Reenact Anything: Semantic Video Motion Transfer Using Motion-Textual Inversion

Supplementary Material

MANUEL KANSY, ETH Zürich, Switzerland and DisneyResearch|Studios, Switzerland
JACEK NARUNIEC, DisneyResearch|Studios, Switzerland
CHRISTOPHER SCHROERS, DisneyResearch|Studios, Switzerland
MARKUS GROSS, ETH Zürich, Switzerland and DisneyResearch|Studios, Switzerland
ROMANN M. WEBER, DisneyResearch|Studios, Switzerland

A Broader Impact and Ethics

To the best of our knowledge, our method is the first that can reenact a wide array of objects and motions given a target image and motion reference video without training domain-specific models. We believe this represents a significant advancement in controllable video generation, as our approach can address multiple existing domain-specific scenarios within a single framework and even facilitate entirely new applications. That said, we acknowledge the potential for misuse of reenactment methods like ours, such as creating realistic deepfakes or videos depicting individuals or objects performing specified, potentially inappropriate actions. We strongly condemn such misuse and advocate for implementing safety mechanisms and procedures in real-world applications. Additionally, we support ongoing research into detecting fake videos to mitigate these risks.

For legal reasons, we cannot show images or videos from public data sets in the paper without individuals' written consents. For the qualitative evaluation, we therefore use motion reference videos and target images from internal data sets as well as target images generated with Stable Diffusion XL [Podell et al. 2024].

B Extended Related Work

In this section, we provide an extended description of related work for interested readers.

Authors' Contact Information: Manuel Kansy, ETH Zürich, Zürich, Switzerland and DisneyResearch|Studios, Zürich, Switzerland, manuel.kansy@disneyresearch.com; Jacek Naruniec, DisneyResearch|Studios, Zürich, Switzerland, jacek.naruniec@disneyresearch.com; Christopher Schroers, DisneyResearch|Studios, Zürich, Switzerland, christopher.schroers@disneyresearch.com; Markus Gross, ETH Zürich, Zürich, Switzerland and DisneyResearch|Studios, Zürich, Switzerland, grossm@inf.ethz.ch; Romann M. Weber, DisneyResearch|Studios, Zürich, Switzerland, romann.weber@disneyresearch.com.

SIGGRAPH Conference Papers '25, Vancouver, BC, Canada

ACM ISBN 979-8-4007-1540-2/2025/08
https://doi.org/10.1145/3721238.3730668

B.1 Domain-Specific Reenactment

Reenactment has been a significant research area, but much of the focus has been on domain-specific approaches like face reenactment [Drobyshev et al. 2022; Guo et al. 2024b; Hsu et al. 2022; Li et al. 2023; Nirkin et al. 2019; Wang et al. 2021] and human full-body motion transfer [Chan et al. 2019; Hu 2024; Karras et al. 2023; Ma et al. 2024a; Tu et al. 2024a,b; Wang et al. 2024a; Yang et al. 2020; Zhu et al. 2024; Zuo et al. 2024]. While these methods perform well, their architectures and training data are tailored to specific domains, making it challenging to adapt them for use across multiple domains.

B.2 Keypoint-Based Motion Transfer

Keypoint-based motion transfer has been a popular approach in reenactment, spanning both domain-specific and more general methods. Many techniques extract keypoints using pre-trained, domain-specific landmark detectors [Chan et al. 2019; Hsu et al. 2022; Hu 2024; Ma et al. 2024a; Ni et al. 2023; Nirkin et al. 2019; Tu et al. 2024a; Yang et al. 2020; Zuo et al. 2024], which limits their applicability to specific object categories like human bodies or faces. To move toward general motion transfer, other approaches learn keypoints in an unsupervised manner [Drobyshev et al. 2022; Guo et al. 2024b; Siarohin et al. 2019, 2021; Tanveer et al. 2024; Wang et al. 2021; Zhao and Zhang 2022]. Although this strategy increases flexibility, it still typically requires a separate model per domain, making it impractical for applications involving diverse object types.

Several methods first find meaningful common keypoints and then warp features [Ni et al. 2023; Siarohin et al. 2019, 2021; Zhao and Zhang 2022] or latents [Tanveer et al. 2024] to transfer motion from the driving to the target object. However, such warping becomes nontrivial in the presence of 3D rotations, and methods like AnaMoDiff [Tanveer et al. 2024] are thus limited to flat 2D motions. JOKR [Mokady et al. 2022], while not relying on explicit warping, also focuses on relatively planar 2D motions and requires an affine alignment between the target and the driving video. Crucially, both JOKR and AnaMoDiff require a target video to learn target object motions, whereas our method works well even with a single target image by leveraging motion priors from a pre-trained image-to-video model.

Keypoint-based approaches also face challenges when applied to unseen domains or extreme cross-domain transfers (e.g., from animal to inanimate object). While recent advances in deep features from diffusion models [Hedlin et al. 2023; Luo et al. 2023; Tang et al. 2023; Zhang et al. 2024a, 2023a] have made it easier to find correspondences between points across different images, a more fundamental problem remains: where to place keypoints in the first place to meaningfully capture motion. This becomes especially difficult for hand-crafted motions or for motion transfers with large structural differences between objects (e.g., Fig. 7), where there may be no obvious semantically meaningful anchors. To address these challenges, we propose using an implicit motion representation instead of relying on explicit keypoints. We show that priors from pre-trained diffusion models can be used more directly, rather than only as a tool to find keypoint correspondences.

B.3 Video Generation

Following the rise of text-to-image diffusion models [Ramesh et al. 2022; Rombach et al. 2022; Saharia et al. 2022], video generation models have also greatly improved in quality in recent years. Many text-to-video methods start with a pre-trained text-to-image model and inflate it by adding and training temporal convolution and attention blocks after each corresponding spatial block [Bar-Tal et al. 2024; Blattmann et al. 2023b; Guo et al. 2024a; Wang et al. 2023b]. Similarly, many image-to-video diffusion models use a pre-trained text-to-image [Zhang et al. 2023c] or text-to-video [Blattmann et al. 2023a] model as a starting point. They then adapt the model to the image-to-video task by conditioning the model on the image, e.g., by adding [Zhang et al. 2023c] or concatenating [Blattmann et al. 2023a] it to the noisy input. The text embedding input from the pre-trained model is either kept [Zhang et al. 2023c] or replaced with an image embedding input [Blattmann et al. 2023a]. Recently, video generation models [Brooks et al. 2024; Kong et al. 2024; Yang et al. 2025] based on diffusion transformers [Peebles and Xie 2023] have gained significant popularity. While training a custom video generation model provides the most freedom in terms of design choices, it is very expensive in terms of computation and data. Even fine-tuning video models requires substantial resources, so we decided to use a pre-trained diffusion model, Stable Video Diffusion [Blattmann et al. 2023a], and keep it frozen. Additionally, we aim for our method to be applicable to a wide range of motions and subjects. In contrast, approaches that involve training the model often focus on a single type of motion, such as human full-body motion [Hu 2024; Ma et al. 2024a].

B.4 Video Motion Editing with Explicit Motions

B.4.1 Based on Sparse Control Signals. In theory, the motion of all video generation models that have a text input can simply be controlled by text [Dai et al. 2023; Li et al. 2024b; Molad et al. 2023; Yan et al. 2023], but this approach struggles with complex motions in practice. For more precise spatial control, recent methods use bounding boxes, either with training [Li et al. 2024b; Wang et al. 2024e] or without [Chen et al. 2024; Jain et al. 2024; Ma et al. 2024b], and trajectories [Chen et al. 2023a; Geng et al. 2024; Li et al. 2024c, 2025; Mou et al. 2024; Niu et al. 2024; Qiu et al. 2024; Wu et al. 2024b; Yin et al. 2023; Zhou et al. 2024], but they rely on consistent spatial alignment for effective motion transfer. Similarly, keypoints are another option for describing motions [Gu et al. 2024; Niu et al. 2024; Tanveer et al. 2024], but they suffer from the challenges outlined in Section B.1. Additionally, some methods focus specifically on camera motions [Bahmani et al. 2024; Cheong et al. 2024; He et al. 2024; Hou et al. 2024; Hu et al. 2024; Xu et al. 2024; Zheng et al. 2024] or combine camera and bounding box motions [Wang et al. 2024c; Wu et al. 2024c; Yang et al. 2024]. However, all these approaches are either limited to simple motions or require significant effort to specify complex ones. For instance, a bounding box can specify an object's location (e.g., a person) but not the detailed motion within it (e.g., doing jumping jacks). Modeling complex motion with part-based boxes or trajectories [Li et al. 2024c] quickly becomes impractical, especially if a precise temporal alignment to a reference motion is desired.

B.4.2 Based on Dense Control Signals. Dense control signals, such as motion vectors [Wang et al. 2024d], 3D tracking videos [Gu et al. 2025], warped noise [Burgert et al. 2025], and depth maps [Chen et al. 2023b; Wang et al. 2024d; Zhang et al. 2024b] allow for a more precise motion specification. However, using them for general motion transfer is challenging because they also encode information about image and object structure. This can result in unnatural motions when there is a mismatch between the structures of the target image and the reference video as shown in MotionCtrl [Wang et al. 2024c].

B.5 Video Motion Editing with Implicit Motions

This subsection covers methods for implicitly representing and transferring motion from a reference video. We thereby focus on the two main paradigms: fine-tuning approaches, which encode motion into model weights, and inversion-then-generation methods, which capture motion in model features and attention maps. Additionally, some techniques integrate elements of both paradigms.

When the layout of the subjects in the reference and generated videos match, a given transfer can be seen as either changing the appearance to match the target image or altering the motion to match the reference video. Our focus is on motion transfer where the layouts do not align, a less explored area in the literature, as discussed in Section B.5.3.

B.5.1 Fine-Tuning. Many fine-tuning methods are inspired by image customization techniques like DreamBooth [Ruiz et al. 2023] and LoRA [Hu et al. 2022]. Loosely speaking, the idea is to fine-tune the parts of the model responsible for motion but avoid training the parts responsible for appearance. Tune-A-Video [Wu et al. 2023] inflates a text-to-image model by adding spatio-temporal attention and only trains some parts of the attention layers. Similarly, Materzyńska et al. [2024] only fine-tune parts of the model and further focus the training more on earlier denoising steps to emphasize learning the general motion rather than fine appearance details. MotionDirector [Zhao et al. 2024] proposes a dual-path LoRA architecture and an appearance-debiased temporal loss to disentangle appearance from motion. Similarly, DreamVideo [Wei et al. 2024], MotionCrafter [Zhang et al. 2023b], Customize-A-Video [Ren et al. 2024], and CustomTTT [Bi et al. 2025] have separate branches for appearance and motion. CustomTTT [Bi et al. 2025] further proposes a test-time training method to improve the results when combing the appearance and motion information. VMC [Jeong et al. 2024] adapts temporal attention layers using a motion distillation strategy with residual vectors between consecutive noisy latent frames as the motion reference.

