VIDMP3: Video Editing by Representing Motion with Pose and Position Priors

Sandeep Mishra
University of Texas at Austin
sandy.mishra@utexas.edu

Oindrila Saha
University of Massachusetts Amherst
osaha@umass.edu

Alan C. Bovik
University of Texas at Austin
bovik@ece.utexas.edu

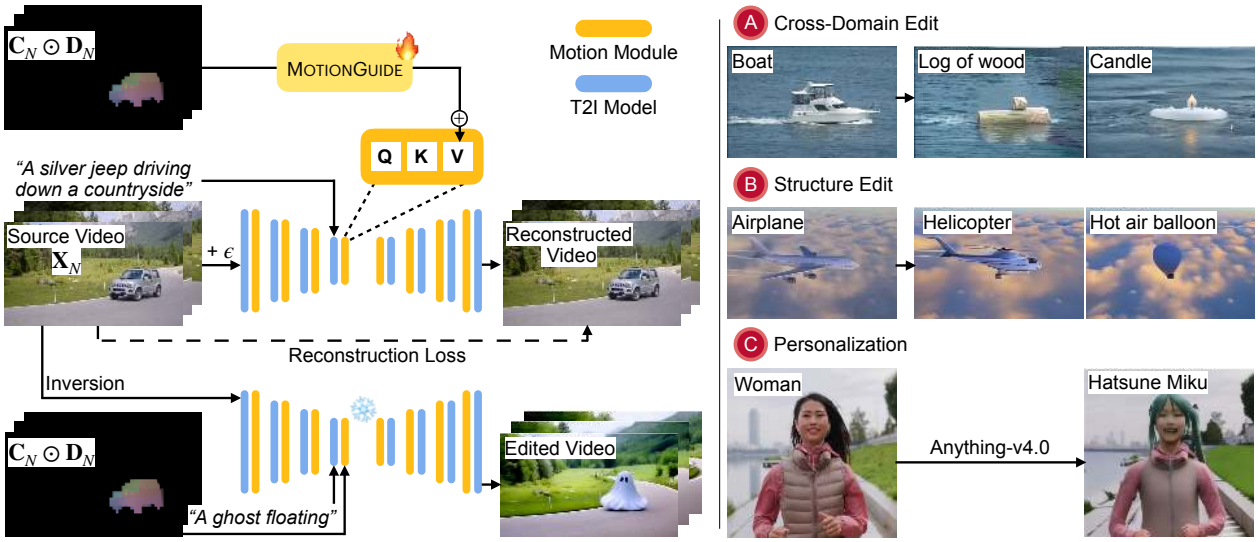

Figure 1. **VIDMP3.** We present a novel video editing technique that can perform challenging video editing tasks guided by pose and position priors. We introduce a MOTIONGUIDE module that learns a generalized motion representation from correspondence and depth maps. We inject the features of this module to the "Value"s of the temporal self-attention layer of a T2V initialized with a T2I model. During inference, we use the correspondence and depth maps of the source video to generate a novel motion-preserved video. VIDMP3 enables the generation of challenging edits, including Ⓐ Cross-Domain editing, where objects with vastly different semantic meanings can be generated, Ⓑ Structure editing, where structure of the object can be changed significantly, and Ⓒ adaptation to various T2V editing tasks such as personalized editing.

Abstract

Motion-preserved video editing is crucial for creators, particularly in scenarios that demand flexibility in both the structure and semantics of swapped objects. Despite its potential, this area remains underexplored. Existing diffusion-based editing methods excel in structure-preserving tasks, using dense guidance signals to ensure content integrity. While some recent methods attempt to address structure-variable editing, they often suffer from issues such as temporal inconsistency, subject identity drift, and the need for human intervention. To address these challenges, we introduce **VIDMP3***, a novel approach that leverages pose and position priors to learn a generalized motion representation from source videos. Our method enables the generation of new videos that maintain the original motion while allowing for structural and semantic flexibility. Both qualitative and quantitative evaluations demonstrate the superiority of our approach over existing methods.*

1. Introduction

The strong generation capabilities of text-to-image (T2I) diffusion models have encouraged the adoption of these models for video generation and editing tasks, owing to the simple architectural changes required over T2I models to enable them to generate videos. Inclusion of temporal self-attention layers and inflating 2D convolutions to pseudo 3D convolutions facilitates the generation of videos conditioned on text. While some approaches train text-to-video (T2V) models on large-scale text-video paired

datasets [4, 15, 16, 37, 52], others explore a more data-efficient technique. These methods [13, 38, 42, 48] train a T2V model on a single video and use the learned priors to generate novel videos using edited text prompts. T2I models have also been used for zero-shot video editing [6, 8, 11, 27, 34] by utilizing structure from a specific source video.

Generative video editing is a task of remarkable interest to creators which enables them to create novel videos which can borrow information from a captured real video. One of the most important and under-explored sub-areas is where only motion is preserved from a source video and mimicked to generated a new video. This is the most general use-case of generative video editing, whereby the motion of the subject in the source video is preserved but structure, appearance, and semantics remain modifiable. Apart from the clear benefits of reducing costs and time for video creators, this serves an important case where a creator would want to imitate the motion of a real subject and transfer it to subjects that might be hard to capture following that specific motion e.g., imaginary concepts following the motion in a real video.

