint Training	Cosine RGB (↑)	Cosine Flow (↑)	CD-FVD (↓)	Text Alignment (↑)	Image Alignment (↑)	Temporal Consistency (↑)
#1	16	✘	83.80	68.06	1023.30	22.53	77.10	95.43
#2	16	✘	83.98	70.61	932.92	22.48	76.72	94.91
#5	16	✘	84.46	72.24	906.31	22.52	76.35	95.01
#1	8	✘	82.50	68.13	995.43	22.70	76.05	94.79
#2	8	✘	83.30	70.09	962.82	22.62	74.37	94.40
#5	8	✘	83.40	72.01	943.54	22.66	75.02	94.51
#1	4	✘	81.40	68.02	1050.03	22.22	72.81	94.24
#2	4	✘	81.88	70.15	1045.49	22.60	72.00	93.83
#5	4	✘	82.22	71.83	1031.87	22.46	72.56	94.22
#5	16	✘	84.46	72.24	906.31	22.52	76.35	95.01
#5	16	✔	85.01	72.32	897.05	22.61	77.47	95.39

Table A3. Evaluation of the Detail Enhancement Decoder with DINO-V2.

Variant	Image Alignment (↑)	Temporal Consistency (↑)
#5	87.02	97.48
#6	87.75	97.85

ule, we conduct experiments using 4 and 8 videos randomly sampled from each of the following action classes: sprint run, soccer shoot, canoeing sprint, and hugging human. The results are shown in Table A2. Across different numbers of training videos per action class, Variant #5 consistently outperforms Variants #1 and #2 on CD-FVD, Cosine-RGB and Cosine-Flow. The results show that the Motion Alignment Module enhances the motion quality of animated videos in different few-shot configurations.

Joint Training with Multiple Action Classes. We examine whether the model benefits from joint training across multiple action classes. We use all the training videos from four action classes (sprint run, soccer shoot, canoeing sprint, and hugging human) to train a single model with 100,000 steps. The results in Table A2 show improvements across all metrics. The improvements in Image Alignment, Temporal Consistency, and Cosine RGB are considerable. The results suggest that joint training with multiple action classes enhances the quality of the generated videos. This makes our technique more practical for applications that have accessible example videos of multiple delicate or customized human actions. In practice, the model can be trained per action or on multiple actions, depending on the application.

To further verify the applicability of our multi-action model across diverse actions, we train a new model on 8 new actions with 100,000 steps. These actions include atlatl throw, axe throwing, backward roll, balance beam flip, balance beam rotation, baseball swing, bench dip, and burpee. We conduct a qualitative comparison between the action videos generated by Wan[7] and FLASH, with the results accessible on the webpage[8]. The results show that FLASH generates actions that are more realistic and plausible than Wan, validating the broad applicability of multi-action FLASH across a variety of human actions.

Evaluation of the Detail Enhancement Decoder with DINO model. The CLIP vision encoder, trained on vision-language tasks, may have limited ability to perceive fine-grained visual details [86], which can affect the evaluation of Image Alignment and Temporal Consistency. Therefore, we use the DINO-V2 [64] vision encoder, which excels at capturing rich, fine-grained details at the pixel level, to assess Image Alignment and Temporal Consistency. The results in Table A3 demonstrate that the Detail Enhancement Decoder enhances both Image Alignment and Temporal Consistency, illustrating its effectiveness in improving transition smoothness.

Effects of Text Input Alterations on Video Generation. To assess how modifications to text input influence video generation outputs, we present results from experiments where subject-related terms in the text prompt were varied (see Figure A6). We observe that when the subject term in the input text misaligns with the subject of the reference image, the visual appearance of the generated subject exhibits distortion. For example, in Figure A6(b), the facial features of the panda are clearly distorted, highlighted by red boxes.

[7] https://tongyi.aliyun.com/wanxiang/
[8] https://lihaoxin05.github.io/human_action_animation/

Table A4. Ablation studies on different values of τ for motion feature alignment, different values of p for inter-frame correspondence alignment, and the impact of the Warping Branch and Patch Attention Branch in the Detail Enhancement Decoder.