Fine-tuning a model carries the risk of appearance leakage, which is why many of the aforementioned methods focus on preventing it. We find that using an image-to-video model instead of a text-to-video model largely avoids these problems. LAMP [Wu et al. 2024a] is the most similar method to ours in that sense, but they adapt a pre-trained text-to-image model to the image-to-video task and fine-tune it only briefly. In contrast, we employ a pre-trained, large-scale image-to-video model to leverage stronger priors for better generalization.

B.5.2 Inversion-then-Generation. The inversion-then-generation framework, initially developed for image editing [Hertz et al. 2023; Parmar et al. 2023; Tumanyan et al. 2023], involves first inverting a reference video into "noise" using methods like DDIM [Song et al. 2020] to enable reconstruction through backward diffusion. Thereby, features such as self-attention maps are extracted from the reference video and then injected into the diffusion process of the video being generated. These features either directly replace existing features [Tumanyan et al. 2023] or are incorporated into a loss function [Parmar et al. 2023], ensuring the generated video has a similar structure. Numerous methods have been proposed within this framework for video appearance editing [Bai et al. 2024; Ceylan et al. 2023; Geyer et al. 2024; Harsha et al. 2024; Liu et al. 2024; Meral et al. 2024; Wang et al. 2023a; Yang et al. 2023; Zhao et al. 2023] and video motion editing [Bai et al. 2024; Yatim et al. 2023], mainly differing in their inversion techniques and feature choices.

The methods mentioned above face several inherent issues in motion transfer tasks. Most notably, they often assume or enforce that the features of the reference and target videos are identical, which leads to problems when generating videos with different geometries or spatial layouts. Some methods attempt to address this by collapsing the spatial dimension of features before using them in a loss [Yatim et al. 2023], but they still typically produce motions with similar directions in pixel space. This limits control and diversity and can produce less natural results. Furthermore, these approaches require tuning numerous hyperparameters (choice of feature, layers, time steps) and necessitate inverting the video, which is challenging for high guidance scales [Mokady et al. 2023] and when using few time steps [Garibi et al. 2024].

Another recent line of work [Ling et al. 2024; Pondaven et al. 2024; Xiao et al. 2024] extracts features from a reference video in line with the inversion-then-generation framework but without inversion. While these approaches bypass the costly inversion process, they still suffer from issues related to primarily replicating the spatial rather than semantic motion.

B.5.3 With Different Spatial Layout. To avoid being restricted to the layout of a single motion reference video, some methods use multiple motion videos [Materzyńska et al. 2024; Wei et al. 2024; Wu et al. 2024a; Zhao et al. 2024]. However, our goal is to transfer motion with precise temporal alignment to the reference video. This would require multiple temporally-aligned videos, which are often impractical to obtain. Additionally, many motion editing methods with spatial variations [Li et al. 2024a; Materzyńska et al. 2024; Ren et al. 2024; Wang et al. 2024b] use text to define the subject's appearance instead of an image, resulting in videos that only roughly match the input image. The concurrent work by Wang et al. [2024b] is most similar to ours as it keeps the model frozen and learns a motion embedding like we do, but it also suffers from the above limitation.

C Implementation Details

C.1 High-Level Overview of the Implementation

To aid in reproducibility, we list the main steps of our method's implementation below:

(1) [Only once] Take pre-trained Stable Video Diffusion (SVD) [Blattmann et al. 2023a] and adapt code to inflate motion-text embedding and cross-attention. See high-level description in the method section of the main paper and details in Section C.4.
(2) Initialize motion-text embedding of shape $(F+1) \times N \times d$. See Section C.2.
(3) Repeat until convergence:
   - Load same F frames of reference video in data loader for each iteration.
   - Augment data. See Section C.2.
   - Input noisy version of frames, motion-text embedding, and other inputs into SVD.
   - Apply loss to update motion-text embedding.
(4) Save motion-text embedding.
(5) For all target images:
   - Input learned motion-text embedding along with new target image to inflated SVD during inference to generate video with motion from reference video.

C.2 Hyperparameters

Our implementation builds up on the diffusers implementation [von Platen et al. 2022] of Stable Video Diffusion (SVD) [Blattmann et al. 2023a]. We use the default parameters of the 14-frame version of SVD (e.g., micro-conditionings) unless specified otherwise. Like SVD, we generally employ a classifier-free guidance [Ho and Salimans 2021] scale that increases linearly from 1 to 3 across the frame axis. For the motion visualization (unconditional image input), however, we use a higher scale, i.e., increasing linearly from 1 to 10, to improve the visibility of the objects. We initialize the $F = 14$ sets of $N = 5$ tokens for the spatial cross-attention with the CLIP image embedding token of each corresponding frame and the $N = 5$ tokens for the temporal cross-attention with the mean of the CLIP image embedding tokens across all frames. We additionally add Gaussian noise $\mathcal{N}(0, 0.1)$ to the combined motion-text embedding during initialization. In our experience, the initialization does not affect the results significantly, so other initializations are equally reasonable. During optimization, we always pick the same F frames of a given video and apply the same spatial and color augmentations to all frames.[1] Since most of the video motion is determined in noisy diffusion steps, we shift the noise schedule towards higher noise values (from $P_{\text{mean}} = 1.0, P_{\text{std}} = 1.6$ to $P_{\text{mean}} = 2.8, P_{\text{std}} = 1.6$ where $\log \sigma \sim \mathcal{N}(P_{\text{mean}}, P_{\text{std}}^2)$ to speed up the optimization. We use

[1]For horizontal camera motions, we turn of horizontal flipping

Adam [Kingma and Ba 2015] with a learning rate of 10^{-2} for 1000 iterations with a batch size of 1.

C.3 Hardware Requirements and Runtime

The optimization for a motion reference video with a resolution of 1024×576 takes around 55 GB of GPU memory and around one hour on an NVIDIA Tesla A100 (80 GB) GPU. The inference takes less than one minute per video. While the peak memory usage was measured at 55 GB on the A100, we have also successfully run the method on a 48 GB RTX A6000 GPU. Our current implementation has not been optimized extensively for memory efficiency or runtime, and further engineering could reduce the resource requirements.

C.4 Motion-Text Embedding and Cross-Attention Inflation

This section provides more implementation details for the motion-text embedding and cross-attention inflation described in the main paper. Fig. 1 shows the spatial and temporal cross-attention layers of the default Stable Video Diffusion (SVD) [Blattmann et al. 2023a] and our inflated version along with their tensor dimensions.

The image embedding of the default SVD consists of a single token and has dimensions $B \times 1 \times d$, where B is the batch size (in our implementation typically 1 when optimizing the motion-text embedding and 2 during inference because of classifier-free guidance) and d is the CLIP [Radford et al. 2021] embedding dimension. For spatial cross-attention, the image embedding is broadcast to dimensions $(B * F) \times 1 \times d$, i.e., the same token is used for all F frames. This results in an attention map M of dimensions $(B * F) \times (H_i * W_i) \times 1$ where H_i and W_i are the spatial heights and widths respectively, and C_i is the number of channels of level i of the diffusion model. Notably, due to the softmax operation and the last dimension being 1, every value of the attention map is 1. This means that each spatial location attends 100% to the single token. Similarly, for temporal cross-attention, the image embedding is broadcast from dimensions of $B \times 1 \times d$ to dimensions $(B * H_i * W_i) \times 1 \times d$, eventually leading to an attention map M of dimensions $(B * H_i * W_i) \times F \times 1$ where every value is 1. Having only one token thus leads to a degenerate case of the cross-attention where $\text{Attention}(Q, K, V) = V$ (broadcasted) and many of the components (e.g., queries and keys) have no effect on the result.

C.4.1 Multiple Tokens. To avoid the above degenerate case and instead be able to dynamically attend to different tokens, we extend the token dimension from 1 to N where N is a hyperparameter. For spatial cross-attention, this results in an attention map M of dimensions $(B * F) \times (H_i * W_i) \times N$ where, in general, each spatial location has different values $\neq 1$ for the N different tokens. Similarly, the temporal cross-attention map M has dimensions $(B * H_i * W_i) \times F \times N$ with values $\neq 1$. Since SVD was pre-trained using multiple text embedding tokens as input, the code can already handle multiple tokens, so mainly the initialization of the motion-text embedding as well as some input dimensions have to be adapted slightly.

C.4.2 Different Tokens per Frame. As explained in the main paper, we propose to learn different sets of tokens per frame for the *spatial* cross-attention to obtain a higher temporal granularity of the motion. The default SVD implementation broadcasts the embedding from dimensions $B \times N \times d$ across all frames to $(B * F) \times N \times d$ (where $N = 1$ originally). We instead learn a larger spatial motion-text embedding of dimensions $B \times F \times N \times d$ and reshape it to $(B * F) \times N \times d$. We keep the dimensions of the temporal motion-text embedding at $B \times N \times d$ and learn it separately. Therefore, the dimensions of the combined spatial and temporal motion-text embedding is $B \times (F + 1) \times N \times d$.

C.4.3 Analogy. To give an intuitive analogy for our motion-text embedding inflation, think of building a house. Instead of using a single tool for every part of the house, it is more efficient to have N different tools depending on the spatial location on a given floor—like a hammer for the floor and a drill for the wall. Moreover, each of the F floors of the house might need a different set of tools. For example, the roof requires different tools compared to the walls. Similarly, in our approach, we use multiple tokens to handle different aspects of the motion.

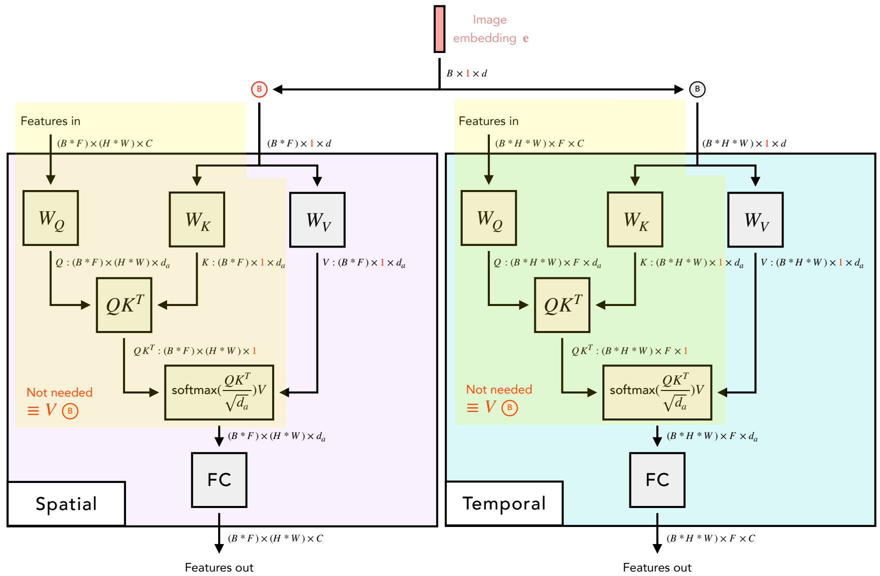
D Motion-Text Embedding Analysis

SVD was pre-trained as a text-to-video model and dropped the image (latent) input for some percentage of training iterations for classifier-free guidance [Ho and Salimans 2021]. We find that SVD can produce somewhat reasonable videos with the image (latent) input zeroed out and only the CLIP [Radford et al. 2021] image embedding as input, especially if we increase the classifier-free guidance scale (e.g., to 10). We can use this to visualize our learned motion-text embedding with an unconditional appearance.