In a data efficient setting where we want to use only a single source video to generate an edited novel video, changing the structure and domain of the subject has been a challenging task. Zero-shot video editing techniques heavily rely on the structure of the source video, and are thus unable to deviate much from the source concept. One-shot tuning techniques have shown sufficient promise, but struggle with either shape leakage, quality issues, or fail in cases of cross-domain editing. This can be attributed to unconstrained optimization over the source video [42] or too sparse external control [13].

We embark on learning a generalized motion representation that distentangles spatial properties of subjects from their motion. Motion of subjects is perceived by humans as the combination of their position in a 3D space and their internal pose. Thus, we choose to inject an external representation learned from pose and position priors to guide the T2I diffusion model. We hypothesize that motion can be represented as a combination of spatial correspondence maps, depth maps and 2D positional encodings. The correspondence maps provide signals for the internal pose variation of a subject over video frames, while the depth maps and positional encoding signify the 3D positions of the subject in each frame. We introduce a novel MOTIONGUIDE module which utilizes these maps to learn a generalized representation of motion. First, we show a proof of concept where MOTIONGUIDE can be used to learn the 3D trajectory and rotations of a simple moving cube. We show that the learned module is invariant to shape changes of the object but sensitive to motion changes. This shows that this module can be effectively used to induce motion-preservation with variations in shape when appropriately injected into a T2V diffusion model initialized with a T2I model. We present VIDMP3 where we inject the spatially pooled features of MOTIONGUIDE into the "Value"s of the temporal self-attention layers of the T2V model. Essentially, this allows the model to understand added context in frame-to-frame correspondence, thus boosting temporal consistency. We show that VIDMP3 robustly edits subjects with significant structure and semantic shift from the subject in the source video. We also scale our method to Stable-Diffusion-XL [32], which has not been explored previously for video editing. We show that we are able to generate more diverse concepts with VIDMP3 SD-XL. In summary, our contributions are as follows:

- A MOTIONGUIDE module that learns generalized motion representations from pose and position priors
- VIDMP3, which utilizes the MOTIONGUIDE module to inject external guidance to the "Value"s of the temporal self-attention module
- Adaptation to various T2I diffusion models including scaling to SD-XL.

2. Related Work

Diffusion models have been extensively explored for video editing due to their strong generation capability and ability to conform to various kinds of conditions. Previous video editing techniques can be classified into two general categories: 1) Structure-preserved Video Editing, and 2) Motion-preserved Video Editing. We discuss prior work in these two domains in detail below.

2.1. Structure-preserved Video Editing

These techniques aim to edit the video while preserving structural information from the original video by relying on various cues such as depth, edge, optical flow, or attention map information. Gen-1 [9], Ground-a-video [18], and RAVE [20] utilize depth maps for guidance, while CCEdit [10], ControlVideo [50], and MAsk-INT [28] extend to the use of various controls including depth, boundary, and line drawing. MoCa [44], Rerender A Video [46], and FlowVid [25] use optical flow as guidance. VideoP2P [27], FateZero [34], Vid2Vid-Zero [41], and Edit-A-Video [36] inject attention map information from the original video while denoising the edited video. Token-Flow [11], COVE [40] and DreamMotion [19] use dense spatial correspondences among frames to ensure consistency. VidTome [24] develops a method that uses any of the above discussed types of guidance techniques. Codef [30], VidEdit [8], and StableVideo [6] learn a canonical representation of the video. Editing this representation allows high temporal consistency, but restricts changes in low-level features. In contrast to these methods, VIDMP3 allows significant structural and semantic changes in the subject of the

given source video.

2.2. Motion-preserved Video Editing

These methods aim to extract the motion from the source video while allowing significant structural changes in the edited video generated with the same motion.

One-shot tuning. Tune-a-video [42] attaches a motion module to a pre-trained T2I model, and introduces sparse causal self-attention which uses features from other frames to compute self-attention on each frame. Tune-A-Video overfits the motion module to a single video, which is then used to generate novel videos at test-time. We find that Tune-a-Video suffers from severe structure leakage and temporal inconsistency, due to unconstrained training of the motion module on the input video.

VideoSwap [13] alleviates structure leakage by injecting keypoint correspondence information and keeping the motion module frozen. However, VideoSwap requires human effort in selecting or editing the keypoint positions. For cases which require significant size changes, VideoSwap creates a Layered Neural Atlas [22] of the video, in which the user is required to make desired edits. Training this LNA is significantly time consuming. Additionally, as a result of using keypoint correspondence, VideoSwap is ineffective at swapping semantically different objects. By contrast, VIDMP3 is able to swap objects with considerable structure and semantic variation, due to injecting a generalized representation of external pose and position guidance. Most importantly, VIDMP3 relies neither on human effort nor the time-intensive LNA creation process.

SAVE [38] aims to disentangle the structure and motion of a subject by using a motion prompt that focuses on moving areas, but suffers from temporal inconsistencies due to leakage in areas surrounding the moving object, as evidenced in their results. CAMEL [48] injects motion prompts into the temporal attention module, which is then learned from the video. By contrast, our method uses external pose and position guidance to learn a more consistent representation of motion.