Variant	τ	p	Warping Branch	Patch Attention Branch	Cosine RGB (↑)	Cosine Flow (↑)	CD-FVD (↓)	Text Alignment (↑)	Image Alignment (↑)	Temporal Consistency (↑)
#3	90	-	-	-	84.44	71.40	920.39	22.64	76.48	95.06
#3	75	-	-	-	84.38	71.19	904.25	22.58	76.63	95.16
#3	50	-	-	-	84.30	70.31	934.84	22.57	77.29	95.14
#3	25	-	-	-	84.71	69.79	930.53	22.33	76.52	94.85
#4	-	1.0	-	-	84.22	69.34	914.12	22.50	76.43	94.91
#4	-	0.5	-	-	84.32	71.72	938.21	22.70	76.31	94.84
#5	90	0.5	✘	✘	84.46	72.24	906.31	22.52	76.35	95.01
#6	90	0.5	✔	✘	84.63	71.96	918.61	22.54	76.21	95.35
#6	90	0.5	✘	✔	83.32	72.26	888.05	22.71	74.97	95.13
#6	90	0.5	✔	✔	84.51	72.33	908.39	22.77	76.22	95.31

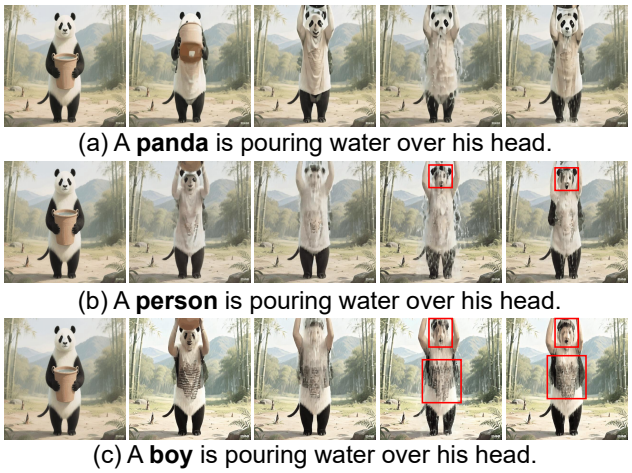


Figure A6. Effects of text input alterations on FLASH video generation. The reference image is kept constant, while the subject in the text prompt varies among "panda", "person" and "boy".

The results indicate the text encoder does encode the appearance of subject.

A4.3. Hyper-parameter Analysis

We tune hyperparameters on four action classes: sprint run, soccer shoot, canoeing sprint, and hugging human. The selected hyperparameters are then applied across all actions without action-specific tuning. The definitions of FLASH variants are consistent with those in the main paper (Table 2).

Motion Alignment Module. In Table A4, we compare the performance of different τ values in Variant #3 and different p values in Variant #4. For τ, we observe that decreasing τ reduces performance in Temporal Consistency, CD-FVD, and Cosine Flow, especially in Temporal Consistency (94.85 for $\tau = 25$) and Cosine Flow (69.79 for $\tau = 25$). This suggests that including more channels as motion channels degrades video quality, likely because motion information is only encoded in a limited number of channels [99], and aligning too many channels hampers feature learning. Thus, we set $\tau = 90$ for the remaining experiments. Regarding p, substituting inter-frame correspondence relations in all cross-frame attention layers ($p = 1.0$) lowers Cosine Flow significantly (*e.g.*, 69.34 for $p = 1.0$) but doesn't affect other metrics obviously. This might be due to the excessive regularization from substituting inter-frame correspondence relations in every layer, which makes learning difficult. Therefore, we use $p = 0.5$ in the remaining experiments.

Detail Enhancement Decoder. In Table A4, we compare the effects of the Warping Branch and the Patch Attention Branch in Variant #6. Using only the Warping Branch leads to a notable improvement in Temporal Consistency (from 95.01 to 95.35). In contrast, the Patch Attention Branch provides a modest increase in Text Alignment (from 22.52 to 22.71) but results in a significant drop in Image Alignment (from 76.35 to 74.97). When both branches are combined, there is an enhancement in both Text Alignment and Temporal Consistency, accompanied by only a slight decrease in Image Alignment. These results suggest that the two branches have complementary effects. Therefore, we use the two branches in the Detail Enhancement Decoder.

A4.4. Mean and Standard Deviations

To account for the variability introduced by different random seeds, we repeat experiments twice. We report the means and standard deviations of the results in Table A5.

Figure A7. Failure cases of FLASH.

Table A5. Means and standard deviations of FLASH performance.

Method	Cosine RGB (↑)	Cosine Flow (↑)	CD-FVD (↓)	Text Alignment (↑)	Image Alignment (↑)	Temporal Consistency (↑)
FLASH	85.48±0.03	77.42±0.02	815.11±20.46	23.21±0.01	79.66±0.31	95.81±0.22

Table A6. Quantitative comparison of different methods on the UCF Sports Action Dataset. The best and second-best results are **bolded** and underlined.