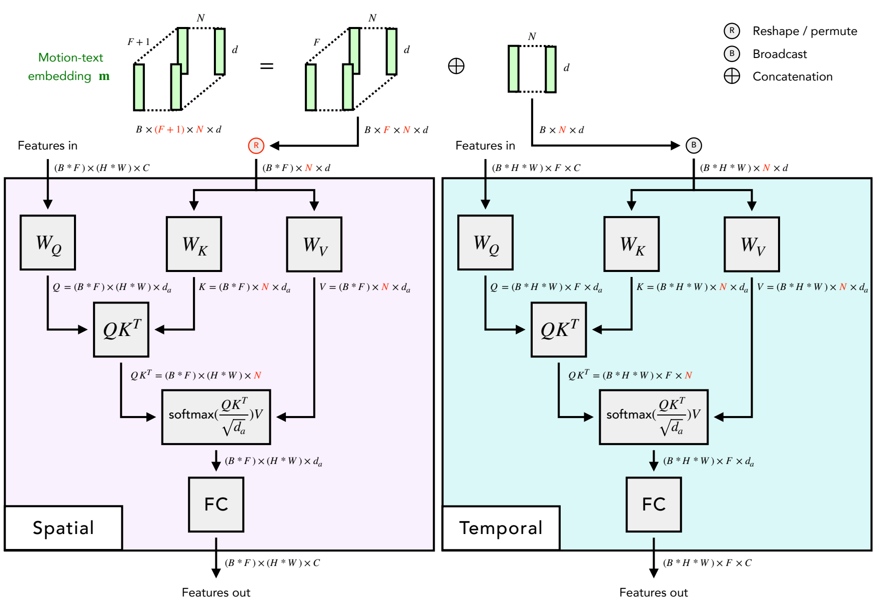
Fig. 2 shows motion visualizations of our motion-text embedding for a "jumping jacks" motion after different numbers of optimization iterations and the generated videos for a given target image side-by-side. Starting around iteration 500, a person doing a "jumping jacks" motion can be seen in the visualizations. Beyond 1000 iterations, the motion visualizations become more abstract, but the generated motions in the conditional case remain of high quality. Notably, the appearance and position of the people do not match those of the motion reference video (from Fig. 3). Furthermore, the position of the people is different in the conditional and unconditional videos, but all videos have a similar semantic motion. This demonstrates that our motion-text embedding neither encodes the appearance nor the exact spatial positioning of the objects extensively, likely for reasons described in the motivation section of the main paper.

E Applicability to Other Video Diffusion Models

We believe our approach should generalize to other architectures, including ones based on transformers, as long as the image-to-video model mainly extracts appearance from the image input and motion from text/image embeddings. This appears to hold for HunyuanVideo-I2V [Kong et al. 2024]; when we repeated the experiment from observation 1 of the main paper, the horse remained white despite the text input specifying a "pink" horse. For video models with full spatio-temporal attention (e.g., HunyuanVideo-I2V), rather than SVD's separate spatial and temporal attention, it remains to be investigated whether inflating the motion-text embedding to have different tokens per frame is strictly necessary for good performance, as it was for SVD.



(a) Default SVD [Blattmann et al. 2023a]: Since the image embedding **e** has only one token, the softmax operation causes all entries of the cross-attention maps to be 1. Therefore, the section highlighted in yellow simplifies to a broadcasted version of the value vector of that token.



(b) Inflated SVD [Blattmann et al. 2023a] (Ours): We use N tokens instead of 1, so the model now dynamically attends to different tokens depending on the spatial and temporal location. Additionally, we use different sets of tokens per frame for the spatial cross-attention instead of broadcasting the same tokens to all frames.

Fig. 1. Technical diagrams of the motion-text embedding and cross-attention inflation showing the dimensions of the features of the spatial and temporal cross-attention blocks. The changes between the default SVD [Blattmann et al. 2023a] and our inflated version are shown in red font. B = batch size, F = number of frames, C = number of channels, H = height, W = width, d = embedding dimension, d_a = attention dimension, N = token dimension, W_Q = query weight matrix, W_K = key weight matrix, W_V = value weight matrix, Q = queries, K = keys, V = values, FC = fully connected layer. For simplicity, the multiple attention heads and block level i indices are not shown.

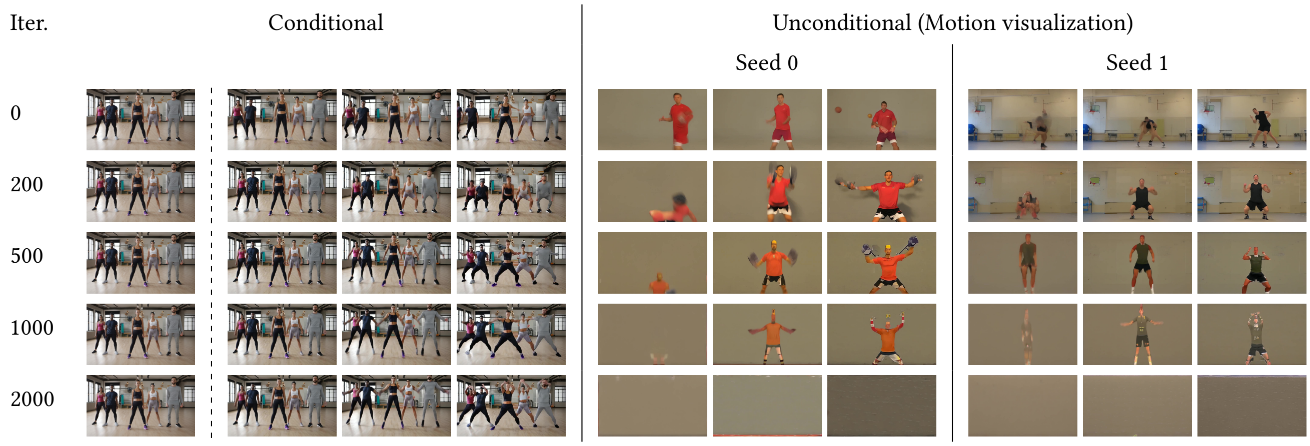


Fig. 2. Motion visualization. We generate videos using our optimized motion-text embedding for a "jumping jacks" motion (reference from Fig. 3) both with the image input (conditional) and without (unconditional) after a different number of optimization iterations. Note how the appearance of the unconditional generations differs from the motion reference video and varies with different seeds. Further observe that our method effectively generates similar semantic motions without needing or enforcing spatial alignment.

F Additional Evaluation

F.1 Additional Information for the Compared Methods

F.1.1 Choice of Compared Methods. To the best of our knowledge, our method is the first to tackle the general motion transfer task in the image-to-video setting. As a result, there are no direct competitor methods. Instead, we evaluate the most closely related general methods, (which were originally designed for slightly different tasks) on our problem. We considered the three most similar classes of methodology and compared our method with a representative of each class:

(1) Image-to-video model with explicit, dense motion representation: VideoComposer [Wang et al. 2024d]
(2) Image-to-video model with implicit motion representation: MotionClone [Ling et al. 2024] (our method falls into this category)
(3) Text-to-video model with implicit motion representation: MotionDirector [Zhao et al. 2024]

Methods within each class tend to have certain inherent drawbacks in common. Specifically, methods based on explicit, dense motion representations (class (1)) transfer spatial but not semantic motion and may leak the reference video's structure; and methods based on text-to-video models (class (3)) do not directly take a target image input, compromising the preservation of the target's appearance and layout. We believe that comparing to one method from each class is sufficient to demonstrate the types of artifacts, as adding more methods would not address the inherent limitations shared within the class.

Additional practical considerations: The following related methods did not have corresponding code publicly available at the time of writing: Diffusion as Shader [Gu et al. 2025] (class (1)), Go-With-The-Flow [Burgert et al. 2025] (class (1)), GenVideo [Harsha et al. 2024], and CustomTTT [Bi et al. 2025] (class (3)). The following methods are computationally infeasible given the size of our evaluation data set and our computational resources available: LAMP [Wu et al. 2024a] (class (2), $\approx$ 14 GPU hours per reference video), and DreamVideo [Wei et al. 2024] (class (3), $\approx$ 1 GPU hours per motion reference video and $\approx$ 2 GPU hours per target image).

Furthermore, we do not compare to methods using explicit, sparse motion representations (see Section B.4.1) because it is unclear how to automatically extract sparse motion inputs from motion reference video. We also do not compare to methods based on text-to-video models without learned appearance [Materzyńska et al. 2024; Wang et al. 2024b; Yatim et al. 2023; Zhang et al. 2023b] because defining appearance solely through text is insufficient to accurately preserve the target image appearance.

F.1.2 Implementation Details. We used the official implementations for all compared methods and followed their installation and usage instructions closely. For the methods requiring a text input, we manually captioned images and videos for the qualitative evaluation. We initially tried several image and video captioning methods, but their captions all led to worse results than manual captions that follow the captions used in the papers more closely. For the quantitative evaluation, we used the corresponding caption from the Something-Something V2 data set [Goyal et al. 2017].

F.2 Additional Qualitative Comparisons to Baseline

Fig. 3 compares our method with the Stable Video Diffusion (SVD) [Blattmann et al. 2023a] baseline for multiple motions and seeds. It further visualizes our motion-text embeddings and SVD's image embeddings with unconditional appearances. While this is not a fair comparison—since SVD does not incorporate the motion reference video—the goal is to analyze and better understand the capabilities of both methods.