Emu-Video [12] attaches an image editing and video generation adapter over a pre-trained T2I model, which is then tuned on a dataset of several videos. VIDMP3 instead extracts various kinds of information from a single video to generate a novel edited video.

Pose-guided video editing. 2D/3D pose-guided video editing has been explored specifically for humans and human-like entities in Follow-Your-Pose [29], DreamPose [21], DeCo [51], MagicPose [7], MagicAnimate [43], AnimateAnyone [17], EVA [47], and DynVideo-E [26]. VIDMP3 instead explores pose-guided editing in a more general context with pose being represented using correspondence maps. This representation allows us to generate subjects which are highly semantically and structurally different from the subject in the source video, while accurately following the motion of the source video.

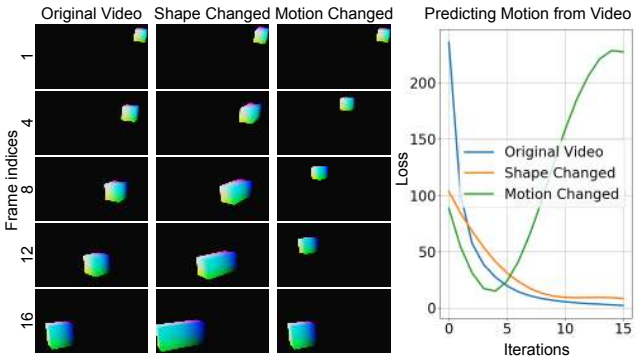

Figure 2. **Toy experiment on learning shape-invariant motion.** We trained our MOTIONGUIDE module on the original video and tested it on videos with 1) shape changes, and 2) motion changes. We show the frames for each video to the left. From the graph at the right, it may be observed that the MOTIONGUIDE module is invariant to shape change but sensitive to motion change.

Propagation from first frame editing. AnyV2V [23] and I2VEdit [31] use a separate model for editing the first frame of the video and then propagate the edit to the other frames. While these methods can significantly change the structure of the subject, they are limited by the image-editing technique they utilize. AnyV2V suffers from severe temporal inconsistencies when modeling videos with significant motion (see Appendix). VIDMP3 instead learns the motion representation from the source video and jointly models it across frames.

3. Method

The motion of any object can be represented as a combination of pose and position in 3D space. Given a video $\mathbf{X}_N = [x_1, x_2 \dots x_N]$ of N frames, we wish to learn only the motion of the subject in the video. We want to build a **generalized representation of motion** using the 3D pose and position of an object. This representation enables us to swap objects with significantly different shapes or semantics. We hypothesize that motion can be extracted only using the dense correspondences within frames $\mathbf{C}_N$ and the depth maps per frame $\mathbf{D}_N$, without using the frames of the video $\mathbf{X}_N$. $\mathbf{C}_N$ is useful for representing the 2D position and pose of the object, while $\mathbf{D}_N$ represents the 3D position. First, we present a proof of concept, whereby we introduce a MOTIONGUIDE module to learn motion using $\mathbf{C}_N$ and $\mathbf{D}_N$, and show that the learned representation of the module is invariant to shape changes but sensitive to changes in motion. Next, we formally describe how the representations of this MOTIONGUIDE module can be injected into a diffusion model to edit videos.

Figure 3. **Comparison with prior art on motion-preserved video editing.** We consider the challenging cases of Ⓐ **Cross-Domain Edit** – "silver jeep" → "bulldog on roller blades", and Ⓑ **Structure Edit** – "monkey"→ "tiger". It may be observed that in the case of cross-domain editing, all baselines suffer from severe temporal inconsistencies of the subject. For the case of structure editing, Tune-A-Video produces a highly saturated video with the head pose not correctly following the pose of the input video. Similarly, FateZero also models incorrect head pose (see second row of Ⓑ). For VideoSwap we notice that the tiger has a similar humped shape like the monkey (notice the yellow circled areas), due to the keypoint correspondences being very sparse and spatially constrained signal. The sparsity of this signal results in the orientation of the face being inaccurate, resulting in a wrong head pose of the tiger in the middle row. By comparison, VIDMP3 generates temporally consistent results following the input pose while making necessary changes faithful to the new concept.

3.1. Representing motion with pose and position

We design a MOTIONGUIDE module ϕ_m that takes as input dense correspondence maps $\mathbf{C}_N$ and depth maps $\mathbf{D}_N$ of the subject of interest in a video. We present the design of this lightweight module in the Appendix. Essentially, the module processes $\mathbf{C}_N \odot \mathbf{D}_N$ with convolution layers, then concatenates a positional encoding $\mathbf{P}$ to each frame. After another convolution, we average pool in the spatial dimensions and divide by α to form a single-dimensional vector for each frame $\mathbf{M}_{N,d}$, where α is the ratio of pixels occupied by the object in the frame to the total number of pixels in the frame. This is then processed by a final linear layer. The pooling is crucial to our method as it prevents shape and size leakage. The positional encoding $\mathbf{P}$ provides information on the 2D location of the values in $\mathbf{C}_N \odot \mathbf{D}_N$, making the representation sensitive to the average 2D position, even after spatial pooling.

