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

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

Fig. 1. We encode the motion of a reference video into a novel motion-text embedding using a frozen, pre-trained image-to-video diffusion model. This optimized motion-text embedding can then be applied to different starting images to generate videos with semantically similar motions. The general nature of our motion representation allows for successful motion transfer even when objects are not spatially aligned, across various domains, and for multiple objects. Additionally, our method supports multiple types of motions, including full-body, face, camera, and even hand-crafted motions.
Please refer to https://mkansy.github.io/reenact-anything/ for corresponding videos for all figures of this paper.

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

SIGGRAPH Conference Papers '25, Vancouver, BC, Canada

ACM ISBN 979-8-4007-1540-2/25/08
https://doi.org/10.1145/3721238.3730668

Recent years have seen a tremendous improvement in the quality of video generation and editing approaches. While several techniques focus on editing appearance, few address motion. Current approaches using text, trajectories, or bounding boxes are limited to simple motions, so we specify motions with a single motion reference video instead. We further propose to use a pre-trained image-to-video model rather than a text-to-video model. This approach allows us to preserve the exact appearance and position of a target object or scene and helps disentangle appearance from motion.

Our method, called *motion-textual inversion*, leverages our observation that image-to-video models extract appearance mainly from the (latent) image input, while the text/image embedding injected via cross-attention predominantly controls motion. We thus represent motion using text/image embedding tokens. By operating on an inflated motion-text embedding containing multiple text/image embedding tokens per frame, we achieve a high temporal motion granularity. Once optimized on the motion reference video, this embedding can be applied to various target images to generate videos with semantically similar motions.

Our approach does not require spatial alignment between the motion reference video and target image, generalizes across various domains, and can be applied to various tasks such as full-body and face reenactment, as well as controlling the motion of inanimate objects and the camera. We empirically demonstrate the effectiveness of our method in the semantic video motion transfer task, significantly outperforming existing methods in this context.

Project website: https://mkansy.github.io/reenact-anything/

CCS Concepts: • **Computing methodologies** → **Image processing**; Motion processing; Learning from demonstrations.

Additional Key Words and Phrases: Video motion transfer, reenactment, video editing, image-to-video diffusion model

ACM Reference Format:
Manuel Kansy, Jacek Naruniec, Christopher Schroers, Markus Gross, and Romann M. Weber. 2025. Reenact Anything: Semantic Video Motion Transfer Using Motion-Textual Inversion. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '25), August 10–14, 2025, Vancouver, BC, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3721238.3730668

1 Introduction

The ability to generate and edit videos has rapidly advanced thanks to diffusion models, enabling applications in filmmaking, marketing, and beyond. However, controlling *how* objects move in generated videos—the semantics of motion—remains challenging and largely underexplored. Many existing methods excel at editing *appearance* but struggle to intuitively control *motion*. For example, even state-of-the-art image-to-video models like Stable Video Diffusion [Blattmann et al. 2023a] offer little control over motion, i.e., only by modifying the random seed or adjusting micro-conditioning inputs like frame rate, neither of which is easily interpretable.

To make motion control more intuitive, we propose a new task: semantic video motion transfer from a reference video to a target image. Specifically, we aim to generate a video that replicates the semantic motion of a motion reference video while preserving the appearance and spatial layout of a target image. Crucially, we do not aim to copy pixel-wise trajectories but rather to transfer the meaning of the motion, even when objects are misaligned — for instance, producing a subject performing jumping jacks on the left side of the frame even if the motion reference was centered.

We identify two key challenges for this task: appearance leakage from the motion reference video and object misalignment. To tackle appearance leakage, we employ an image-to-video rather than a text-to-video model and do not fine-tune the model. To the best of our knowledge, we are the first to use an image-to-video model for general motion transfer. To address object misalignments between the motion reference video and the target image, we introduce a novel motion representation that eliminates the need for spatial alignment by not having a spatial dimension in the first place.

Our motion representation is based on our observation that image-to-video models extract the appearance predominantly from the image (latent) input, whereas the text/image embedding injected via cross-attention mostly controls the motion. We therefore propose to represent motion with several text/image embedding tokens, together referred to as *motion-text embedding*, that we optimize on a given motion reference video. Thereby, our inflated motion-text embedding enables us to preserve the timing of the motion video very precisely, which is crucial for applications such as visual dubbing. Our approach, named *motion-textual inversion*, is general in nature and works for various types of motions and objects.[1] Perhaps surprising at first, it turns out that while words are not ideal for describing motions, their embeddings can describe motions exceptionally well. Fig. 1 shows exemplary results of our method, including motion transfers to multiple (misaligned) objects.

To summarize, our contributions are:

(1) We introduce the semantic video motion transfer task in an image-to-video setting.
(2) We observe that text/image embeddings of image-to-video diffusion models store and affect motion and leverage them as a general and compact motion representation.
(3) We propose *motion-textual inversion*, a novel method that optimizes multiple text/image embedding tokens on a motion reference video and transfers the learned motion to target images.
(4) We demonstrate superior performance over existing motion transfer approaches.

2 Related Work

Our goal is to develop a general reenactment method that requires no large-scale domain-specific training. Given the impressive cross-domain translation capabilities of diffusion models [Hertz et al. 2023; Parmar et al. 2023; Tumanyan et al. 2023] and the rise of video generation models [Bar-Tal et al. 2024; Blattmann et al. 2023a; Brooks et al. 2024; Chefer et al. 2025; Kong et al. 2024; Yang et al. 2025], we employ a diffusion-based video model for our general task to capitalize on its broad and general priors. In contrast, the most related non-diffusion methods, JOKR [Mokady et al. 2022] and AnaMoDiff [Tanveer et al. 2024], operate under more constrained conditions, typically requiring a target video, assuming mostly planar 2D motions, and lacking support for natural backgrounds.