As expected, SVD's generated results generally do not follow the reference motions. In rare cases, the motion does match somewhat, likely because the expected motion of the target image is similar

Fig. 3. Comparison to Stable Video Diffusion [Blattmann et al. 2023a] baseline. We compare our method to Stable Video Diffusion (SVD) for multiple motions and seeds. While SVD often fails to align with the motion reference and is highly influenced by the seed, our motion-text embedding guides the model to generate videos with matching motion, minimizing variations caused by the seed.

to reference motion, as seen in the horse/dog example. However, close inspection reveals that the gaits of the generated videos do differ and that the dog's tail wiggles in the third example. Our method's motion-text embeddings seem to capture the motion of the reference videos well, i.e., replacing the image embedding of the target image with the motion-text embedding leads to successful motion transfers for all three seeds. In our method, different seeds produce varying artifacts (e.g., arms for the jumping jacks example) while maintaining largely consistent motions. For the horse/dog example, our method generates videos where the motion closely follows the horse's gait, as explored further in Fig. 7.

Generating results with an unconditional appearance, i.e., where the image (latent) input is zeroed out, provides insight into the information encoded in the embeddings. However, note that the visualization is not always easily interpretable, depending on the motion, the optimization iteration, and the seed. SVD uses the CLIP [Radford et al. 2021] image embedding of the target image, resulting in videos that depict characters semantically similar to those in the target image. The motions vary with the seed and do not consistently align with those in videos generated with the image (latent) condition. In contrast, our method uses the motion-text embeddings optimized on the motion reference video. While the exact appearance (e.g., colors) varies with the seed, the object types seem to resemble those of the motion reference video. This may stem from initializing the motion-text embedding with image embeddings extracted from the motion reference video. The encoding of object types in the motion-text embedding may also explain the occasional structure leakage noted in the limitations section.

Results generated with SVD frequently exhibit significant artifacts (e.g., first two seeds for the jumping jacks example) and appearance changes (e.g., last two seeds for the yawning example). As our method builds on SVD's frozen weights, we inherit some of SVD's issues, as described in the limitations section. However, by conditioning the model on a reference motion, our results tend to appear more realistic and contain fewer artifacts. We hypothesize that this improvement arises because the model leverages the provided (realistic) motion rather than needing to hallucinate it from scratch, simplifying the overall task. Additionally, SVD often generates static objects with moving cameras in our experience. We suggest that motion transfer methods, like ours, can help generate more natural and diverse motions.

F.3 Additional Qualitative Comparisons to State-of-the-Art Methods

To further demonstrate the effectiveness of our method in transferring semantic motion from a reference video to target images, we generated videos using state-of-the-art competing methods for the same examples presented in the main paper. These results, covering a range of motion types and complexities, are provided in Fig. 4 and Fig. 5. As before, competing methods suffer from problems inherent to their class of methods. Stable Video Diffusion [Blattmann et al. 2023a], lacking a motion input, typically fails to follow the reference motion. VideoComposer [Wang et al. 2024d], an image-to-video method with dense motion inputs, struggles when the reference video's motions are not aligned with the input image. In such cases, the method applies the spatial but not semantic motion, leading to either unwanted background movement or the foreground object morphing into the spatial position where the motion occurs in the reference video. MotionDirector [Zhao et al. 2024], based on a text-to-video model, cannot directly use the target image as input and must instead learn its appearance. As a result, the generated videos often deviate in appearance and spatial layout from the target image.

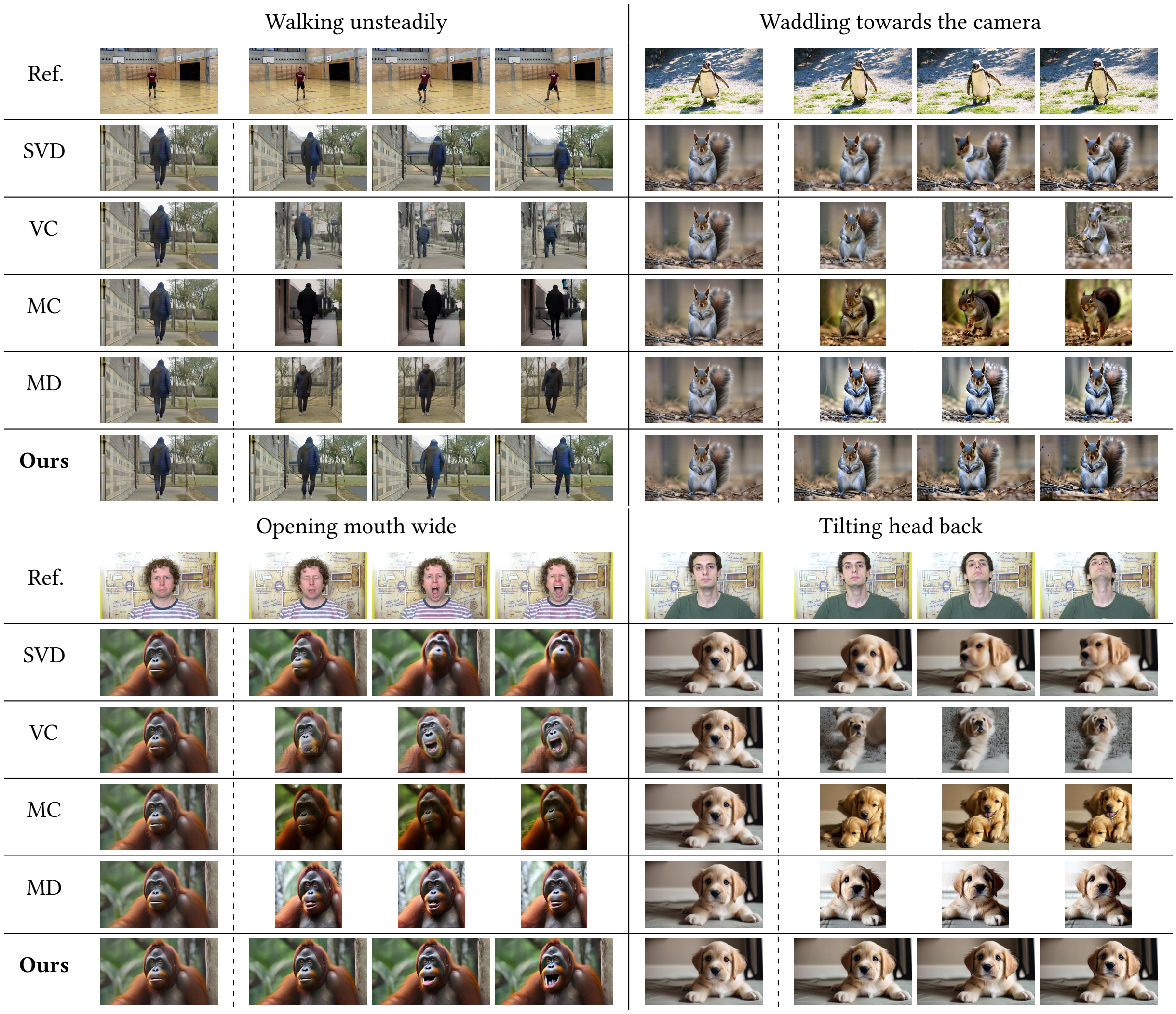


Fig. 4. Qualitative evaluation for additional examples (1/2). We compare our method to SVD = Stable Video Diffusion [Blattmann et al. 2023a] (baseline, no motion input), VC = VideoComposer [Wang et al. 2024d], MC = MotionClone [Ling et al. 2024], and MD = MotionDirector [Zhao et al. 2024] for four different motions and target images.

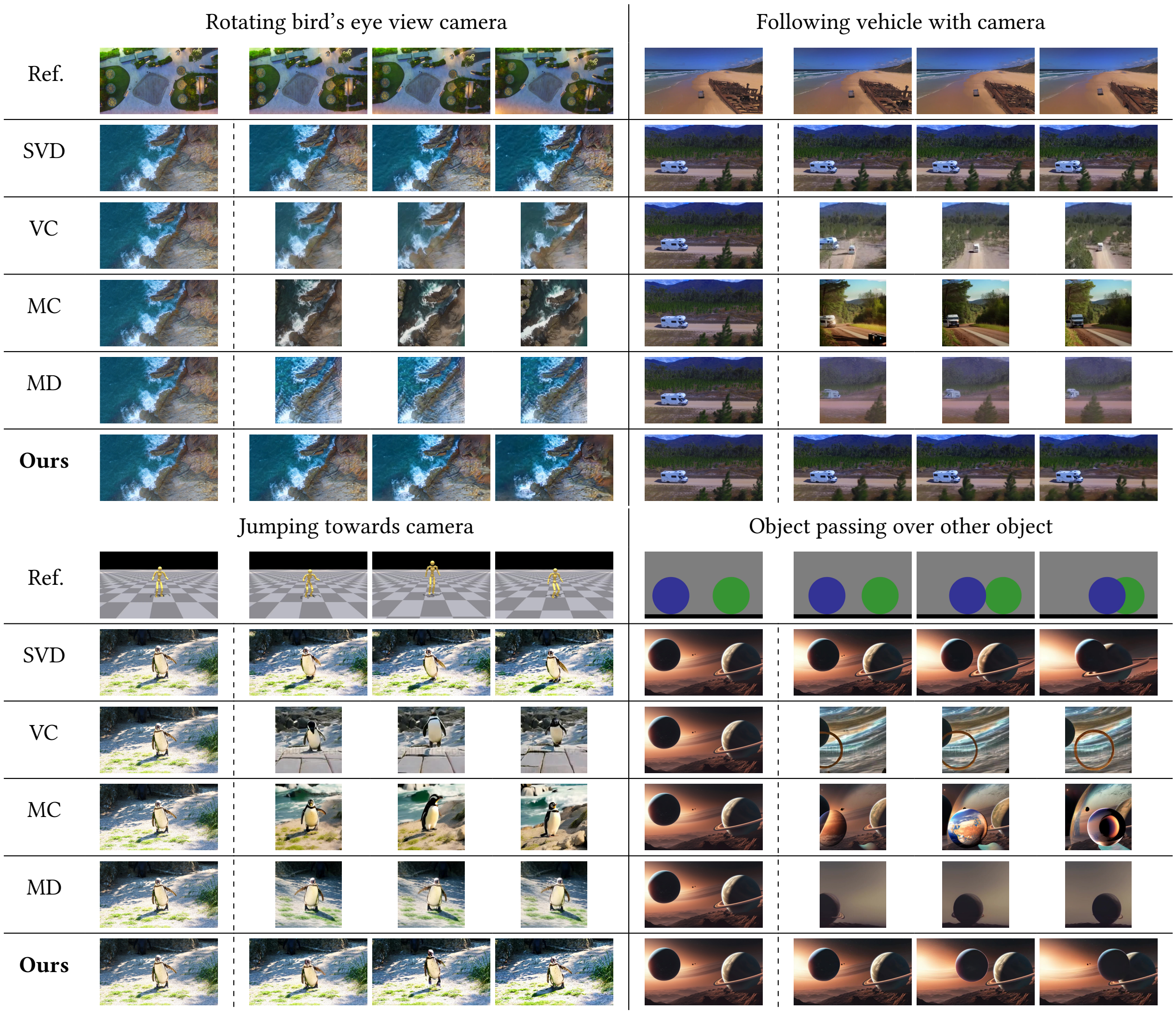


Fig. 5. Qualitative evaluation for additional examples (2/2). We compare our method to SVD = Stable Video Diffusion [Blattmann et al. 2023a] (baseline, no motion input), VC = VideoComposer [Wang et al. 2024d], MC = MotionClone [Ling et al. 2024], and MD = MotionDirector [Zhao et al. 2024] for four different motions and target images.