For proof of concept, we designed a toy experiment where the MOTIONGUIDE module ϕ_m was attached with a final linear layer to predict the 3D trajectory and rotations of an object. We rendered a video of a cube following a specific trajectory and rotations $\mathbf{T}_{N,6}$. The cube is rendered with different gradient colors on its faces to mimic correspondence maps. We treated the rendered frames of the

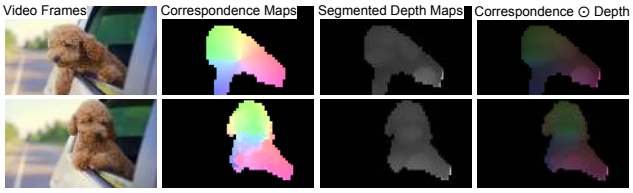

Figure 4. **Visualization of correspondence and depth maps.** For two frames of the video of "a dog looking out the window of a car", we show the corresponding correspondence and depth maps we obtain from off-the-shelf models. The depth segmented using the correspondence map is multiplied with the correspondence map (right-most column) and provided as input to MOTIONGUIDE.

cube as correspondence maps $\mathbf{C}_N$ and found depth maps of each frame, denoted as $\mathbf{D}_N$. We trained ϕ_m on this single video of the cube to predict $\hat{\mathbf{T}}_{N,6}$ by optimizing:

$$\min_{\phi_m} \|\mathbf{T}_{N,6} - \phi_m(\mathbf{C}_N, \mathbf{D}_N)\|^2 . \quad (1)$$

Given this trained MOTIONGUIDE module ϕ_m, we used it to infer on 1) ${\mathbf{C}^1}_N$, ${\mathbf{D}^1}_N$ of a positive sample where the shape of the cube was changed, but followed the same motion, and 2) ${\mathbf{C}^2}_N$, ${\mathbf{D}^2}_N$ of a negative sample where the original cube followed a different motion. We present frames of the original video, and the test videos along with prediction loss in Fig. 2. It may be observed that the training loss and loss of the positive sample follow a similar reducing trend, while that of the negative sample diverges. This shows 1) that motion can be predicted reasonably using correspondence and depth maps, 2) the learned representation is invariant to shape change, and 3) the learned representation is sensitive to motion changes.

3.2. VIDMP3

VIDMP3, depicted in Fig. 1, utilizes the MOTIONGUIDE module formulated in the previous section to learn motion from a source video $\mathbf{X}_N$, to generate a new video having the same motion. We fine-tuned our model on the single source video $\mathbf{X}_N$. We followed the paradigm of Tune-A-Video [42], where motion modules are inserted into a pre-trained T2I diffusion model. The motion module consists of temporal self-attention layers which are computed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}, \quad (2)$$

$$\mathbf{Q} = \mathbf{W}^{\mathbf{Q}}\mathbf{z}_{i,j}, \quad \mathbf{K} = \mathbf{W}^{\mathbf{K}}\mathbf{z}_{i,j}, \quad \mathbf{V} = \mathbf{W}^{\mathbf{V}}\mathbf{z}_{i,j}, \quad (3)$$

where $\mathbf{z}_{i,j}$ is the latent representation of the video at a spatial location (i, j) before the temporal self-attention. We inject the output of our MOTIONGUIDE module into the values of the temporal self-attentions such that:

$$\mathbf{V} = \mathbf{W}^{\mathbf{V}}(\mathbf{z}_{i,j} + \lambda\phi_m(\mathbf{C}_N, \mathbf{D}_N)), \quad (4)$$

where λ is a weighting factor. We chose to inject the external features into the values, to add extra context to the locations the self-attention focuses on. We used the pre-trained weights of the motion module from AnimateDiff [14].

We updated the spatial self-attention to the sparse causal variant of Tune-A-Video, where for a specific frame the attention is calculated using the first and previous frame of the video. Unlike Tune-A-Video which suffers from severe shape leakage because of over-fitting the full motion modules on the source video, we chose to keep the motion module frozen and inject motion only using the external adapter MOTIONGUIDE module. This enables us to learn a representation space of pure motion disentangled from appearance. We trained this modified network by optimizing:

$$\min_{\phi_m, \phi_u} \mathbb{E}_{z_0,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(z_t; t, y, \phi_m(\mathbf{C}_N, \mathbf{D}_N)\|^2\right], \quad (5)$$

where t represents the time-step, z_t the latents diffused at time t, y the prompt for the source video, and ϵ_θ represents the denoising diffusion model. We optimized only over ϕ_m and ϕ_u. ϕ_u represents other trainable parameters, namely $\mathbf{W}^{\mathbf{Q}}$ of the spatial self- and cross-attention layers, and $\mathbf{W}^{\mathbf{V}}$ of the motion modules. Finally, after training, we used the inverted latents of the source video to sample a new video with an edited prompt, while using $\mathbf{C}_N$ and $\mathbf{D}_N$ of the source video. We show that this simple formulation is highly robust and quite general, enabling us to generate subjects that are significantly different in shape and semantics as compared to the subject in the original video.