In the following sections, we focus on video motion editing approaches based on video diffusion models. In the supplementary material, we discuss additional related works on domain-specific reenactment [Chan et al. 2019; Drobyshev et al. 2022; Guo et al. 2024b; Hsu et al. 2022; Karras et al. 2023; Li et al. 2023; Ma et al. 2024a; Nirkin et al. 2019; Tu et al. 2024a,b; Wang et al. 2024a, 2021; Yang et al. 2020; Zhu et al. 2024; Zuo et al. 2024], keypoint-based motion transfer [Hedlin et al. 2023; Luo et al. 2023; Ni et al. 2023; Siarohin et al. 2019, 2021; Tang et al. 2023; Zhang et al. 2024a, 2023a; Zhao and Zhang 2022], image and video generation [Guo et al. 2024a; Ramesh et al. 2022; Saharia et al. 2022; Wang et al. 2023b], and the inversion-then-generation framework [Ceylan et al. 2023; Garibi et al. 2024; Geyer et al. 2024; Harsha et al. 2024; Liu et al. 2024; Meral et al. 2024; Mokady et al. 2023; Pondaven et al. 2024; Wang et al. 2023a; Xiao et al. 2024; Yang et al. 2023; Zhao et al. 2023].

[1] Independently, a concurrent work, LEAD [Andreou et al. 2024], introduced the term *motion textual inversion* to describe their approach of applying textual inversion [Gal et al. 2023] to a text-to-motion model. While the names are similar, the underlying methods differ significantly.

2.1 Video Motion Editing with Explicit Motions

Existing methods for controlling motion with sparse control signals like text [Dai et al. 2023; Li et al. 2024b; Molad et al. 2023; Yan et al. 2023], boxes [Chen et al. 2024; Jain et al. 2024; Li et al. 2024b; Ma et al. 2024b; Wang et al. 2024e], 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], keypoints [Gu et al. 2024; Niu et al. 2024; Tanveer et al. 2024], or camera motions [Bahmani et al. 2024; Cheong et al. 2024; He et al. 2024; Hou et al. 2024; Hu et al. 2024; Li et al. 2024c; Wang et al. 2024c; Wu et al. 2024c; Xu et al. 2024; Yang et al. 2024; Zheng et al. 2024] are limited to simple motions in most practical scenarios and may require manual prompting. On the other hand, dense motion trajectories [Burgert et al. 2025; Chen et al. 2023b; Gu et al. 2025; Wang et al. 2024d; Zhang et al. 2024b] may leak the motion reference video's spatial structure, thus often failing in unaligned scenarios.

2.2 Video Motion Editing with Implicit Motions

In contrast to the methods discussed above, the methods in this section use less interpretable motion representations. Specifically, fine-tuning approaches encode motions in model weights, and inversion-then-generation approaches extract motions from model features or attention maps.

2.2.1 Fine-Tuning. Approaches based on fine-tuning [Bi et al. 2025; Jeong et al. 2024; Materzyńska et al. 2024; Ren et al. 2024; Wei et al. 2024; Wu et al. 2023; Zhang et al. 2023b; Zhao et al. 2024] involve fine-tuning a model on one or several motion reference videos, similar to DreamBooth [Ruiz et al. 2023]. The methods primarily differ in the parts of the model they fine-tune and the techniques they use, such as LoRA [Hu et al. 2022], to train only the components responsible for motion. However, in practice, they often inadvertently learn the reference video's appearance as well, which can hinder generalization to new target object appearances. We make a similar observation to Wu et al. [2024a], namely that conditioning the diffusion model on the image helps the model concentrate on learning motion.

2.2.2 Inversion-then-Generation. Approaches based on the inversion-then-generation paradigm [Bai et al. 2024; Ling et al. 2024; Yatim et al. 2023] extract model features such as attention maps from the motion reference video (e.g., via DDIM inversion [Song et al. 2020]), which are then incorporated into the diffusion process of the generated video. This helps replicate the reference video's structure in the output. However, these approaches struggle when there are significant differences between the locations and geometries of the reference and target objects, leading to misaligned semantic features being injected or enforced.

2.2.3 With Different Spatial Layout. Most of the one-shot reference-based methods produce videos with motions that are mostly spatially aligned with the motion reference video, i.e., they follow the layout as well as the subject scale and position of the reference video. We thus argue that many of these works [Jeong et al. 2024; Yatim et al. 2023; Zhang et al. 2023b] can be considered as an advanced form of appearance transfer rather than motion transfer. We focus on the general case where layouts may not align, a less explored scenario. Unlike existing methods [Materzyńska et al. 2024; Wei et al. 2024; Wu et al. 2024a; Zhao et al. 2024], which use multiple motion videos to avoid overfitting to a single layout, we transfer motion from a single reference video with precise temporal alignment. Also, instead of relying on text to loosely define the subject's appearance [Li et al. 2024a; Materzyńska et al. 2024; Ren et al. 2024; Wang et al. 2024b], we aim to generate videos that seamlessly continue from a given target image. Concurrently, Wang et al. [2024b] propose an approach that also learns a motion embedding while keeping the model frozen, but they do not incorporate a target image and appear to overfit to the reference video's layout.

3 Method

We propose to transfer the semantic motion of a motion reference video to a given target image by *motion-textual inversion*. We thereby optimize a set of text/image embedding tokens, which we refer to as *motion-text embedding*, for the motion reference video using a pre-trained image-to-video diffusion model.

