GenVideo: One-shot target-image and shape aware video editing using T2I diffusion models

Sai Sree Harsha
Adobe Inc .
ssree@adobe.com

Ambareesh Revanur
Adobe Inc.
arevanur@adobe.com

Dhwanit Agarwal
Adobe Inc.
dhagarwa@adobe.com

Shradha Agrawal
Adobe Inc.
shradagr@adobe.com

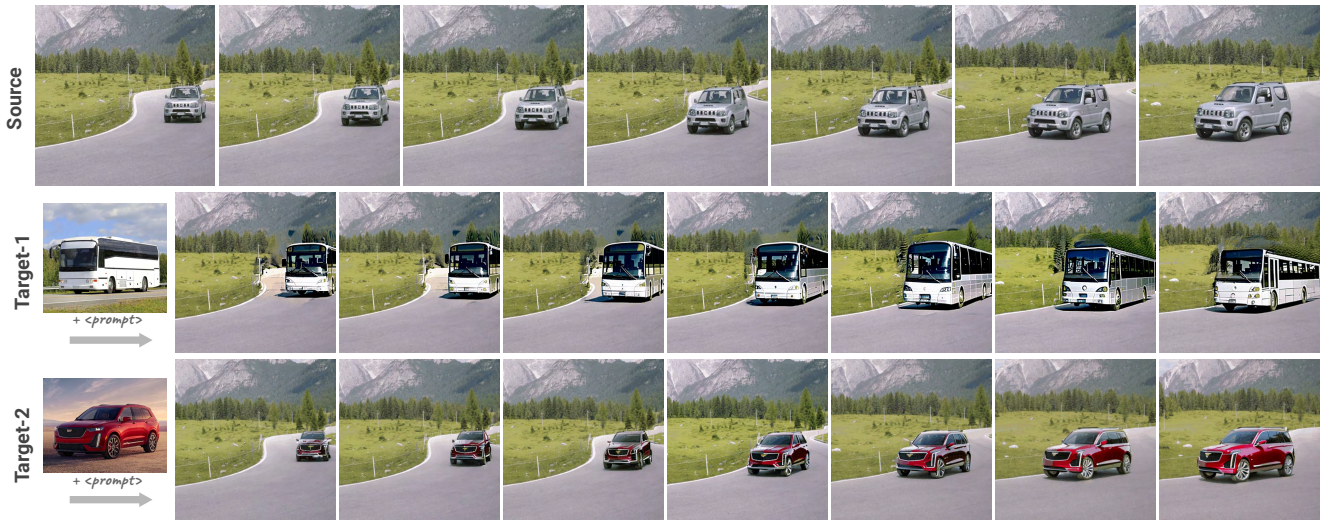

Figure 1. Editing the source car video based on a **target-image** and a text prompt *Driving a "white bus"/"red cadillac" down a mountain road* with GenVideo method. GenVideo can accurately replace the object in the source video, even with a target object of a different shape.

Abstract

Video editing methods based on diffusion models that rely solely on a text prompt for the edit are hindered by the limited expressive power of text prompts. Thus, incorporating a reference target image as a visual guide becomes desirable for precise control over edit. Also, most existing methods struggle to accurately edit a video when the shape and size of the object in the target image differ from the source object. To address these challenges, we propose "GenVideo" for editing videos leveraging target-image aware T2I models. Our approach handles edits with target objects of varying shapes and sizes while maintaining the temporal consistency of the edit using our novel target and shape aware InvEdit masks. Further, we propose a novel target-image aware latent noise correction strategy during inference to improve the temporal consistency of the edits. Experimental analyses indicate that GenVideo can effectively handle edits with objects of varying shapes, where existing approaches fail.