4. Experiments

Datasets. We used the same set of 30 videos provided by VideoSwap which were selected from Shutterstock and DAVIS [33]. The videos are divided into three categories – human, animal, and object – where each category comprises of 10 videos. For each source video we used three predefined concepts and three customized concepts, resulting in a total of 180 edited videos. Unlike VideoSwap, our customized concepts involve significant semantic changes.

Implementation Details. We used Stable Diffusion 1.5 as the foundation model for baseline comparisons and also extended our method to use SDXL for generating more diverse concepts. For the SD-1.5 architecture, we primarily use Chilloutmix [3] pre-trained weights, except for 1) style editing where we used the original SD-1.5 weights, or 2) personalized editing tasks. We used the pre-trained motion modules of AnimateDiff [14] for the temporal self-attention layers. We uniformly sampled frames at a sampling rate of 4 at their original resolution from the input video to fine-tune the models. All experiments were conducted on Nvidia A100 (40GB) and H100 GPUs. We used Adam with a learning rate of $5e^{-4}$ when optimizing the fine-tuning stage

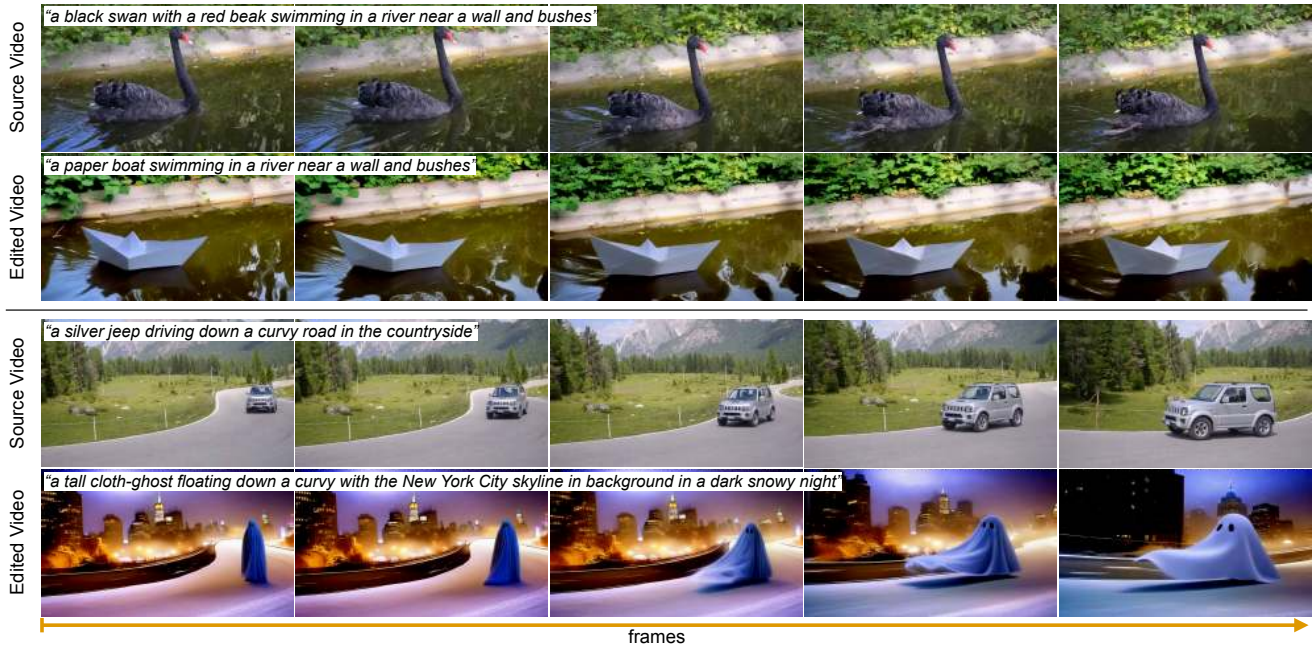

Figure 5. **Scaling VIDMP3 to SDXL.** As a novel initiative, we scaled up the T2V model to utilize SDXL as the foundation model. We show that we can model more diverse concepts using this setup, owing to the stronger generation capabilities of SDXL.

over 100 iterations. We set the MOTIONGUIDE weighting factor λ to a value of 0.1 for videos with higher ranges of motion and 0.05 for videos with lower ranges of motion. The weights of the final linear layer of the MOTIONGUIDE module are zero-initialized when training so that the output of the MOTIONGUIDE module is zero for the first iteration. We also disabled the bias of the convolution layers of the MOTIONGUIDE, since we are overfitting on one video without the need to have any regularization.

To compute spatial correspondence maps, we used the implementation of SD-Dino [49], which utlizes the internal deep features of Dino [5] and Stable Diffusion [35] for this task. For classes not present in the COCO dataset e.g. "monkey", we used the off-the-shelf figure-ground segmentation tool RMBG-1.4 [2]. We found correspondence maps for each frame using the first frame as reference. Depth maps were found using DepthAnythingV2 [45], which are then segmented to only contain the subject aided by the obtained correspondence maps. Finally we multiplied the correspondence and segmented depth maps to form the input to MOTIONGUIDE. We show examples of the computed correspondence and depth maps for a video in Fig. 4.