3.1 Preliminaries

3.1.1 Diffusion. Diffusion models [Ho et al. 2020; Song et al. 2021] consist of two processes. In the *forward process*, Gaussian noise is iteratively added to a clean data sample $\mathbf{x}_0$ until it is approximately pure noise. In the *reverse process*, starting with pure noise $\mathbf{x}_T$, a learnable denoiser $D_{\boldsymbol{\theta}}$ iteratively removes noise to obtain a sample that matches the original data distribution p_{data}. We follow the continuous-time framework [Karras et al. 2022; Song et al. 2021], where the denoiser is trained via *denoising score matching*:

$$\mathbb{E}_{(\mathbf{x}_0,\mathbf{c})\sim p_{\text{data}}(\mathbf{x}_0,\mathbf{c}),(\sigma,\mathbf{n})\sim p(\sigma,\mathbf{n})}[\lambda_\sigma ||D_{\boldsymbol{\theta}}(\mathbf{x}_0+\mathbf{n};\sigma,\mathbf{c})-\mathbf{x}_0||_2^2], \quad (1)$$

where $\mathbf{x}_0$ is a clean data sample and $\mathbf{c}$ an arbitrary conditioning signal from the original data distribution p_{data}; $p(\sigma,\mathbf{n}) = p(\sigma)\mathcal{N}(\mathbf{n};\mathbf{0},\sigma^2)$, where $p(\sigma)$ is a probability distribution over noise levels σ, and $\mathbf{n}$ is noise; and $\lambda_\sigma : \mathbb{R}_+ \to \mathbb{R}_+$ is a weighting function. The denoiser $D_{\boldsymbol{\theta}}$ is parameterized as

$$D_{\boldsymbol{\theta}}(\mathbf{x};\sigma) = c_{\text{skip}}(\sigma)\mathbf{x} + c_{\text{out}}(\sigma)F_{\boldsymbol{\theta}}(c_{\text{in}}(\sigma)\mathbf{x};c_{\text{noise}}(\sigma)), \quad (2)$$

where $F_{\boldsymbol{\theta}}$ is the neural network to be trained; $c_{\text{skip}}(\sigma)$ modulates the skip connection; $c_{\text{out}}(\sigma)$ and $c_{\text{in}}(\sigma)$ scale the output and input magnitudes respectively; and $c_{\text{noise}}(\sigma)$ maps noise level σ into a conditioning input for $F_{\boldsymbol{\theta}}$. For more details, please refer to EDM [Karras et al. 2022].

3.1.2 Latent Diffusion. Latent diffusion models [Rombach et al. 2022] operate in the latent space rather than in pixel space to reduce computation and thus enable higher resolutions. First, an encoder $\mathcal{E}$ produces a compressed latent $\mathbf{z} = \mathcal{E}(\mathbf{x})$. Then, we perform the diffusion process over $\mathbf{z}$. Lastly, a decoder $\mathcal{D}$ reconstructs the latent features back into pixel space.[2]

3.1.3 Baseline. Stable Video Diffusion (SVD) [Blattmann et al. 2023a] is a video latent diffusion model trained in three stages: 1. A text-to-image model [Rombach et al. 2022] is trained or fine-tuned on (image, text) pairs. 2. The diffusion model is inflated by inserting temporal convolution and attention layers following Blattmann

[2]To maintain consistency in notation, we use $\mathbf{x}$ for the diagrams and method description, even though the diffusion process actually occurs in latent space.

Fig. 2. Observation 1. In image-to-video models, the image input primarily dictates the appearance of the generated videos. For example, I2VGen-XL [Zhang et al. 2023c] generates a video of a predominantly white horse from a white horse image, even when the input text specifies the horse's color as "pink."

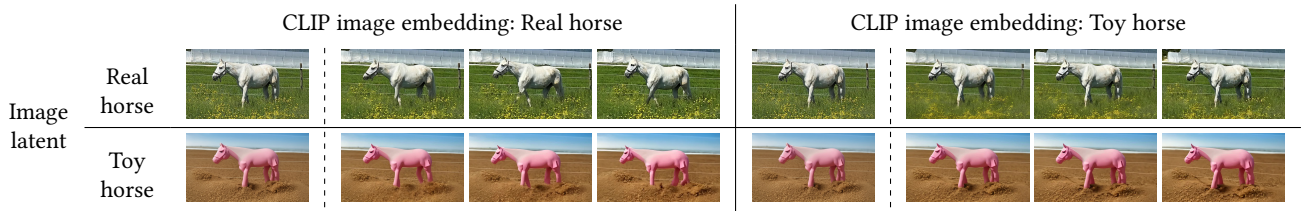


Fig. 3. Observation 2. In image-to-video models, text/image embeddings significantly influence the generated motions. Swapping the CLIP [Radford et al. 2021] image embeddings of a real horse and a toy horse in Stable Video Diffusion [Blattmann et al. 2023a] results in a swap of the motions in the output videos. This suggests that the real horse's embedding encodes a walking motion, while the toy horse's embedding encodes camera motion without object movement.

et al. [2023b] and then trained on (video, text) pairs. 3. The diffusion model is refined on a smaller subset of high-quality videos with exact model adaptations and inputs depending on the task (text-to-video, image-to-video, frame interpolation, multi-view generation). For image-to-video generation, the task is to produce a video given its starting frame. The starting frame is supplied to the model in two places: as a CLIP [Radford et al. 2021] image embedding via cross-attention (replacing the CLIP text embedding from the text-to-video pre-training) and as a latent repeated across frames and concatenated channel-wise to the video input. Additionally, the model is micro-conditioned on the frame rate, motion amount, and strength of the noise augmentation (applied to first frame latent).

3.2 Motivation

Transferring the motion of a reference video to a given target poses two key challenges, which our design solves quite naturally.

3.2.1 Challenge 1: Appearance Leakage. Fine-tuning a text-to-video model on a single reference video to learn its motion risks overfitting to its appearance, hindering the generation of correct target appearances during inference. We demonstrate that using a frozen image-to-video model can preserve the target appearance without any of the special mechanisms from the literature.

By design, image-to-video models generate videos from a starting frame, naturally preserving the input appearance. We observe that image-to-video models primarily derive the appearance from the image (latent) input, even with an additional text input, as shown in Fig. 2. This is likely because the model can directly copy (latent) pixels from the first frame instead of hallucinating them from the sparse text input. This strong reliance on the image input reduces the chance of the reference video's appearance leaking through. To further minimize the risk of appearance leakage, we keep the model's weights frozen, so they cannot possibly store the reference video appearance. This also helps retain the rich video understanding and generalization capabilities of the pre-trained model.