Method	Cosine RGB (↑)	Cosine Flow (↑)	CD-FVD (↓)	Text Alignment (↑)	Image Alignment (↑)	Temporal Consistency (↑)
TI2V-Zero	71.90	64.43	1222.35	24.62	70.16	88.87
SparseCtrl	71.56	63.21	1574.69	23.26	61.16	89.69
PIA	70.05	58.51	1385.54	23.93	66.03	94.41
DynamiCrafter	77.83	63.16	1630.83	23.95	**87.84**	96.75
DreamVideo	68.60	70.20	949.72	**26.04**	78.20	96.19
MotionDirector	75.60	63.01	1315.36	23.88	76.92	**97.07**
LAMP	74.15	73.78	1076.77	24.02	81.17	95.17
FLASH	**86.80**	**79.36**	**480.70**	24.11	85.75	96.22

Table A7. Quantitative comparison of different methods on non-human motion videos. The best and second-best results are **bolded** and underlined.

Method	Cosine RGB (↑)	Cosine Flow (↑)	CD-FVD (↓)	Text Alignment (↑)	Image Alignment (↑)	Temporal Consistency (↑)
TI2V-Zero	58.96	45.31	1562.76	21.95	79.05	93.00
SparseCtrl	67.89	59.02	1441.31	21.92	76.89	93.75
PIA	68.11	57.02	1591.11	21.83	79.44	96.83
DynamiCrafter	77.14	69.90	1371.39	22.18	**87.27**	**98.08**
DreamVideo	68.80	61.44	1222.22	**22.75**	84.40	96.96
MotionDirector	74.57	68.41	1302.02	20.75	79.09	96.68
LAMP	79.48	71.89	1210.55	22.14	86.68	97.49
FLASH	**79.53**	**75.48**	**1204.72**	22.05	85.42	97.51

A4.5. Experiments on UCF Sports Action Dataset

To evaluate the effectiveness of FLASH on additional datasets, we conducted experiments on the UCF Sports Action Dataset [84], focusing on two actions: golf swing and lifting. Due to the limited number of videos in this dataset, we use only 6 golf swing videos and 4 lifting videos for training. For each class, we use the first frames of two videos for testing.

Table A6 compares the performance of FLASH with baseline methods. FLASH achieves superior results on CD-FVD, Cosine RGB, and Cosine Flow, highlighting its ability to generate realistic motions. While DynamiCrafter performs better on Image Alignment and Temporal Consistency, this is primarily because it fails to animate the reference images and instead repeats it across frames, which represents a failure in animation. This limitation is further reflected in the poor scores of DynamiCrafter on CD-FVD and Cosine Flow. For Text Alignment, DreamVideo and TI2V-Zero outperform FLASH, but their inability to generate smooth transitions from reference images is evident from their low Image Alignment scores. These observations, consistent with results on the HAA dataset, demonstrate the effectiveness of FLASH in scenarios with fewer training videos.

A4.6. Experiments on Non-human Motion Videos

To assess the performance of FLASH on non-human motion videos, we conducted experiments using two categories of natural motion: firework and raining. The videos were sourced from [98]. For each category, we selected two videos for testing and used the remaining videos for training.

Table A7 presents a comparison between FLASH and baseline methods. FLASH achieves superior performance in CD-FVD, Cosine RGB, and Cosine Flow while not showing a obvious decline in CLIP scores. Although DynamiCrafter performs better in Image Alignment and Temporal Consistency, it struggles with CD-FVD and Cosine Flow. Similarly, DreamVideo excels in Text Alignment but performs poorly in Cosine RGB and Cosine Flow. These results indicate that FLASH can also animate images into videos depicting natural scene motion.

A5. Limitations

Although FLASH can animate diverse reference images, it encounters challenges in accurately generating interactions involving human and objects, particularly when multiple objects are present. For example, in Figure A7(a), while a chopping action is depicted, the object being chopped is not the wood. Furthermore, if the initial action states in the reference images differ noticeably in motion patterns from those in the training videos, the model may struggle with animation. For example, in Figure A7(b), the initial action status suggests a small-scale motion for chopping wood, which differs from the large-scale motion in training videos; in Figure A7(c), the knee elevation motion contrasts with the steadier motion of running in place observed in the training videos; and in Figure A7(d), a baby holding a cup with both hands deviates from the adult actions in the training videos, where one hand is used to hold the cup while drinking water. These results suggest that the model still lacks a thorough understanding of motion and interactions. Leveraging advanced multi-modal large language models to improve the understanding of human-object interactions could be a promising approach to addressing these challenges.