Baselines. We qualitatively and quantitatively compared our model to Tune-A-Video [42], VideoSwap [13], and FateZero [34]. We found these baselines to be the most relevant ones delivering the strongest results for motion-preserved editing tasks using a single video for training.[1] We show in the Appendix that first-frame editing methods like AnyV2V struggle to capture considerable levels of motion and are highly dependent on the quality of the first frame generated by their image editing method.

5. Results

Here we showcase some of the various capabilities of VIDMP3, comparison to baselines, adaptability of our model to various video editing tasks, scaling to SDXL, ablations over the components of our method, and discuss implementation choices.

Cross-domain Edit. The most important contribution of VIDMP3 lies in the challenging case of Cross-domain Editing, where previous methods suffer. In this case, we show that the subject in the source video can be swapped with a semantically different subject in the edited video, while correctly preserving motion. In Fig. 3 we show the instance "silver jeep" $\rightarrow$ "bulldog on roller blades," where VIDMP3 can generate a video where the motion is preserved and the subject is temporally consistent. We attribute these results to the external strong motion signal we inject, which allows the model to understand a general sense of position and pose. We present additional results in the Appendix.

[1] CAMEL [48] is a related work but does not provide sufficient results, and omits dependencies required to run their code in their repository.

Structure Edit. Previous methods have shown good performance for the case of structure editing, while keeping the edited subject in the same domain, e.g., "silver jeep" $\rightarrow$ "Porsche." This case is much simpler as compared to cross-domain editing, due to the internal semantic understanding of the diffusion model. We show the case of "monkey" $\rightarrow$ "tiger" in Fig. 3, where the edited tiger generated by VIDMP3 follows the exact same head and hand motion as the monkey, allowing freedom for the different body shapes of the tiger as compared to the monkey. We present additional results for structure editing in the Appendix.

Comparison to baselines. For the two previously described cases of **Cross-Domain Edit** and **Structure Edit**, we compared to the previous methods, Tune-A-Video, VideoSwap and FateZero. For fair comparison, we initialized all baselines with the same pre-trained T2I weights [3] as ours. Tune-A-Video and FateZero don't explicitly provide any external guidance to the model, which lead to high temporal inconsistencies in the case of Cross-Domain Editing, where the pre-trained T2I model is not confident in its outputs owing to semantic changes of the object to be edited with respect to the source object. On the other hand, VideoSwap uses explict keypoint correspondences and guides the model to change the object, but it fails when the semantic meanings do not remain relevant (e.g.: "silver jeep" $\rightarrow$ "brown bulldog"). VideoSwap requires human effort in marking the positions of 2D keypoints that should be tracked in the video. It also involves significant time and human effort to manually edit the position of the keypoints for the target video when there are significant shape changes. Tune-A-Video generates saturated videos on both Cross-Domain and Structure Editing, possibly due to overfitting the entire motion module on the source video. This is not true for either VideoSwap or VIDMP3, as all or most parts of the motion module are kept frozen while learning an external adapter that has a fixed input. For the case of Structure Editing, it may be observed that VideoSwap generated the tiger to be in a bent posture like the monkey, because fitting to the keypoint correspondence signal was too constrictive. None of the baselines were able to follow the head pose of the monkey accurately as can be observed especially in the second and third row of the generated videos of all baselines in Fig. 3 Ⓑ.

By contrast, VIDMP3 generates temporally consistent videos in both cases while preserving motion from the source video. This is achieved by computing a pose and position representative value for each frame, using dense correspondence and depth maps to learn generalized representation of motion. During inference, these representations help guide motion in the generated videos, and allowing the text-to-image model more room to explore appearances.

Adaptability of VIDMP3. Since VIDMP3 is based on an existing T2I model, it can be effectively applied to tasks other than subject swapping. We show results of using VIDMP3 for 1) background change, 2) style change, and 3) personalization. For personalization, we attempted both per-subject personalization, as well as using pre-trained T2I models that are personalized on more general concepts, such as Anything-v4.0 [1]. We refer the readers to the Appendix for qualitative results of these tasks.

Scaling VIDMP3 to SDXL. We also studied scaling to StableDiffusion-XL which is a stronger T2I model that is able to represent more diverse concepts. We use AnimateDiff's SDXL motion module, and found that it less effectively models motion than the motion module of the SD-1.5 version. Thus, we identify the specific parameters of the motion module that contribute to shape leakage and kept them frozen. More specifically, we found that the feed-forward layers of temporal self-attention blocks contribute to the highest leakage. We trained the other parameters namely, $\mathbf{W^Q}$, $\mathbf{W^K}$, and $\mathbf{W^V}$ and projection matrices of the temporal self-attention modules, in addition to the parameters that we kept trainable in the SD-1.5 version. This enabled our model to better learn motion while still avoiding leakage. We present results of using SDXL within VIDMP3 in Fig. 5. We show generated concepts that we were not able to represent consistently using SD-1.5 VIDMP3, such as a "paper boat" and a "cloth-ghost". We provide additional results generated by VIDMP3 SDXL in the Appendix.

Ablations. We conducted ablations over various components of our method and implementation choices. Fig. 6 depicts the effect of using only correspondence maps, only depth maps, or concatenated depth and correspondence maps as input to MOTIONGUIDE. We also show the effect of disabling the MOTIONGUIDE. For all these cases, we find that the motion is modeled incorrectly, with a much subdued range and incorrect orientations per frame.