3.2.2 Challenge 2: Handling Object Misalignment. Our goal is to generate videos where subjects perform the same semantic actions, even if they are in different spatial locations or orientations. Handling misaligned objects is especially important when using image-to-video models because the subject's position is determined by the input image, which typically does not match the position in the motion reference video.

As discussed in Section 2.2.2, existing methods using the inversion-then-generation framework inject features from the motion reference video into the generated video, making it closely follow the reference structure. Arguably, these methods do not copy the motion at its origin but rather the *per-frame structure* that results from a motion (e.g., rough object positions). For the general, unaligned case, these features would first need to be aligned spatially to avoid injecting the structure in the wrong place. This alignment is challenging since the final positions in the generated video are unknown during the diffusion process as they depend on the motion.

We forgo the alignment problem by representing motions with text or image embedding tokens that do not have a spatial dimension in the first place. Our novel motion representation was motivated by the observation shown in Fig. 3. While SVD generated walking motions for an image of a real horse, it generated no object but mostly camera motion for an image of a pink toy horse, perhaps because the model learned that toys do not move.[3] Recall that SVD has the first frame as input in two places: as image latent and as CLIP [Radford et al. 2021] image embedding. When using the image latent of the real horse but the CLIP embedding of the toy horse, the horse in the generated video does not move. Inversely, the toy horse starts walking when using the CLIP embedding of the real horse, implying that the CLIP embedding affects the motion. We believe that these embeddings are *not just affecting* the motion but are actually the main *origin* of the motion.

[3]Image was generated using the method by Tumanyan et al. [2023].

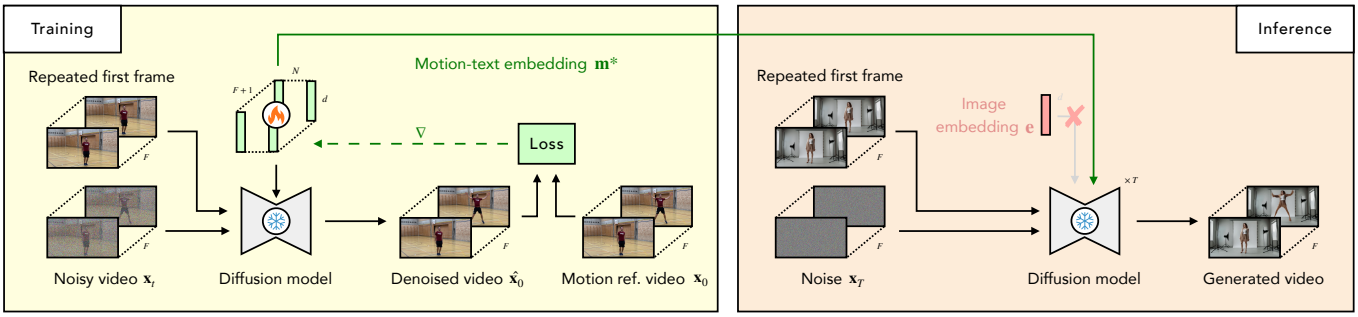


Fig. 4. Method overview. The baseline image-to-video diffusion model, Stable Video Diffusion [Blattmann et al. 2023a] in our case, inputs the first frame in two places: as image (latent) concatenated with the noisy video and as image embedding (some other image-to-video diffusion models may input text embeddings here instead). We propose to replace the image embedding $\mathbf{e}$ (shown in red in the inference block) with a learned motion-text embedding $\mathbf{m}^*$ (green). The motion-text embedding is optimized directly with a regular diffusion model loss on one given motion reference video $\mathbf{x}_0$ while keeping the diffusion model frozen. For best results, the motion-text embedding is inflated prior to optimization to $(F+1) \times N$ tokens, where F is the number of frames and N is a hyperparameter, while keeping the embedding dimension d the same to stay compatible with the pre-trained diffusion model. Note that the diffusion process operates in latent space in practice, and other conditionings and model parameterizations [Karras et al. 2022] are omitted for clarity.

Our intuition for why the text/image embeddings determine the motion (which may be surprising at first) is as follows: Videos can be divided into appearance and motion. Appearance is tied to the spatial arrangement of pixels, making it easier to extract it from spatial inputs like image latents. Motion depends on how pixels change over time, requiring a more global, semantic understanding. Thus, it is more effective to modify motion using image embeddings, which contain more semantic information, have no spatial dimension, and are injected in multiple places of the model. Furthermore, SVD was initially trained as a text-to-video model, with CLIP text embeddings describing motions like "standing," "walking," or "running," incentivizing the model to control motion through cross-attention inputs to effectively denoise training videos.

3.3 Motion-Textual Inversion

While using embeddings from different images can alter the generated motion, it does not transfer the motion robustly. Moreover, selecting a specific frame to define a desired motion is difficult since motion is rarely captured by a single frame. To address this, we propose optimizing the embedding based on a given motion reference video, which bears some resemblance to textual inversion [Gal et al. 2023]. In analogy to textual inversion, we name our method *motion-textual inversion*.[4] Note, however, that our method has a completely different goal: using embeddings to encode video motion rather than image appearance.