A6. Ethics Statement

We firmly oppose the misuse of generative AI for creating harmful content or spreading false information. We do not assume any responsibility for potential misuse by users. Nonetheless, we recognize that our approach, which focuses on animation human images, carries the risk of potential misuse. To address these risks, we are committed to maintaining the highest ethical standards in our research by complying with legal requirements and protecting privacy. Additionally, we will explore implementing a content safety mechanism, similar to the one used in Stable Diffusion [72], as an effective way to address these concerns.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 2

[2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 5

[4] Di Chang et al. Magicpose. *arXiv:2311.12052*, 2023. 2

[5] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. *arXiv preprint arXiv:2401.09047*, 2024. 2

[6] J Chen, A. B. Solinger, J. F. Poncet, and C. A. Lantz. Meta-analysis of normative cervical motion. *Spine*, 24:1571–1578, 1999. 1

[7] Ricky TQ Chen, Jens Behrmann, David K Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[8] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024. 2

[9] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36:30286–30305, 2023. 3

[10] Jihoon Chung, Cheng-hsin Wuu, Hsuan-ru Yang, Yu-Wing Tai, and Chi-Keung Tang. Haa500: Human-centric atomic action dataset with curated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13465–13474, 2021. 5, 3

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3

[12] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 1, 2, 3

[13] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 7

[14] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Empowering dynamics-aware text-to-video diffusion with large language models. *arXiv preprint arXiv:2308.13812*, 2023. 2

[15] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[16] Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. On the content bias in fréchet video distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7277–7288, 2024. 6, 4

[17] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 3

[18] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023. 3

[19] Litong Gong, Yiran Zhu, Weijie Li, Xiaoyang Kang, Biao Wang, Tiezheng Ge, and Bo Zheng. Atomovideo: High fidelity image-to-video generation. *arXiv preprint arXiv:2403.01800*, 2024. 2, 3

[20] 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

[21] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 4

[22] Xianfan Gu, Chuan Wen, Weirui Ye, Jiaming Song, and Yang Gao. Seer: Language instructed video prediction with latent diffusion models. *arXiv preprint arXiv:2303.14897*, 2023. 2

[23] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36:15890–15902, 2023. 3

[24] Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Pengfei Wan, Di Zhang, Yufan Liu, Weiming Hu, Zhengjun Zha, et al. I2v-adapter: A general image-to-video adapter for diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 2, 3

[25] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. *arXiv preprint arXiv:2311.16933*, 2023. 2, 3, 6, 1

[26] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 1, 2, 3, 4

[27] Zekun Hao, Xun Huang, and Serge Belongie. Controllable video generation with sparse trajectories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[28] Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940*, 2023. 1, 2, 4

[29] Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*, 2024. 6, 4

[30] Katherine Hermann et al. The origins and prevalence of texture bias in convolutional neural networks. *NeurIPS*, pages 19000–19015, 2020. 4

[31] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[32] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[33] 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. 2

[34] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 1, 2, 3

[35] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 2

[36] Hanzhuo Huang, Yufan Feng, Cheng Shi, Lan Xu, Jingyi Yu, and Sibei Yang. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[37] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 4

[38] Xiaohu Huang, Hao Zhou, Kun Yao, and Kai Han. Froster: Frozen clip is a strong teacher for open-vocabulary action recognition. *arXiv preprint arXiv:2402.03241*, 2024. 1

[39] Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin CK Chan, and Ziwei Liu. Reversion: Diffusion-based relation inversion from images. *arXiv preprint arXiv:2303.13495*, 2023. 4

[40] Hyeonho Jeong, Geon Yeong Park, and Jong Chul Ye. Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9212–9221, 2024. 3

[41] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. *arXiv preprint arXiv:2312.00777*, 2023. 2, 3

[42] Hitesh Kandala, Jianfeng Gao, and Jianwei Yang. Pix2gif: Motion-guided diffusion for gif generation. *arXiv preprint arXiv:2403.04634*, 2024. 2, 3

[43] Manuel Kansy, Jacek Naruniec, Christopher Schroers, Markus Gross, and Romann M Weber. Reenact anything: Semantic video motion transfer using motion-textual inversion. *arXiv preprint arXiv:2408.00458*, 2024. 3

[44] 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. 3, 1

[45] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[46] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 1, 2