1. Introduction

Image and video generation using diffusion models [7, 8, 12, 18, 25, 26, 30] has gained immense popularity in recent times. Recently, many methods have been proposed for editing visual content like images [33, 35, 36], and videos [5, 37, 43, 47] using text prompts. Although several works have showcased high-quality results for the task of image editing [1, 6, 11, 28, 41], the utilization of diffusion models for the task of video editing has been limited. Works such as [8, 24], use text-to-video (T2V) diffusion models for video editing. However, these methods have limited utility since they demand a large computational infrastructure for training along with large and diverse video dataset. Due to these limitations, more recent methods have started exploring inflated text-to-image (T2I) diffusion models as an alternative to T2V models for video editing in a one-shot or zero-shot fashion [2, 4, 5, 9, 21, 31, 43, 46]. Nonetheless, with the exception of Make-A-Protagonist [46] which can be additionally conditioned on a target-image, existing and concurrent video editing methods that are based on T2I diffusion models focus exclusively on text-driven video editing. There-

fore, these methods are not suitable for scenarios where the nature of the edit cannot be accurately expressed using only text. In other words, they are not target-image aware. We emphasize here that target-image awareness provides a precise visual guide for the desired edit. This approach enables exact replication of shapes, sizes, and textures, essential for creating content that serves use cases like creating content that aligns with a particular product specifics (like a particular model of a car). Further, previous approaches lack shape awareness and therefore, fail to make edits where the shape and size of the target-object is substantially different from the source video.

In this paper, we propose a novel target-image and shape aware *InvEdit* method to effectively identify a region-of-interest in the video for a given text prompt and the target-image. This is done using the underlying T2I diffusion model similar to [6] but for videos and using additional image conditioning. The strength of our method lies in InvEdit and its simplicity with which it can be adapted to any image conditioned diffusion model to get target-image and shape aware mask for video editing. Localized mask guidance is often crucial for precise edits and avoiding leakage of edits to the unmasked region but obtaining target-image and shape aware masks is a challenging problem due to which previous methods are not able to handle edits of substantially different shapes and sizes. We further address the key challenge of maintaining the temporal consistency in the edited video. Some of the previous works in this research area fail to effectively address this challenge because of their reliance on source-based Neural Layer Atlases (NLA) [5] or source-based inter-frame feature propagation [9]. Hence, the temporal consistency of the video generated by these methods is primarily dictated by the source video. Naturally, these methods fail when the target-object is substantially different from the source. To address these concerns, we introduce a novel target-image aware *latent correction* strategy to blend the inter-frame latents of the diffusion model on the fly during inference to improve the inter-frame temporal consistency of the target-object in the edited video. Our approach uses *InvEdit* mask guidance to enable temporal consistency even when the target-object has a different shape and size than the object in the source video. We show extensive quantitative and qualitative comparisons with state-of-the-art video editing methods and conclude that our approach outperforms previous and concurrent approaches on a diverse set of videos. Our contributions can be summarized as follows:

- We introduce *GenVideo*, the first pipeline for target-image and shape aware editing of videos using image diffusion models that can make temporally consistent edits for target objects of varied shapes and structures.
- We introduce *InvEdit*, a novel target-image and shape aware, zero-shot approach for generating video masks using image diffusion models. This mask enables us to identify the region of interest for editing objects of varying shapes and sizes.
- We introduce a target-image and shape aware *latent correction* strategy to enforce inter-frame temporal consistency in the edited video, where existing approaches fail.

2. Related Work

Text Driven Image and Video Editing: Before the advent of the diffusion models [12, 38], generative adversarial networks (GANs) have been used to edit images [15, 27, 29, 34, 42, 44]. Recently, generative diffusion models [33, 35, 36] trained on internet scale image and text data have achieved significantly higher quality and diversity of image generation than GANs. Inspired by the success of these T2I diffusion models, text-to-video diffusion models [13, 14, 22, 37] have been developed for video generation. Authors in [37] inflate the pretrained T2I diffusion models with additional temporal attention layers to generate videos, while the recent works like [13] use a 3D-UNet for video generation. Along with generation, editing of existing images and videos has also received significant interest. SDEdit [23] carefully adds noise to an input image and then denoises them using diffusion models to achieve the desired edits. Prompt-to-prompt (P2P) [11] allows text guided editing of images by controlling the cross-attention maps. In the realm of video editing, Tune-A-Video (TAV) [43] finetunes the inflated T2I model on an individual source video to generate edited videos with similar motion. Video-P2P [21] and FateZero [31] achieve controllable edits in videos by altering the cross-attention maps. Pix2Video [4] achieves video editing by altering the self-attention features during inference. Text2Live [2] utilizes NLA and then edits each layer independently using text guidance. Text2Video-Zero [16] uses inflated self-attention layers along with ControlNet [45] to do text guided video editing. Nonetheless, these works using T2I diffusion models exclusively focus on image and video editing via text, and cannot be used for edits which are indescribable by text alone. Hence, they are not target-image aware which is often a crucial aspect for accurate video editing [46].