Fig. 4 shows a high-level overview of our method. Given a single motion reference video $\mathbf{x}_0$ containing F frames, we optimize the motion-text embedding $\mathbf{m}$ directly by minimizing the diffusion model loss from Equation 1, keeping the diffusion model frozen:

$$\mathbf{m}^* = \arg\min_{\mathbf{m}} \mathbb{E}_{(\mathbf{x}_0,\mathbf{c})\sim p_{\text{data}}(\mathbf{x}_0,\mathbf{c}),(\sigma,\mathbf{n})\sim p(\sigma,\mathbf{n})} [\lambda_\sigma ||D_{\boldsymbol{\theta}}(\mathbf{x}_0 + \mathbf{n}; \sigma, \mathbf{m}, \mathbf{c}) - \mathbf{x}_0||_2^2], \tag{3}$$

[4]In our implementation, it is actually an image embedding, but we refer to it as "motion-*textual* inversion" since SVD's image and text embeddings share the same CLIP space, and other I2V methods use text embeddings instead. Also, it feels more intuitive to represent motions as text rather than an image.

where $\mathbf{c}$ encompasses all remaining conditionings of SVD (e.g., first frame latent, time/noise step, and micro-conditionings). All other symbols are defined in Equations 1 and 2.

The optimized motion-text embedding can be visualized with an unconditional appearance as seen in Fig. 1 and further described in the supplementary material.

3.4 Motion-Text Embedding and Cross-Attention Inflation

Cross-attention allows the model to dynamically attend to different tokens (~ words in text-to-image and text-to-video) depending on the current features or context. It is computed as follows:

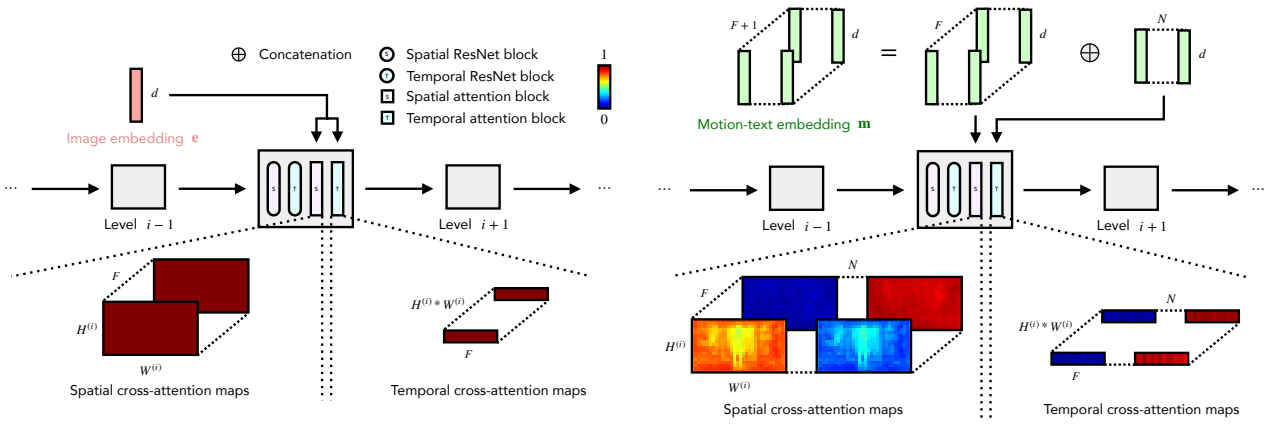
$$\begin{aligned} \text{Attention}(Q,K,V) = MV = \text{softmax}(\frac{QK^T}{\sqrt{d_a}})V, \\ Q = \varphi_i(\mathbf{z}_t)W_{Q,i},\ K = \mathbf{m}W_{K,i},\ V = \mathbf{m}W_{V,i}, \end{aligned} \tag{4}$$

where Q, K, V are the queries, keys, and values respectively; M is the attention map; d_a is the dimension used in the attention operation; $\varphi_i(\mathbf{z}_t)$ is an intermediate representation of the level i features with C_i channels; $\mathbf{m}$ is the motion-text embedding (or text/image embedding $\mathbf{e}$ in case of baseline SVD) with embedding dimension d; and $W_{Q,i} \in \mathbb{R}^{C_i \times d_a}$, $W_{K,i} \in \mathbb{R}^{d \times d_a}$, and $W_{V,i} \in \mathbb{R}^{d \times d_a}$ are learned weight matrices for queries, keys, and values respectively.

SVD's image embedding only has one token. This leads to a degenerate cross-attention where all entries of the attention map M are 1, as shown in Fig. 5a. The model thus attends 100% to that single token and applies its value to all spatial and temporal locations.

3.4.1 Multiple Tokens. To enable richer motion control, we replace the single token with N tokens, recovering the scenario from the text-to-image or text-to-video pre-training. This allows the model to dynamically attend to different tokens depending on the features, e.g., using different values for the background and foreground as seen in the spatial cross-attention maps in Fig. 5b.

3.4.2 Different Tokens per Frame. For spatial cross-attention, SVD broadcasts the image embedding *across all frames*. Instead, we use a different set of tokens per frame, i.e., $F \times N$ tokens, to obtain a



(a) Default SVD: Since the image embedding **e** has only one token, every spatial and temporal location attends 100% to that single token. The cross-attention operation thus degenerates to a simple addition of a single broadcasted vector to the feature tensor.

(b) Inflated SVD (Ours): By introducing more tokens in the token dimension (N), every spatial and temporal location can dynamically attend to different tokens, e.g., different tokens for the foreground vs. background. For the spatial cross-attention, we use different tokens per frame, resulting in different keys and values per frame. This enables a higher temporal granularity of the motion.

Fig. 5. High-level visualization of our motion-text embedding and cross-attention inflation. The SVD [Blattmann et al. 2023a] UNet is composed of several levels of blocks, shown in gray, that have similar structure. We visualize the sub-blocks of level i and their cross-attention maps in more detail. Our inflated motion-text embedding produces more meaningful cross-attention maps, resulting in improved motion learning. The cross-attention maps were extracted from the example of the woman doing jumping jacks in Fig. 4.

higher temporal motion granularity.[5] This yields distinct keys and values for each frame: different keys enable attention to different spatial regions over time (e.g., arm vs. leg), while different values allow frame-specific feature modifications (e.g., shifting pixels in different directions). This is visualized in Fig. 5b, where the spatial cross-attention maps differ greatly between frames because they use different tokens.