Evaluation. We quantitatively compared our method against previous SOTA models using both automatic and human evaluations. We provide a detailed discussion of the evaluation settings in the Appendix. We conducted a voluntary, controlled laboratory human study to gather opinions expressive of 1) Subject Identity, 2) Motion Alignment, 3) Temporal Consistency, and 4) Overall Preference for video subject swapping. The results of this evaluation, shown in Fig. 7, indicate a clear preference for our method.

Time Cost Analysis We recorded the time required to run each component of VIDMP3 to edit a 16 frame video clip on an Nvidia A100 GPU. This includes 1) *Preprocessing*:

Figure 6. **Ablation over various components of our method.** For the case of "car" → "bike", we see that all other implementations including disabling MotionGuide, providing different inputs to MotionGuide such as only correspondence maps, only depth maps or concatenation of depth and correspondence maps, results in incorrect and lower range of motion. VidMP3 uses MotionGuide with multiplied correspondence and depth maps as input and imitates the motion the subject in the source video correctly.

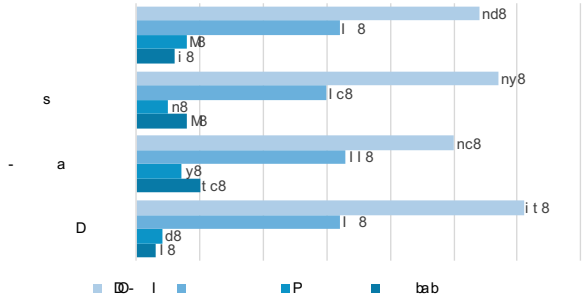

Figure 7. **Human opinions** on 1) Subject Identity, 2) Motion Alignment, 3) Temporal Consistency, and 4) Overall Preference, averaged over 10 participants and 180 edited videos.

which involves computing the correspondence and depth maps. The correspondence map computation required approximately 4s per frame, or 64 seconds over 16 frames. The depth map computation required approximately 2s per frame, or 32 seconds over 16 frames. The preprocessing step used about 100 seconds overall. 2) *Training*: where the MotionGuide module was trained over 100 iterations which expended about 3 minutes of compute time. And, lastly 3) *Editing*: when we generated the edited video using inverted noise from the source video. The DDIM inversion process of 50 steps required about 30 seconds. The backward process to generate the edited video consumed about 30 seconds as well, resulting in a total of 60 seconds. Overall, the complete process, from preprocessing to generating the final edited video required about 6-7 minutes.

6. Limitations and Discussions

There can be multiple choices of features that could represent the pose and position of subjects, and that can be injected externally to a diffusion model to guide motion, e.g. injecting Diffusion Correspondence (DIFT [39]) features. However, our choice of 2D correspondence and depth maps is highly efficient since it only requires three channels of input, and is also a cleaner signal of explicit motion without any leakage of extra information. While VidMP3 can generate subjects with significant structural and semantic differences relative to the source video, we cannot explicitly control the size of the subject. For example, in the case of "black swan" → "paper boat" in Fig. 5, observe that the generated paper boat is large and of a similar size as the swan. Additionaly, our method is dependant on the quality of correspondence and depth maps obtained. However, for all of our evaluation videos, we find that the off-the-shelf methods for obtaining these maps perform well. Scaling video editing to multiple subjects has been studied in previous work, but has not been explored here. For such scenarios, one approach would be to generate separate correspondence maps for each of the various subjects of interest, and inject each using separate MotionGuide modules. We leave this direction of research for future work.

7. Conclusion

We presented VidMP3, a novel video editing technique based on T2I models, which utilizes pose and position priors to generate motion-preserved videos based on a source video. We introduce the MotionGuide module, which learns generalized motion representations from spatial correspondence and depth maps. These representations are injected into the temporal self-attention layers of a T2V model initialized from a T2I model, thus forming VidMP3. We evaluated VidMP3 on challenging video editing tasks: 1) Cross-Domain Editing, and 2) Structure Editing. We observed that VidMP3 can generate objects with significant structural and semantic changes relative to the subject in the source video, while maintaining temporal consistency. We show qualitatively and quantitatively that our method improves over previous strong baselines on the task of motion-preserved video editing. Additionally, we scaled our method to use SDXL as the base T2I model, which is a novel effort in the area of video editing. We explored the adaptibility of our method on various video editing tasks, including personalized editing, background editing, and style editing. Despite its potential to enhance creative

workflows, motion-preserved video editing without rigid structural constraints remains a relatively under-explored domain. VIDMP3 addresses this gap by introducing a novel approach that maintains temporal coherence while allowing flexible content modification, laying the groundwork for future research and developments in this area.