Table 1. Quantitative evaluation data. List of video IDs from the Something-Something V2 data set [Goyal et al. 2017] used in our quantitative evaluation.

Class ID: Label	Video ID for Motion Reference Video: Video IDs for Target Images
0: Approaching something with your camera	31416: 174027, 49364, 179191, 58108, 219270, 124642, 18253, 112846, 75372, 201968
23: Letting something roll down a slanted surface	97908 : 220450, 22070, 46282, 136926, 216643, 109913, 137160, 69704, 19903, 86892
27: Lifting something up completely without letting it drop down	144105: 181548, 167709, 81608, 132100, 167837, 46057, 158390, 41755, 93247, 106014
32: Moving away from something with your camera	121394: 3201, 100064, 35438, 44298, 123636, 4328, 178356, 76980, 71173, 33210
36: Moving something and something away from each other	51295: 4443, 88084, 76718, 132951, 49285, 43627, 45186, 18456, 18788, 142654
37: Moving something and something closer to each other	87711: 180193, 137350, 39979, 150128, 10055, 16205, 208340, 97632, 94171, 99258
41: Moving something away from the camera	207150: 205156, 108506, 139808, 44794, 68922, 197965, 201362, 153856, 21809, 211202
44: Moving something towards the camera	160529: 145447, 30260, 118270, 10405, 66666, 154312, 157137, 106357, 164212, 176798
92: Pulling two ends of something so that it separates into two pieces	187909: 162071, 51196, 87892, 11780, 75398, 148274, 113149, 177507, 47061, 28237
165: Turning the camera downwards while filming something	169117: 120585, 131318, 68372, 104829, 162135, 124382, 108641, 98914, 197549, 213899

Table 2. Quantitative evaluation aggregated by motion category (camera/object). As in the main paper, we compare our method to Stable Video Diffusion [Blattmann et al. 2023a] (baseline, no motion input), VideoComposer [Wang et al. 2024d], MotionClone [Ling et al. 2024], and MotionDirector [Zhao et al. 2024]. The first value in each cell corresponds to camera motions and the second to object motions. The best performing method per column is marked in bold.

Method	Image Appearance Preservation			Video Motion Fidelity				Overall
	CLIP-Avg ↑	CLIP-1st ↑	User rank ↓	Acc-Top-1 ↑	Acc-Top-5 ↑	Cos-Sim ↑	User rank ↓	User rank ↓
Stable Video Diffusion	**0.837**/**0.849**	0.842/0.857	**1.215**/**1.378**	4%/2%	4%/6%	0.398/0.342	4.689/3.733	3.311/2.333
VideoComposer	0.713/0.726	0.853/0.860	3.867/3.704	64%/24%	82%/42%	0.575/0.419	2.941/3.119	3.407/3.696
MotionClone	0.610/0.664	**0.881**/0.890	4.778/4.393	48%/26%	80%/44%	0.555/0.491	3.215/3.059	4.385/4.015
MotionDirector	0.738/0.762	0.752/0.774	3.185/3.859	38%/24%	58%/58%	0.545/0.501	3.067/2.733	2.785/3.333
Ours	0.745/0.813	0.873/**0.894**	1.956/1.667	**72%**/**36%**	**86%**/**66%**	**0.785**/**0.606**	**1.089**/2.356	**1.111**/**1.622**

F.4 Additional Information for the Quantitative Evaluation

We selected the action classes from the Something-Something V2 data set [Goyal et al. 2017] according to the following criteria:

- All interacting objects typically appear in the start frame.
- The action is typically long enough, so that it appears in most of the frames.
- The class is sufficiently different from other classes.

We then extracted the first 11 examples of the given class (with some manual filtering in case the above criteria is not met) and took the first video as motion reference video and the first frames of the other 10 for the target images. Table 1 lists the final class IDs and video IDs used.

The 10 action classes used in our evaluation can be grouped into two categories: five involving camera motion (IDs: 0, 32, 41, 44, 165) and five involving object motion (IDs: 23, 27, 36, 37, 92). Table 2 provides the quantitative results from the main paper, aggregated by motion category. We observe that image appearance preservation is generally worse for camera motions. This is likely because strong camera movements cause significant changes in the visual content. In contrast, video motion fidelity is typically higher for camera motions, possibly because the movements are more uniform and linear, and spatial alignment between the motion reference video and target image is less critical. As a result, methods that mostly transfer spatial rather than semantic motion (e.g., VideoComposer [Wang et al. 2024d]) can still perform well for camera motions.

Our method consistently outperforms all compared methods across both motion categories in terms of video motion fidelity. Notably, for object motions, the advantage over MotionDirector [Zhao et al. 2024] is even more pronounced than the mean user rank suggests: our method was selected as the best in 58% of comparisons, compared to only 22% for MotionDirector. The relatively high mean rank of our method can be attributed to occasional failure cases (further discussed in Section I) which greatly affect the average. In terms of appearance preservation, Stable Video Diffusion (SVD) [Blattmann et al. 2023a] slightly outperforms our approach, though this may be because SVD often produces very limited motion, making it easier to maintain the appearance of the input image. When considering the overall user preference, our method shows a substantial lead: it was voted best among the five compared methods in 90% of the evaluations for camera motions and 65% for object motions. Notably, for object motions, Stable Video Diffusion, despite lacking any motion input, was voted best in 33% of cases, while all other methods combined accounted for just 2%. We believe this can be explained as follows: when our method succeeds, it significantly outperforms all other methods; when it fails, e.g., due to challenging motion reference videos or target images, SVD's conservative, low-motion outputs tend to be the most visually coherent and thus the preferred choice.

G Additional Ablation Study Results

In the main paper, we show results for different settings of the motion-text embedding size for one motion. In Fig. 6, we show two more examples for this ablation. As previously stated, the biggest performance improvement can be seen between rows 2 and 3 for each example, i.e., once there are *different tokens per frame*. Note that the differences for the horse/dog example are best seen in the attached videos. While the dog is always moving to the right, the

Table 3. Quantitative results for our ablation. Here, we compare various settings for the dimensions of the motion-text embedding. Table (a) shows the overall scores aggregated over all motion categories, whereas (b) shows the scores aggregated by the motion category of the motion reference videos, where the first value in each cell corresponds to camera motions and the second to object motions. The best performing method per column is marked in bold.

(a) Overall

Method	Image Appearance Preservation		Video Motion Fidelity		
	CLIP-Avg ↑	CLIP-1st ↑	Acc-Top-1 ↑	Acc-Top-5 ↑	Cos-Sim ↑
Ours ($F' = 1, N = 1$)	**0.788**	0.875	44%	62%	0.619
Ours ($F' = 1, N = 15$)	0.785	0.878	44%	65%	0.637
Ours ($F' = 15, N = 1$)	0.776	0.883	52%	**77%**	0.704
Ours ($F' = 15, N = 15$)	0.776	**0.886**	**56%**	**77%**	**0.705**
Ours ($F' = 15, N = 5$, Default)	0.779	0.884	54%	76%	0.696

(b) By motion category (camera/object)

Method	Image Appearance Preservation		Video Motion Fidelity		
	CLIP-Avg ↑	CLIP-1st ↑	Acc-Top-1 ↑	Acc-Top-5 ↑	Cos-Sim ↑
Ours ($F' = 1, N = 1$)	**0.755**/**0.821**	0.865/0.885	64%/24%	76%/48%	0.722/0.516
Ours ($F' = 1, N = 15$)	0.754/0.817	**0.874**/0.881	70%/18%	82%/48%	0.758/0.517
Ours ($F' = 15, N = 1$)	0.743/0.810	0.872/0.894	74%/30%	**86%**/**68%**	0.807/0.600
Ours ($F' = 15, N = 15$)	0.740/0.813	**0.874**/**0.899**	**78%**/34%	**86%**/**68%**	**0.810**/0.601
Ours ($F' = 15, N = 5$, Default)	0.745/0.813	0.873/0.894	72%/**36%**	**86%**/66%	0.785/**0.606**

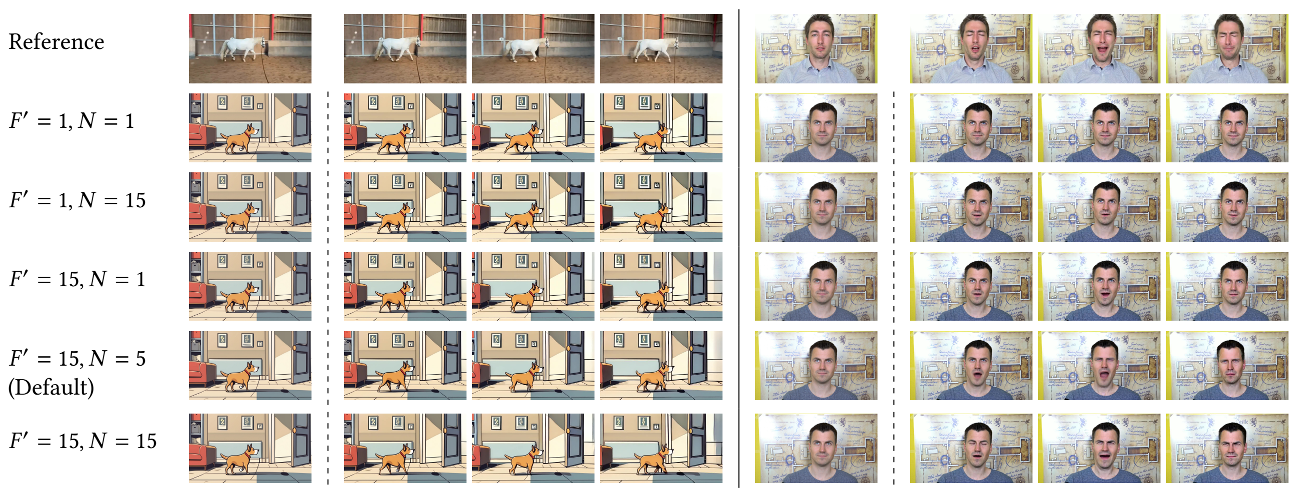


Fig. 6. Ablation with additional examples. Inflating the motion-text embedding, by having more tokens N or by having different tokens for each frame (where $F' = F + 1 = 15$), greatly improves the motion transfer.

speed and style of the gait does not match the reference for the first two rows.