[47] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023. 3

[48] Xiaomin Li, Xu Jia, Qinghe Wang, Haiwen Diao, Pengxiang Li, You He, Huchuan Lu, et al. Motrans: Customized motion transfer with text-driven video diffusion models. In *ACM Multimedia 2024*, 2024. 2, 3

[49] Yumeng Li, William Beluch, Margret Keuper, Dan Zhang, and Anna Khoreva. Vstar: Generative temporal nursing for longer dynamic video synthesis. *arXiv preprint arXiv:2403.13501*, 2024. 2

[50] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. Llm-grounded video diffusion models. *arXiv preprint arXiv:2309.17444*, 2023. 2

[51] Jun Hao Liew, Hanshu Yan, Jianfeng Zhang, Zhongcong Xu, and Jiashi Feng. Magicedit: High-fidelity and temporally coherent video editing. *arXiv preprint arXiv:2308.14749*, 2023. 1, 2

[52] Kun-Yu Lin, Henghui Ding, Jiaming Zhou, Yu-Ming Tang, Yi-Xing Peng, Zhilin Zhao, Chen Change Loy, and Wei-Shi Zheng. Rethinking clip-based video learners in cross-domain open-vocabulary action recognition. *arXiv preprint arXiv:2403.01560*, 2024. 1

[53] Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaian Chen, Jiaqi Wang, and Yi Jin. Motionclone: Training-free motion cloning for controllable video generation. *arXiv preprint arXiv:2406.05338*, 2024. 3

[54] Yu Lu, Linchao Zhu, Hehe Fan, and Yi Yang. Flowzero: Zero-shot text-to-video synthesis with llm-driven dynamic scene syntax. *arXiv preprint arXiv:2311.15813*, 2023. 2

[55] Jiaxi Lv, Yi Huang, Mingfu Yan, Jiancheng Huang, Jianzhuang Liu, Yifan Liu, Yafei Wen, Xiaoxin Chen, and Shifeng Chen. Gpt4motion: Scripting physical motions in text-to-video generation via blender-oriented gpt planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1430–1440, 2024. 2

[56] Wan-Duo Kurt Ma, John P Lewis, and W Bastiaan Kleijn. Trailblazer: Trajectory control for diffusion-based video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 1, 2

[57] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Tien-Tsin Wong, Yuan-Fang Li, and Cunjian Chen. Consistent and controllable image animation with motion diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7288–7298, 2025. 2, 3

[58] Yue Ma, Yingqing He, Hongfa Wang, Andong Wang, Chenyang Qi, Chengfei Cai, Xiu Li, Zhifeng Li, Heung-Yeung Shum, Wei Liu, et al. Follow-your-click: Open-domain regional image animation via short prompts. *arXiv preprint arXiv:2403.08268*, 2024. 2, 3

[59] Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Implicit warping for animation with image sets. *Advances in Neural Information Processing Systems*, 35:22438–22450, 2022. 5

[60] Joanna Materzynska, Josef Sivic, Eli Shechtman, Antonio Torralba, Richard Zhang, and Bryan Russell. Customizing motion in text-to-video diffusion models. *arXiv preprint arXiv:2312.04966*, 2023. 2, 3, 4

[61] 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. 3

[62] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18444–18455, 2023. 2, 3

[63] Haomiao Ni, Bernhard Egger, Suhas Lohit, Anoop Cherian, Ye Wang, Toshiaki Koike-Akino, Sharon X Huang, and Tim K Marks. Ti2v-zero: Zero-shot image conditioning for text-to-video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9015–9025, 2024. 6

[64] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 7

[65] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 1

[66] Geon Yeong Park, Hyeonho Jeong, Sang Wan Lee, and Jong Chul Ye. Spectral motion alignment for video motion transfer using diffusion models. *arXiv preprint arXiv:2403.15249*, 2024. 3

[67] Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Temporal-relational crosstransformers for few-shot action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 475–484, 2021. 1

[68] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15932–15942, 2023. 3

[69] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6, 4

[70] Weiming Ren, Harry Yang, Ge Zhang, Cong Wei, Xinrun Du, Stephen Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024. 2, 3, 1

[71] Yixuan Ren, Yang Zhou, Jimei Yang, Jing Shi, Difan Liu, Feng Liu, Mingi Kwon, and Abhinav Shrivastava. Customize-a-video: One-shot motion customization of text-to-video diffusion models. *arXiv preprint arXiv:2402.14780*, 2024. 3

[72] 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. 3, 10

[73] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 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*, pages 22500–22510, 2023. 3