For temporal cross-attention, SVD broadcasts the image embedding *across all spatial locations*. Inflating this analogously to the spatial case would require learning distinct tokens per spatial location, which is nontrivial due to resolution- and level-dependent spatial dimensions and may cause alignment issues (see Section 3.2.2). Furthermore, temporal cross-attention impacted motion less than spatial cross-attention empirically. We thus keep N tokens for the temporal motion-text embedding but learn them independently from the $F \times N$ tokens of the spatial motion-text embedding, yielding a total of $(F+1) \times N$ tokens per reference video. See the supplementary material for an intuitive analogy and detailed tensor shapes.

4 Experiments

4.1 Implementation Details

Our method builds on the 14-frame version of Stable Video Diffusion (SVD) [Blattmann et al. 2023a; von Platen et al. 2022] but can be applied to other image-to-video models with a text/image embedding input. Per default, we use $N = 5$ different tokens for each of the $F = 14$ frames, so a total of $(14 + 1) \times 5 = 75$ tokens for the motion-text embedding. We further use the Adam optimizer [Kingma and Ba 2015] and SVD's default guidance scale [Ho and Salimans 2021] (except for motion visualization). For our qualitative results, we use internal data sets and target images generated with SDXL [Podell et al. 2024]. See the supplementary material for further details.

4.2 Compared Methods

As baseline, we use SVD [Blattmann et al. 2023a] without adaptations. Since it lacks motion conditioning, it rarely follows the correct motion but serves as a reference for typical SVD output quality and dynamics. Our method is the first to tackle general motion transfer in the image-to-video setting. As no direct competitors exist, we apply the most closely related approaches from literature to our task and show issues inherent to the whole class of methodology. Specifically, we compare to VideoComposer [Wang et al. 2024d], an image-to-video method with an explicit, dense motion representation (motion vectors); the image-to-video setting of MotionClone [Ling et al. 2024] which has an implicit motion representation (sparse temporal attention weights); and MotionDirector [Zhao et al. 2024], a text-to-video method with an implicit motion representation (learned model weights). We only compare to general methods that place no constraints on motion types and target images. Domain-specific methods rely on strong assumptions and typically fail when these are not met. For example, a face reenactment method cannot control transfer the motion of a horse to a boat. As domain-specific methods address a different task, a fair comparison is not possible. See the supplementary material for further details.

[5] Note that we always use the same F frames of the motion reference video when optimizing the motion-text embedding.

Table 1. Quantitative evaluation. 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 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	**0.843**	0.850	**1.296**	3%	5%	0.370	4.211	2.822
VideoComposer	0.719	0.857	3.785	44%	62%	0.497	3.030	3.552
MotionClone	0.637	**0.885**	4.585	37%	62%	0.523	3.137	4.200
MotionDirector	0.750	0.763	3.522	31%	58%	0.523	2.900	3.059
Ours	0.779	0.884	1.811	**54%**	**76%**	**0.696**	**1.722**	**1.367**

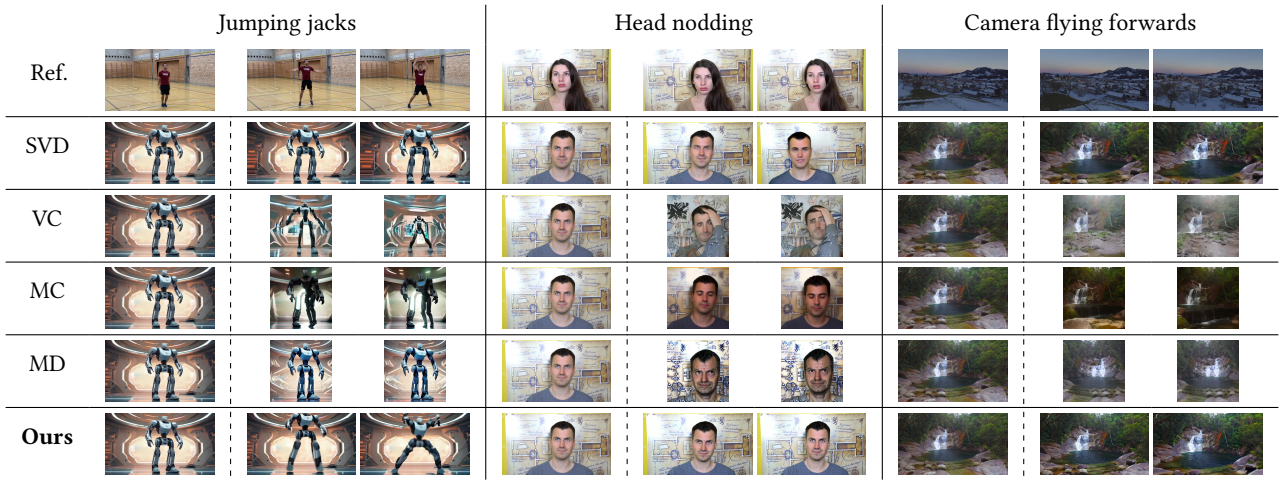


Fig. 6. Qualitative evaluation. We compare our method to SVD = Stable Video Diffusion [Blattmann et al. 2023a] (baseline, no motion input), VC = VideoComposer [Wang et al. 2024d], MC = MotionClone [Ling et al. 2024], and MD = MotionDirector [Zhao et al. 2024] for three different motions and target images: full-body reenactment, face reenactment, and camera motion.