To quantitatively evaluate the settings of the motion-text embedding size, we followed the same protocol as for the quantitative evaluation in the main paper. The results are listed in Table 3 and align well with our observations. Whereas the image appearance preservation is similar throughout, the motion fidelity improves slightly as we increase the token dimension N (when $F' = 1$) and significantly once we use *different tokens per frame* ($F' = 15$). If $F' = 15$, the embedding dimension N does not seem to affect the results much for the tested reference motion videos. In addition to the results aggregated over all evaluation videos in Table 3a, we provide results aggregated by the motion category (camera/object) of the motion reference videos in Table 3b. The results suggest that our proposed motion-text embedding inflation improves the performance for camera and object motions alike.

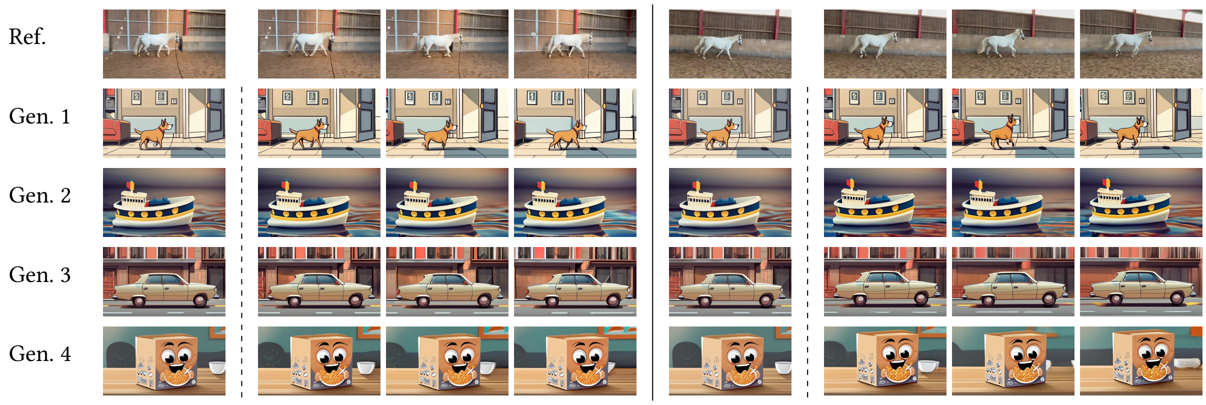


Fig. 7. Motion style transfer. Our learned motion-text embeddings do not only store the rough motion category but also the style of the motion. Here, we apply two different gaits to the same target image: a horse trot (smooth) and a canter (rocking). The resulting videos for the cartoon dog are not only showing the dog moving, but their motions also closely match the motion reference video's gait style. Furthermore, the extreme cross-domain examples with the boat, car, and cereal box show that the essence of the motion style is preserved even across completely different objects.

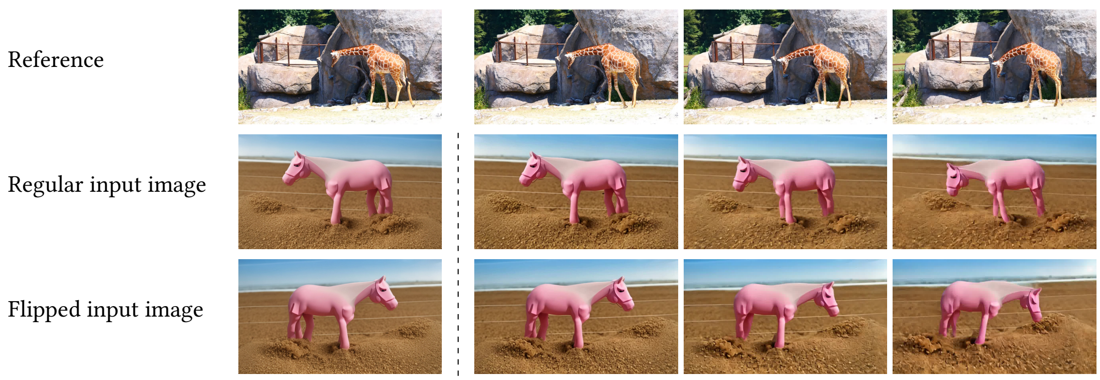


Fig. 8. Semantic motion transfer. Our learned motion-text embeddings store the semantic motion (animal moving in the direction it is facing and moving its head down) rather than the spatial motion (animal moving from right to left and left part is going down). This can be seen in the above example where we apply the same learned motion-text embedding to a flipped input image, and our method produces semantically similar results.

H Additional Results

Fig. 7 shows that our method does not only apply the rough motion category but also its style, even in difficult cases where the domains differ vastly, e.g., transferring the motion of a horse to a cereal box. Furthermore, these examples demonstrate that our method can transfer joint subject and camera motion. Fig. 8 demonstrates that our method transfers the same semantic rather than spatial motion by applying the same learned motion to a flipped target image. Fig. 9 shows additional results of our method, where we apply the same optimized motion to different target images to showcase our method's impressive cross-domain capabilities and temporal alignment. Lastly, Fig. 10 transfers the same four camera motions to four different target images in a grid, demonstrating the robustness of our method for camera motions.

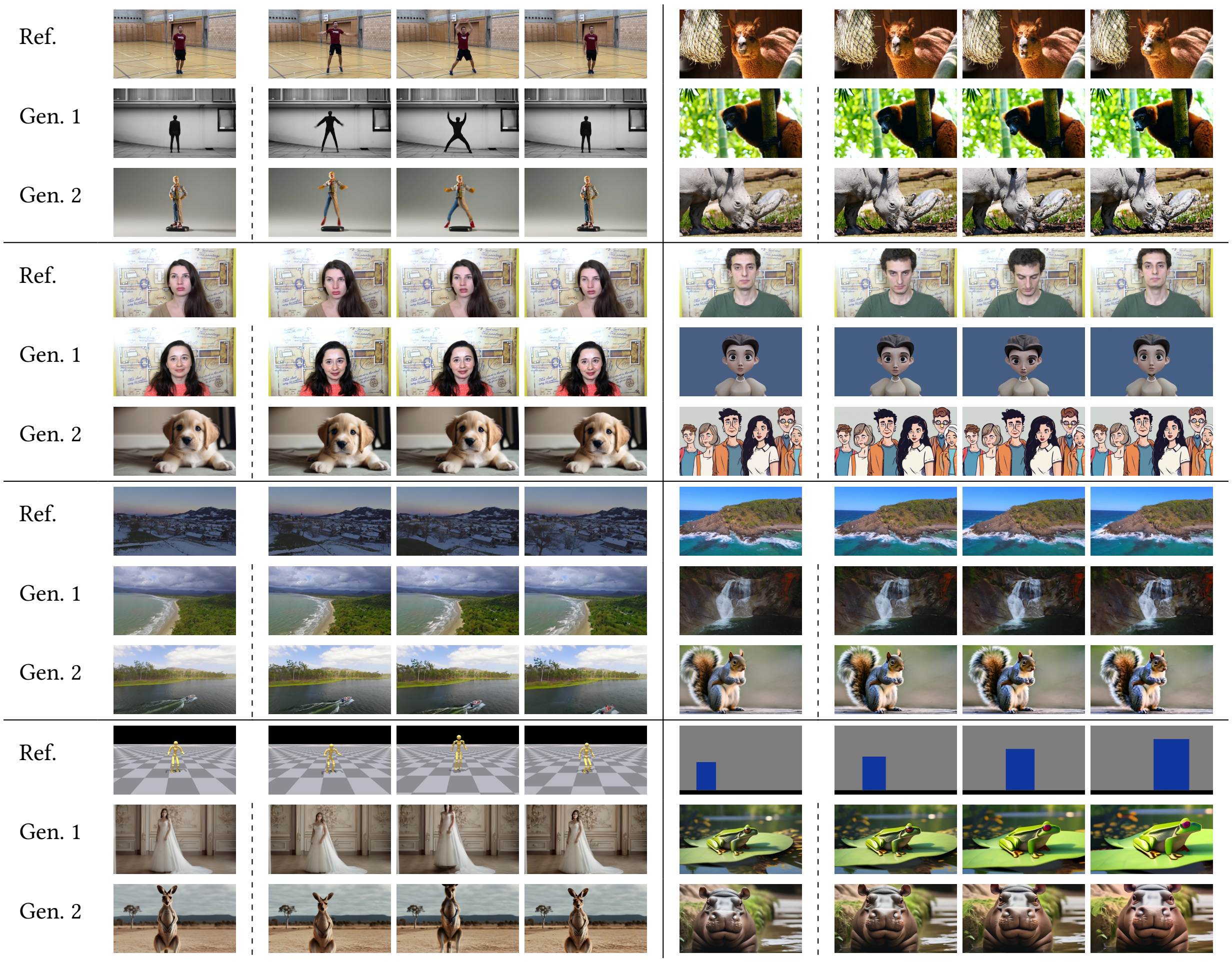


Fig. 9. Additional results. Our learned motion-text embeddings can be applied to multiple target images, resulting in semantically similar motions.