[74] Dian Shao, Mingfei Shi, Shengda Xu, Haodong Chen, Yongle Huang, and Binglu Wang. Finephys: Fine-grained human action generation by explicitly incorporating physical laws for effective skeletal guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1905–1916, 2025. 2

[75] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8543–8552, 2024. 3

[76] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. *arXiv preprint arXiv:2401.15977*, 2024. 2, 3

[77] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. *arXiv preprint arXiv:2306.14435*, 2023. 1

[78] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019. 3

[79] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13653–13662, 2021. 3

[80] 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

[81] James Seale Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia Jin. Continual diffusion: Continual customization of text-to-image diffusion with c-lora. *arXiv preprint arXiv:2304.06027*, 2023. 3

[82] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 4

[83] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[84] Khurram Soomro and Amir R Zamir. Action recognition in realistic sports videos. In *Computer vision in sports*, pages 181–208. Springer, 2015. 9

[85] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 5

[86] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024. 6, 7

[87] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 4

[88] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6, 4

[89] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 2

[90] Cong Wang, Jiaxi Gu, Panwen Hu, Songcen Xu, Hang Xu, and Xiaodan Liang. Dreamvideo: High-fidelity image-to-video generation with image retention and text guidance. *arXiv preprint arXiv:2312.03018*, 2023. 2, 3

[91] Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. *arXiv preprint arXiv:2402.01566*, 2024. 1, 2

[92] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2023. 2

[93] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[94] Yanhui Wang, Jianmin Bao, Wenming Weng, Ruoyu Feng, Dacheng Yin, Tao Yang, Jingxu Zhang, Qi Dai, Zhiyuan Zhao, Chunyu Wang, et al. Microcinema: A divide-and-conquer approach for text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8414–8424, 2024. 3

[95] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2

[96] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6537–6549, 2024. 2, 3, 6

[97] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 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*, pages 7623–7633, 2023. 1, 3, 6, 4

[98] Ruiqi Wu, Liangyu Chen, Tong Yang, Chunle Guo, Chongyi Li, and Xiangyu Zhang. Lamp: Learn a motion pattern for few-shot-based video generation. *arXiv preprint arXiv:2310.10769*, 2023. 3, 6, 1, 4, 9

[99] Zeqi Xiao, Yifan Zhou, Shuai Yang, and Xingang Pan. Video diffusion models are training-free motion interpreter and controller. *arXiv preprint arXiv:2405.14864*, 2024. 3, 4, 8

[100] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023. 2, 3, 6, 4

[101] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5036–5045, 2022. 2

[102] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 3

[103] Xiangpeng Yang, Linchao Zhu, Hehe Fan, and Yi Yang. Eva: Zero-shot accurate attributes and multi-object video editing. *arXiv preprint arXiv:2403.16111*, 2024. 3

[104] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1, 2

[105] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 1, 2

[106] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. *arXiv preprint arXiv:2311.10982*, 2023. 3

[107] Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip HS Torr, and Piotr Koniusz. Few-shot action recognition with permutation-invariant attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 525–542. Springer, 2020. 1

[108] Jianfeng Zhang, Hanshu Yan, Zhongcong Xu, Jiashi Feng, and Jun Hao Liew. Magicavatar: Multimodal avatar generation and animation. *arXiv preprint arXiv:2308.14748*, 2023. 1, 2

[109] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1

[110] Xing Zhang, Zuxuan Wu, Zejia Weng, Huazhu Fu, Jingjing Chen, Yu-Gang Jiang, and Larry S Davis. Videolt: Large-scale long-tailed video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7960–7969, 2021. 1

[111] Yuxin Zhang, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Motioncrafter: One-shot motion customization of diffusion models. *arXiv preprint arXiv:2312.05288*, 2023. 3

[112] Yiming Zhang, Zhening Xing, Yanhong Zeng, Youqing Fang, and Kai Chen. Pia: Your personalized image animator via plug-and-play modules in text-to-image models. *arXiv preprint arXiv:2312.13964*, 2023. 2, 6

[113] Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. *arXiv preprint arXiv:2407.21705*, 2024. 2

[114] Zhongwei Zhang, Fuchen Long, Zhaofan Qiu, Yingwei Pan, Wu Liu, Ting Yao, and Tao Mei. Motionpro: A precise motion controller for image-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27957–27967, 2025. 2

[115] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. *arXiv preprint arXiv:2310.08465*, 2023. 2, 3

[116] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pages 273–290. Springer, 2024. 6

[117] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 1, 2