4.3 Qualitative Evaluation

Fig. 6 shows motion transfer results for three motions. As expected, the SVD baseline typically produces mismatched motions. For certain videos, like the face video, SVD produces significant artifacts and alters the subject identity. Due to its dense motion input, VideoComposer replicates motion in the spatial location of the reference video, leading to incorrect semantic motion and artifacts when structures misalign. MotionClone faces similar issues but handles minor structural differences better in the nodding example and has more high-level artifacts due to its higher-level motion representation. Since MotionDirector is based on a text-to-video model, it must learn the appearance and thus cannot continue naturally from the target image by design. Additionally, the motion is only transferred correctly for the head nodding example. Our method is the only one that preserves the input image's appearance and layout while successfully transferring the semantic motion of the video. The supplementary material provides additional qualitative comparisons, including an in-depth comparison with SVD and its embeddings.

4.4 Quantitative Evaluation and User Study

We evaluate our method on the Something-Something V2 data set [Goyal et al. 2017], selecting 10 classes from the validation set (5 with camera movements, 5 with object movements). For each class, one video serves as the motion reference, and 10 other videos' first frames act as target images, totaling 100 generated videos per method. This data set provides a challenging benchmark, as videos within each class have the same semantic action but vastly different spatial layouts. See the supplementary material for further details.

For image appearance preservation, we calculate the mean cosine similarity between the CLIP [Radford et al. 2021] image embeddings of the target image and the generated video, where **CLIP-Avg** is the average across all frames and **CLIP-1st** refers to the first frame. For video motion fidelity, we avoid metrics like optical flow or Motion-Fidelity-Score [Yatim et al. 2023], which emphasize spatial over semantic motion. Instead, similar to MoTrans [Li et al. 2024a], we use an action recognition network [Tong et al. 2022] trained on Something-Something V2 (174 classes). **Acc-Top-1** is the percentage of videos correctly classified, and **Acc-Top-5** the percentage with the correct class in the top 5 predictions. **Cos-Sim** is the cosine similarity between the logits of the generated and reference videos.

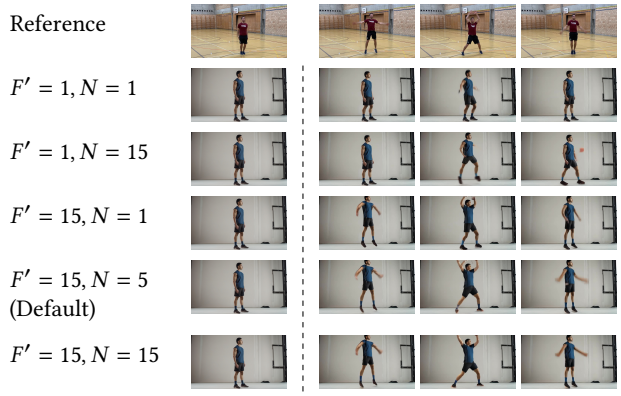


Fig. 7. Ablation. Our proposed motion-text embedding inflation is crucial for successful motion transfer. While adding more tokens (increasing N) improves the results already, the biggest gain comes from having different tokens for each frame (where $F' = F + 1 = 15$).

Fig. 8. Failure cases. Our method is limited by the priors and quality of the pre-trained image-to-video model, which may lead to artifacts (e.g., identity changes as head moves in first example). Furthermore, there may be some structure leakage in some cases, leading to certain characteristics from the motion reference video being visible (e.g., human-like legs on a kangaroo in second example). Lastly, our method struggles to transfer spatially fine-grained motion at times (e.g., typing motion not transferred to dinosaurs in third example).

The results in Table 1 reflect our qualitative findings. SVD preserves the target image but fails to capture the motion. MotionDirector struggles with image preservation in the first frame, whereas image-to-video methods generally excel in this aspect by design. For motion fidelity, all competitor methods (except SVD) perform similarly, while our method outperforms them significantly.

Additionally, we conducted a user study with 27 users on a random subset of the evaluation data (one target image per motion video). For each of the 10 video sets, users ranked the methods from best (1) to worst (5) based on (a) **image appearance preservation**, (b) **video motion fidelity**, and (c) **overall** task fulfillment. The rankings align with the metrics but show an even stronger preference for our method. As seen in Table 1, our method has the best average rank for motion fidelity and overall task fulfillment, voted best 75% and 78% of times respectively. It also performs well on appearance preservation, landing closely behind SVD. Note that this metric is biased towards methods that produce little motion, so it should only be regarded in combination with the motion fidelity.

4.5 Ablation Study

Our motion-text embedding inflation is key to high-quality motion transfer. Fig. 7 shows different embedding configurations. A single token captures only limited motion. Adding more tokens shared across frames helps, but the crucial factor is *having different tokens per frame*. Rows 2 and 3 both use 15 tokens, but allowing the embedding to adapt frame-wise performs significantly better, especially for complex motions. Increasing tokens per frame further improves results slightly before saturating, so we default to $N = 5$. The supplementary material provides two additional qualitative examples for this ablation as well as quantitative results when using the same protocol as for the above state-of-the-art comparison.

4.6 Results

Our motion representation is highly versatile, enabling motion transfer across diverse objects and motions, as demonstrated in Fig. 1 and Fig. 9. Notably, we do not require a spatial alignment, as seen in row 6 (right) of Fig. 9, where the camera follows the moving camper van similar to how it follows the car in the fifth row, despite their misalignment. Our method also applies the motion to all semantically reasonable objects simultaneously "for free." It even supports simple hand-crafted motions, enabling artists to sketch motions (e.g., stick figures) and apply them to complex scenes. For more results, including joint subject and camera motion, extreme cross-domain transfers, and applying the same motion to multiple target images, please refer to the supplementary material.