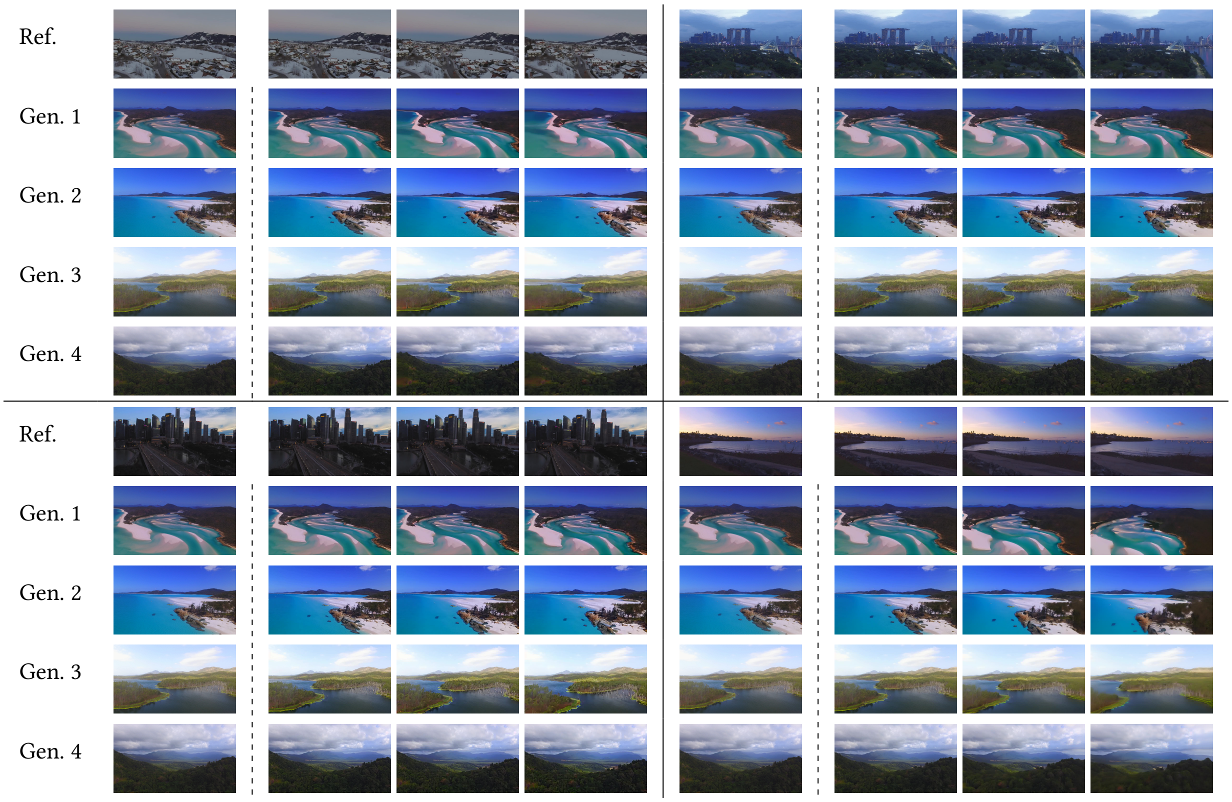


Fig. 10. Camera motion grid. Our learned motion-text embeddings handle camera motions robustly, enabling us to apply a given motion to various target images and various motions to a given target image. The results are best seen in the project website.

I Failure Rate Analysis

As is common practice in diffusion-based video generation, we sampled multiple outputs per input and selected the best for display. Quantifying failure rates is difficult, as success can be subjective and depends heavily on the complexity of the motion. Table 2 shows metrics broken down by motion category. The Acc-Top-1 metric reports the percentage of videos correctly classified by an action recognition model [Tong et al. 2022] and can loosely be interpreted as a success rate for the semantic motion transfer (independent of visual artifacts). Our method achieves much higher accuracy for camera motions (72%) than for object motions (36%). It is worth noting that the main challenge in the quantitative evaluation on Something-Something V2 [Goyal et al. 2017] stems from the domain gap between the motion reference video and the target image—e.g., transferring a toy car rolling down a book to a pen rolling down a rock—rather than the motion complexity itself. In contrast, our qualitative experiments explored more complex motions to better test the limits of our method, and thus had higher failure rates: approximately 1 in 10 motions resulted in good motion transfers for more than half of the tested target images. To give a more intuitive sense of when our method succeeds or fails, we list motion categories based on how reliably they could typically be transferred in Table 4.

In our experiments, we observed that the reconstruction quality of the motion reference video, i.e., applying the optimized motion-text embedding to the first frame of the motion reference video, is a strong indicator of the final motion transfer performance. If the model fails to reconstruct the reference video accurately, it suggests that the optimized motion-text embedding does not effectively capture the semantics of the motion. In such cases, applying the same embedding to a different target image typically also fails. This issue is illustrated in Fig. 11, where the reconstructed video collapses the person into a blob-like shape rather than depicting a realistic forward roll. The same collapse occurs when transferring the motion to a different target image. One contributing factor may be the use of a simple mean-squared error loss, which can lead to pixels being placed in roughly the correct spatial positions, even if the resulting motion does not semantically match the reference. Another potential reason for failure is that some motions may be out-of-domain for the pre-trained Stable Video Diffusion [Blattmann et al. 2023a]. Since our approach optimizes only the input motion-text embedding without fine-tuning the model itself, it is challenging to capture entirely novel or complex motion types that the model has not seen during training. To mitigate these issues, we encourage future work to explore more semantically meaningful loss functions, regularize the embedding to remain closer to the original CLIP [Radford et al. 2021] space, or adopt recent video diffusion models with stronger motion understanding, such as VideoJAM [Chefer et al. 2025].

Table 4. Summary of motion types by performance.

Performance	Motion Types
Motions good Quality good	**Camera motions:** bird's-eye panning/zooming/rotation, panoramas, smooth drone flights, object tracking **Common head motions:** nodding, facial expressions (surprise, yawning, opening mouth) **Some full-body motions:** walking (human to human, four-legged to four-legged), jumping jacks **Handcrafted motions with small domain gap:** colliding/passing circles of similar shapes/colors
Motions good/okay Quality bad	**Fast motions:** boxing, fast running animals (left/right limb confusion) **Head motions with drastic appearance changes:** frontal-to-profile rotations, extremely wide mouth openings, revealing teeth from closed mouth **Some full-body motions:** jumping forward far, walking into jump, karate kicks **Handcrafted motions where target object has many details:** texture-free bouncing ball transferred to soccer ball with many patches, stick figure to detailed human / two-legged animal
Motions bad	**Fine-grained motions:** tongue movement, eyebrow raises, small/distant actions **Emerging objects:** hand entering frame **Large domain gap:** human face motions to minimalistic cartoon or ostrich, human to kangaroo, bouncing ball to landscape scene with sun **Complex full-body motions:** running into forward roll, handstands, swinging arm punch, yoga/stretching

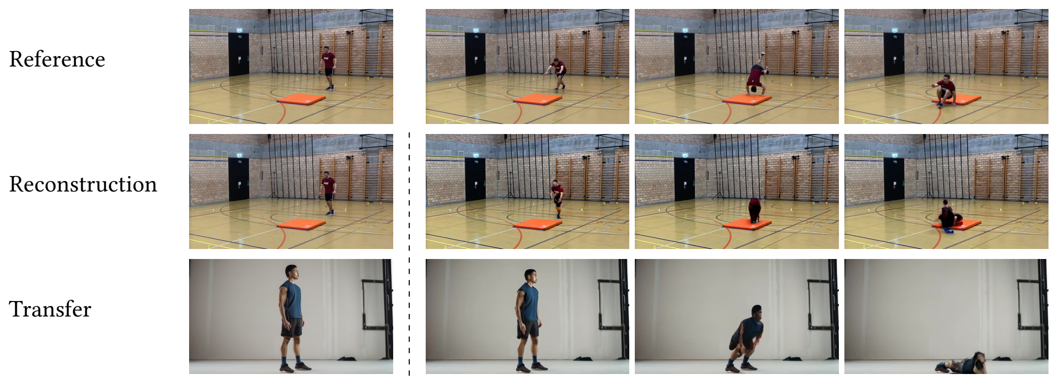


Fig. 11. Failure case with poor reconstruction. When the optimized motion-text embedding fails to accurately reconstruct the reference motion, the subsequent transfer to a new target typically fails as well.

References

Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. 2024. VD3D: Taming Large Video Diffusion Transformers for 3D Camera Control. *arXiv preprint arXiv:2407.12781* (2024).

Jianhong Bai, Tianyu He, Yuchi Wang, Junliang Guo, Haoji Hu, Zuozhu Liu, and Jiang Bian. 2024. UniEdit: A Unified Tuning-Free Framework for Video Motion and Appearance Editing. *arXiv preprint arXiv:2402.13185* (2024).

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. 2024. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*. 1–11.

Xiuli Bi, Jian Lu, Bo Liu, Xiaodong Cun, Yong Zhang, Weisheng Li, and Bin Xiao. 2025. Customttt: Motion and appearance customized video generation via test-time training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 1871–1879.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023a. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023).

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

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. 2024. Video generation models as world simulators. (2024). https://openai.com/research/video-generation-models-as-world-simulators

Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. 2025. Go-with-the-Flow: Motion-Controllable Video Diffusion Models Using Real-Time Warped Noise. *arXiv preprint arXiv:2501.08331* (2025).

Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. 2023. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 23206–23217.

Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. 2019. Everybody dance now. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5933–5942.

Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. 2025. VideoJAM: Joint Appearance-Motion Representations for Enhanced Motion Generation in Video Models. *arXiv preprint arXiv:2502.02492* (2025).

Changgu Chen, Junwei Shu, Lianggangxu Chen, Gaoqi He, Changbo Wang, and Yang Li. 2024. Motion-Zero: Zero-Shot Moving Object Control Framework for Diffusion-Based Video Generation. *arXiv preprint arXiv:2401.10150* (2024).

Tsai-Shien Chen, Chieh Hubert Lin, Hung-Yu Tseng, Tsung-Yi Lin, and Ming-Hsuan Yang. 2023a. Motion-conditioned diffusion model for controllable video synthesis. *arXiv preprint arXiv:2304.14404* (2023).

Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. 2023b. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840* (2023).

Soon Yau Cheong, Duygu Ceylan, Armin Mustafa, Andrew Gilbert, and Chun-Hao Paul Huang. 2024. Boosting camera motion control for video diffusion transformers. *arXiv preprint arXiv:2410.10802* (2024).

Zuozhuo Dai, Zhenghao Zhang, Yao Yao, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. 2023. AnimateAnything: Fine-Grained Open Domain Image Animation with Motion Guidance. *arXiv e-prints* (2023), arXiv–2311.

Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. 2022. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*. 2663–2671.

Daniel Garibi, Or Patashnik, Andrey Voynov, Hadar Averbuch-Elor, and Daniel Cohen-Or. 2024. Renoise: Real image inversion through iterative noising. In *European Conference on Computer Vision*. Springer, 395–413.

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, et al. 2024. Motion Prompting: Controlling Video Generation with Motion Trajectories. *arXiv preprint arXiv:2412.02700* (2024).

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. 2024. TokenFlow: Consistent Diffusion Features for Consistent Video Editing. In *The Twelfth International Conference on Learning Representations*.

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

Yuchao Gu, Yipin Zhou, Bichen Wu, Licheng Yu, Jia-Wei Liu, Rui Zhao, Jay Zhangjie Wu, David Junhao Zhang, Mike Zheng Shou, and Kevin Tang. 2024. Videoswap: Customized video subject swapping with interactive semantic point correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7621–7630.

Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. 2025. Diffusion as Shader: 3D-aware Video Diffusion for Versatile Video Generation Control. *arXiv preprint arXiv:2501.03847* (2025).

Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. 2024b. LivePortrait: Efficient Portrait Animation with Stitching and Retargeting Control. *arXiv preprint arXiv:2407.03168* (2024).

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. 2024a. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In *The Twelfth International Conference on Learning Representations*.