References

[1] Anything-v4.0. https://huggingface.co/xyn-ai/anything-v4.0. [Online; accessed 14-November-2024]. 7

[2] RMBG-1.4. https://huggingface.co/briaai/RMBG-1.4. [Online; accessed 14-November-2024]. 6

[3] chilloutmix. https://huggingface.co/swl-models/chilloutmix?not-for-all-audiences=true. [Online; accessed 14-November-2024]. 5, 7

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

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 6

[6] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23040–23050, 2023. 2

[7] Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In *Forty-first International Conference on Machine Learning*, 2023. 3

[8] Paul Couairon, Clément Rambour, Jean-Emmanuel Haugeard, and Nicolas Thome. Videdit: Zero-shot and spatially aware text-driven video editing. *Transactions on Machine Learning Research*, 2023. 2

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

[10] Ruoyu Feng, Wenming Weng, Yanhui Wang, Yuhui Yuan, Jianmin Bao, Chong Luo, Zhibo Chen, and Baining Guo. Ccedit: Creative and controllable video editing via diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6712–6722, 2024. 2

[11] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 2

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

[13] 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. Videoswap: Customized video subject swapping with interactive semantic point correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7621–7630, 2024. 2, 3, 6

[14] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 5

[15] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

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

[17] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 3

[18] Hyeonho Jeong and Jong Chul Ye. Ground-a-video: Zero-shot grounded video editing using text-to-image diffusion models. *arXiv preprint arXiv:2310.01107*, 2023. 2

[19] Hyeonho Jeong, Jinho Chang, Geon Yeong Park, and Jong Chul Ye. Dreammotion: Space-time self-similar score distillation for zero-shot video editing. *arXiv preprint arXiv:2403.12002*, 2024. 2

[20] Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M Rehg, and Pinar Yanardag. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6507–6516, 2024. 2

[21] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion video synthesis with stable diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22680–22690, 2023. 3

[22] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021. 3

[23] Max Ku, Cong Wei, Weiming Ren, Huan Yang, and Wenhu Chen. Anyv2v: A plug-and-play framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024. 3

[24] Xirui Li, Chao Ma, Xiaokang Yang, and Ming-Hsuan Yang. Vidtome: Video token merging for zero-shot video editing.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7486–7495, 2024. 2

[25] Feng Liang, Bichen Wu, Jialiang Wang, Licheng Yu, Kunpeng Li, Yinan Zhao, Ishan Misra, Jia-Bin Huang, Peizhao Zhang, Peter Vajda, et al. Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8207–8216, 2024. 2

[26] Jia-Wei Liu, Yan-Pei Cao, Jay Zhangjie Wu, Weijia Mao, Yuchao Gu, Rui Zhao, Jussi Keppo, Ying Shan, and Mike Zheng Shou. Dynvideo-e: Harnessing dynamic nerf for large-scale motion-and view-change human-centric video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7664–7674, 2024. 3

[27] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8599–8608, 2024. 2

[28] Haoyu Ma, Shahin Mahdizadehaghdam, Bichen Wu, Zhipeng Fan, Yuchao Gu, Wenliang Zhao, Lior Shapira, and Xiaohui Xie. Maskint: Video editing via interpolative non-autoregressive masked transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7403–7412, 2024. 2

[29] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4117–4125, 2024. 3

[30] Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. Codef: Content deformation fields for temporally consistent video processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8089–8099, 2024. 2

[31] Wenqi Ouyang, Yi Dong, Lei Yang, Jianlou Si, and Xingang Pan. I2vedit: First-frame-guided video editing via image-to-video diffusion models. *arXiv preprint arXiv:2405.16537*, 2024. 3

[32] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[33] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 5

[34] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15932–15942, 2023. 2, 6

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 6

[36] Chaehun Shin, Heeseung Kim, Che Hyun Lee, Sang-gil Lee, and Sungroh Yoon. Edit-a-video: Single video editing with object-aware consistency. In *Asian Conference on Machine Learning*, pages 1215–1230. PMLR, 2024. 2

[37] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[38] Yeji Song, Wonsik Shin, Junsoo Lee, Jeesoo Kim, and Nojun Kwak. Save: Protagonist diversification with structure agnostic video editing. In *European Conference on Computer Vision*, pages 41–57. Springer, 2025. 2, 3

[39] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 8

[40] Jiangshan Wang, Yue Ma, Jiayi Guo, Yicheng Xiao, Gao Huang, and Xiu Li. Cove: Unleashing the diffusion feature correspondence for consistent video editing. *arXiv preprint arXiv:2406.08850*, 2024. 2

[41] Wen Wang, Yan Jiang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023. 2

[42] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 2, 3, 5, 6

[43] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024. 3

[44] Wilson Yan, Andrew Brown, Pieter Abbeel, Rohit Girdhar, and Samaneh Azadi. Motion-conditioned image animation for video editing. *arXiv preprint arXiv:2311.18827*, 2023. 2

[45] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 6

[46] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 2

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

[48] Guiwei Zhang, Tianyu Zhang, Guanglin Niu, Zichang Tan, Yalong Bai, and Qing Yang. Camel: Causal motion enhancement tailored for lifting text-driven video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9079–9088, 2024. 2, 3, 6

[49] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024. 6

[50] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 2

[51] Xiaojing Zhong, Xinyi Huang, Xiaofeng Yang, Guosheng Lin, and Qingyao Wu. Deco: Decoupled human-centered diffusion video editing with motion consistency. In *European Conference on Computer Vision*, pages 352–370. Springer, 2025. 3

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