4.7 Limitations and Future Work

Fig. 8 shows typical failure cases of our method. Since we do not fine-tune the model, our method inherits the priors and quality of our pre-trained image-to-video model. We observed that the SVD baseline often struggles with object motions, as can be seen in the head example in Fig. 6, where the appearance changes throughout the video. Our method's results have similar issues: in the first example of Fig. 8, the identity of the target person changes when he moves his head to the side. We believe our motion-text embedding does not exacerbate these issues or temporal inconsistencies, as it primarily instructs the model on the desired motion without altering the rest of the model. Often, it seems that the model attempts to produce the desired motion, but its priors are insufficient to generate a satisfactory result. SVD also does not seem to be able to handle some combinations of motions and given input images, likely because they fall outside of the range of the training data set. When the domain gap between motion reference video and target image is too large, our method may leak the structure of the motion reference video into the generated video. In the second example of Fig. 8, when applying a laid-back walking style to a kangaroo, the kangaroo starts walking, but its feet and overall structure become more human-like. Lastly, we found that some motions are

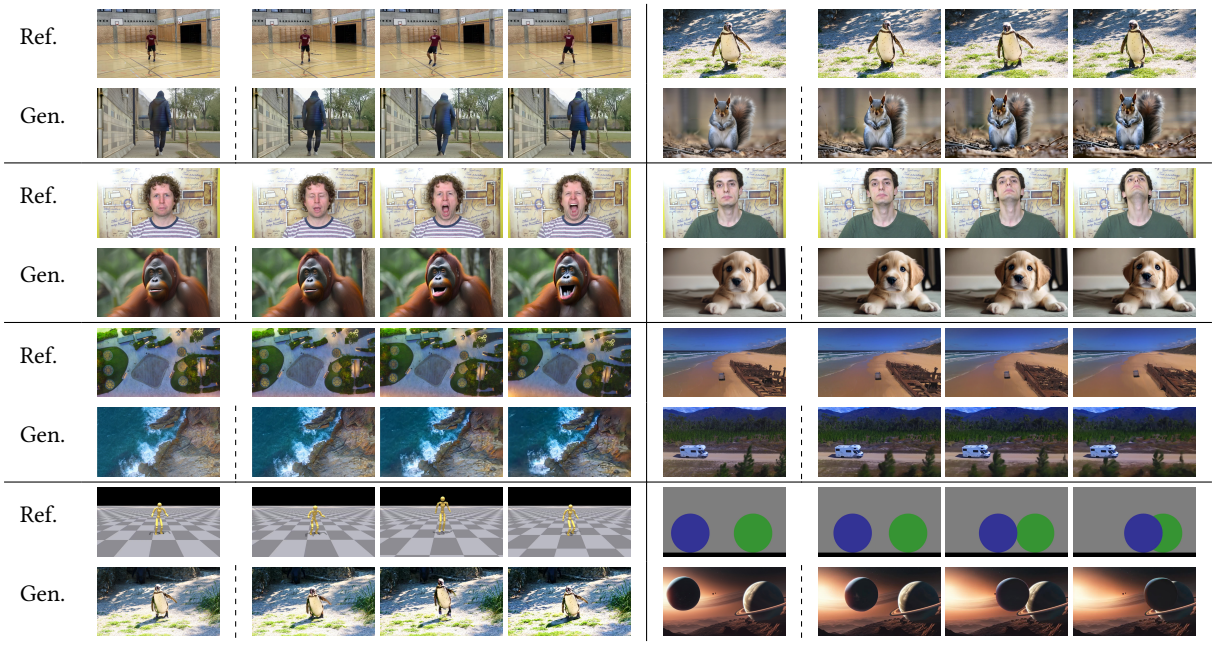


Fig. 9. Results. Our method can successfully transfer semantic video motion across a wide number of domains and motions.

not transferred or to a smaller extent. This is especially visible if a video has multiple motions, where the more fine-grained motion is sometimes not transferred. In the third example of Fig. 8, the person pretends to squat down and type on a keyboard. The dinosaurs in the generated video do squat down, but their hands do not move. We hypothesize that fine-grained motions are also a general limitation of SVD. Overall, we expect better results of our method as image-to-video models improve. In the supplementary material, we analyze our method's failure rate in more detail.

An important practical consideration is that the target image must be temporally aligned with the first frame of the motion reference video, as it serves as the starting frame. This is not a limitation of our method specifically, but rather a consequence of the task formulation. Alternatively, one could treat the image as an appearance reference (as in Animate Anyone [Hu 2024]) and adapt or fine-tune the model accordingly.

While more accessible than methods requiring extensive training or fine-tuning, our approach requires an optimization procedure that takes about one hour per motion on an A100 (80 GB) GPU. We have also run it on 48 GB GPUs, albeit with slightly longer runtimes. We encourage future work on reducing the per-motion optimization time, or eliminating it entirely by learning to predict motion-text embeddings directly from motion reference videos, scaling our method to longer videos, as well as adapting it to newer architectures based on diffusion transformers [Peebles and Xie 2023].

5 Conclusion

We introduce the general task of transferring the semantic motion of a reference video to any target image. We observe and exploit inherent advantages of image-to-video over text-to-video models for this task and find that text/image embedding tokens are well-suited as a motion representation. Specifically, our method, *motion-textual inversion*, optimizes an inflated version of the text/image embedding for a given motion reference video. Due to its general nature, this motion can then be applied to a wide number of objects and domains. Our method thus enables completely novel applications and takes a significant step towards being able to reenact anything.

Acknowledgments

We would like to thank Michael Bernasconi, Dominik Borer, Jakob Buhmann, and Daniela Kansy for providing motion videos as well as all participants featured in our internal data sets. We also want to give special thanks to Bastian Amrhein, Michael Bernasconi, Vukasin Bozic, Karlis Briedis, Pascal Chang, Guilherme Haetinger, Christopher Otto, Lucas Relic, Seyedmorteza Sadat, and Agon Serifi for their valuable and insightful discussions throughout the project.

References

Nefeli Andreou, Xi Wang, Victoria Fernández Abrevaya, Marie-Paule Cani, Yiorgos Chrysanthou, and Vicky Kalogeiton. 2024. Lead: Latent realignment for human motion diffusion. In *Computer Graphics Forum*. Wiley Online Library, e70093.

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.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations*.

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, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

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.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. 2022. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems* 35 (2022), 26565–26577.

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).

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

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).