Sai Sree Harsha, Ambareesh Revanur, Dhwanit Agarwal, and Shradha Agrawal. 2024. GenVideo: One-shot target-image and shape aware video editing using T2I diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7559–7568.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. 2024. CameraCtrl: Enabling Camera Control for Text-to-Video Generation. *arXiv preprint arXiv:2404.02101* (2024).

Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. 2023. Unsupervised semantic correspondence using stable diffusion. *Advances in Neural Information Processing Systems* 36 (2023), 8266–8279.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. 2023. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *The Eleventh International Conference on Learning Representations*.

Jonathan Ho and Tim Salimans. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Chen Hou, Guoqiang Wei, Yan Zeng, and Zhibo Chen. 2024. Training-free Camera Control for Video Generation. *arXiv preprint arXiv:2406.10126* (2024).

Gee-Sern Hsu, Chun-Hung Tsai, and Hung-Yi Wu. 2022. Dual-generator face reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 642–650.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.

Li Hu. 2024. 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*. 8153–8163.

Teng Hu, Jiangning Zhang, Ran Yi, Yating Wang, Hongrui Huang, Jieyu Weng, Yabiao Wang, and Lizhuang Ma. 2024. MotionMaster: Training-free Camera Motion Transfer For Video Generation. *arXiv preprint arXiv:2404.15789* (2024).

Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. 2024. Peekaboo: Interactive video generation via masked-diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8079–8088.

Hyeonho Jeong, Geon Yeong Park, and Jong Chul Ye. 2024. 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*. 9212–9221.

Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. 2023. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 22623–22633.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

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

Mingxiao Li, Bo Wan, Marie-Francine Moens, and Tinne Tuytelaars. 2024b. Animate your motion: Turning still images into dynamic videos. In *European Conference on Computer Vision*. Springer, 409–425.

Ruining Li, Chuanxia Zheng, Christian Rupprecht, and Andrea Vedaldi. 2024c. Puppet-Master: Scaling Interactive Video Generation as a Motion Prior for Part-Level Dynamics. *arXiv preprint arXiv:2408.04631* (2024).

Weichuang Li, Longhao Zhang, Dong Wang, Bin Zhao, Zhigang Wang, Mulin Chen, Bang Zhang, Zhongjian Wang, Liefeng Bo, and Xuelong Li. 2023. One-shot high-fidelity talking-head synthesis with deformable neural radiance field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17969–17978.

Xiaomin Li, Xu Jia, Qinghe Wang, Haiwen Diao, Mengmeng Ge, Pengxiang Li, You He, and Huchuan Lu. 2024a. Motrans: Customized motion transfer with text-driven video diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 3421–3430.

Yaowei Li, Xintao Wang, Zhaoyang Zhang, Zhouxia Wang, Ziyang Yuan, Liangbin Xie, Ying Shan, and Yuexian Zou. 2025. Image conductor: Precision control for interactive video synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 5031–5038.

Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaian Chen, Jiaqi Wang, and Yi Jin. 2024. MotionClone: Training-Free Motion Cloning for Controllable Video Generation. *arXiv preprint arXiv:2406.05338* (2024).

Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. 2024. Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8599–8608.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. 2023. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems* 36 (2023), 47500–47510.

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

Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. 2024a. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 4117–4125.

Joanna Materzyńska, Josef Sivic, Eli Shechtman, Antonio Torralba, Richard Zhang, and Bryan Russell. 2024. NewMove: Customizing text-to-video models with novel motions. In *Proceedings of the Asian Conference on Computer Vision*. 1634–1651.

Tuna Han Salih Meral, Hidir Yesiltepe, Connor Dunlop, and Pinar Yanardag. 2024. MotionFlow: Attention-Driven Motion Transfer in Video Diffusion Models. *arXiv preprint arXiv:2412.05275* (2024).

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6038–6047.

Ron Mokady, Rotem Tzaban, Sagie Benaim, Amit H Bermano, and Daniel Cohen-Or. 2022. JOKR: Joint Keypoint Representation for Unsupervised Video Retargeting. In *Computer Graphics Forum*, Vol. 41. Wiley Online Library, 245–257.

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

Chong Mou, Mingdeng Cao, Xintao Wang, Zhaoyang Zhang, Ying Shan, and Jian Zhang. 2024. Revideo: Remake a video with motion and content control. *Advances in Neural Information Processing Systems* 37 (2024), 18481–18505.

Haomiao Ni, Yihao Liu, Sharon X Huang, and Yuan Xue. 2023. Cross-identity video motion retargeting with joint transformation and synthesis. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 412–422.

Yuval Nirkin, Yosi Keller, and Tal Hassner. 2019. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*. 7184–7193.

Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. 2024. Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. In *European Conference on Computer Vision*. Springer, 111–128.

Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. 2023. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4195–4205.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations*.

Alexander Pondaven, Aliaksandr Siarohin, Sergey Tulyakov, Philip Torr, and Fabio Pizzati. 2024. Video Motion Transfer with Diffusion Transformers. *arXiv preprint arXiv:2412.07776* (2024).

Haonan Qiu, Zhaoxi Chen, Zhouxia Wang, Yingqing He, Menghan Xia, and Ziwei Liu. 2024. FreeTraj: Tuning-Free Trajectory Control in Video Diffusion Models. *arXiv preprint arXiv:2406.16863* (2024).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.

Yixuan Ren, Yang Zhou, Jimei Yang, Jing Shi, Difan Liu, Feng Liu, Mingi Kwon, and Abhinav Shrivastava. 2024. Customize-a-video: One-shot motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*. Springer, 332–349.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

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

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* 35 (2022), 36479–36494.

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

Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. 2021. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13653–13662.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. 2023. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems* 36 (2023), 1363–1389.

Maham Tanveer, Yizhi Wang, Ruiqi Wang, Nanxuan Zhao, Ali Mahdavi-Amiri, and Hao Zhang. 2024. AnaMoDiff: 2D Analogical Motion Diffusion via Disentangled Denoising. *arXiv preprint arXiv:2402.03549* (2024).

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

Shuyuan Tu, Qi Dai, Zhi-Qi Cheng, Han Hu, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. 2024a. Motioneditor: Editing video motion via content-aware diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7882–7891.

Shuyuan Tu, Qi Dai, Zihao Zhang, Sicheng Xie, Zhi-Qi Cheng, Chong Luo, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. 2024b. MotionFollower: Editing Video Motion via Lightweight Score-Guided Diffusion. *arXiv preprint arXiv:2405.20325* (2024).

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1921–1930.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. 2022. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. 2023b. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571* (2023).

Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. 2024e. Boximator: generating rich and controllable motions for video synthesis. In *Proceedings of the 41st International Conference on Machine Learning*. 52274–52289.

Luozhou Wang, Guibao Shen, Yixun Liang, Xin Tao, Pengfei Wan, Di Zhang, Yijun Li, and Yingcong Chen. 2024b. Motion Inversion for Video Customization. *arXiv preprint arXiv:2403.20193* (2024).

Qilin Wang, Zhengkai Jiang, Chengming Xu, Jiangning Zhang, Yabiao Wang, Xinyi Zhang, Yun Cao, Weijian Cao, Chengjie Wang, and Yanwei Fu. 2024a. VividPose: Advancing Stable Video Diffusion for Realistic Human Image Animation. *arXiv preprint arXiv:2405.18156* (2024).

Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. 2021. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10039–10049.

Wen Wang, Yan Jiang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. 2023a. Zero-shot video editing using off-the-shelf image diffusion models.

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

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

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

Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. 2024c. MotionBooth: Motion-Aware Customized Text-to-Video Generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. 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*. 7623–7633.

Ruiqi Wu, Liangyu Chen, Tong Yang, Chunle Guo, Chongyi Li, and Xiangyu Zhang. 2024a. LAMP: Learn A Motion Pattern for Few-Shot Video Generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. 2024b. Draganything: Motion control for anything using entity representation. In *European Conference on Computer Vision*. Springer, 331–348.

Zeqi Xiao, Yifan Zhou, Shuai Yang, and Xingang Pan. 2024. Video Diffusion Models are Training-free Motion Interpreter and Controller. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. 2024. CamCo: Camera-Controllable 3D-Consistent Image-to-Video Generation. *arXiv preprint arXiv:2406.02509* (2024).

Wilson Yan, Andrew Brown, Pieter Abbeel, Rohit Girdhar, and Samaneh Azadi. 2023. Motion-conditioned image animation for video editing. *arXiv preprint arXiv:2311.18827* (2023).

Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. 2024. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH 2024 Conference Papers*. 1–12.

Shuai Yang, Yifan Zhou, Ziwei Liu, , and Chen Change Loy. 2023. Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation. In *ACM SIGGRAPH Asia Conference Proceedings*.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. 2025. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=LQzN6TRFg9

Zhuoqian Yang, Wentao Zhu, Wayne Wu, Chen Qian, Qiang Zhou, Bolei Zhou, and Chen Change Loy. 2020. Transmomo: Invariance-driven unsupervised video motion retargeting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5306–5315.

Danah Yatim, Rafail Fridman, Omer Bar Tal, Yoni Kasten, and Tali Dekel. 2023. Space-Time Diffusion Features for Zero-Shot Text-Driven Motion Transfer. *arXiv preprint arXiv:2311.17009* (2023).

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

Junyi Zhang, Charles Herrmann, Junhwa Hur, Eric Chen, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. 2024a. Telling left from right: Identifying geometry-aware semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3076–3085.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. 2023a. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems* 36 (2023), 45533–45547.

Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. 2023c. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145* (2023).

Yuxin Zhang, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. 2023b. MotionCrafter: One-Shot Motion Customization of Diffusion Models. *arXiv preprint arXiv:2312.05288* (2023).

Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, XIAOPENG ZHANG, Wangmeng Zuo, and Qi Tian. 2024b. ControlVideo: Training-free Controllable Text-to-video Generation. In *The Twelfth International Conference on Learning Representations*.

Jian Zhao and Hui Zhang. 2022. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3657–3666.

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

Yuyang Zhao, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. 2023. Make-a-protagonist: Generic video editing with an ensemble of experts. *arXiv preprint arXiv:2305.08850* (2023).

Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. 2024. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957* (2024).

Haitao Zhou, Chuang Wang, Rui Nie, Jinxiao Lin, Dongdong Yu, Qian Yu, and Changhu Wang. 2024. TrackGo: A Flexible and Efficient Method for Controllable Video Generation. *arXiv preprint arXiv:2408.11475* (2024).

Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. 2024. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*. Springer, 145–162.

Yi Zuo, Lingling Li, Licheng Jiao, Fang Liu, Xu Liu, Wenping Ma, Shuyuan Yang, and Yuwei Guo. 2024. Edit-Your-Motion: Space-Time Diffusion Decoupling Learning for Video Motion Editing. *arXiv preprint arXiv:2405.04496* (2024).