Target-Image Aware Visual Content Editing: To cater to visual content edits that isn't easily or accurately described by text alone, models like DALLE–2 [33] and Gen–1 [8] have been developed to make target-image aware edits to images and videos by leveraging the CLIP [32] image embedding. While DALLE–2 edits images, Gen–1 can edit videos but the drawback here is that the entire image, including the background, influences the edits due to the the lack of mask guidance. Subject driven solutions like Dreamix [24], which can edit videos conditioned on the target-image data, typically require multiple target images

for learning the concept/target object accurately and also require large video diffusion models as used in [13]. While they can generate and edit videos based on a single image, the edited videos fail to maintain the fidelity of the edits to the target-object in the target-image. Make-A-Protagonist [46] addresses these limitations by using an image guided T2I diffusion model called SD-unCLIP [33]. Their pipeline can achieve target-image aware video object editing using a single target-image and employ mask-guided inference to do localized edits. However, the edits still suffers from lack of temporal consistency and the approach fails when the target object differs in shape and size from the source object present in the source video. This is because the mask used for editing the video is generated only using the source video and has no awareness of the target-object.

Video Editing with Temporal Consistency: Aforementioned methods suffer from lack of inter-frame temporal consistency in the edited video. A concurrent work, *TokenFlow* [9] improves the temporal consistency by inter-frame propagation of the self-attention features using a nearest neighbor field in the feature space of source video. However, both the editing and temporal consistency in *TokenFlow* are not target-image and shape aware. Thus, it cannot make edits where the target-object differs a lot in shape and size to the source object. Moreover, it does not offer a way to localize the edits due to the lack of mask guidance and also requires manually choosing key frames for imposing temporal consistency. *StableVideo* [5] uses NLA of the source video to enforce temporal consistency and hence, also lacks target-image and shape awareness for accurate shape changing edits. Another work [19] addresses shape-aware editing by formulating a deformation map which maps the original atlas into edited atlas. While the method achieves favorable results, NLA mapping takes tens of hours of training and an inaccurate NLA mapping results in objects with missing parts and substantial undesirable background changes. Our work *GenVideo* addresses these limitations by providing the first "target-image and shape aware" video editing using T2I models with only *30 mins of training and 2 mins of inference* using localized mask guidance to avoid background changes.

3. Approach

GenVideo aims to edit the given source video based on a target text prompt and a target-image containing an object of any shape while maintaining temporal consistency. More formally, given an input source video $\mathcal{V}^{src} = [I_1^{src}, \cdots, I_N^{src}]$ composed of N frames containing a source object, a source text prompt $\mathcal{P}^{src}$ describing the source video, a target-image I^{trg} containing the target-object and a target text prompt $\mathcal{P}^{trg}$ describing the desired edit to the source video, *GenVideo* generates a target video $\mathcal{V}^{trg} = [I_1^{trg}, \cdots, I_N^{trg}]$ which preserves the motion of the input source video but replaces the source object with a new target-object from the target-image. The entire training and inference pipeline is summarized in Fig. 2. First, we fine-tune an inflated Stable Diffusion unCLIP (SD-unCLIP) model [33, 46] on the source video using the reconstruction loss from standard LDMs [35] (Sec. 3.1). Then, we employ our novel target-image and shape aware mask generation approach, called *InvEdit*, where we use the fine-tuned model to infer a region-of-interest where the edits need to be localized (Sec. 3.2). Finally, we introduce a novel *latent correction* approach to improve the inter-frame temporal consistency (Sec. 3.3).

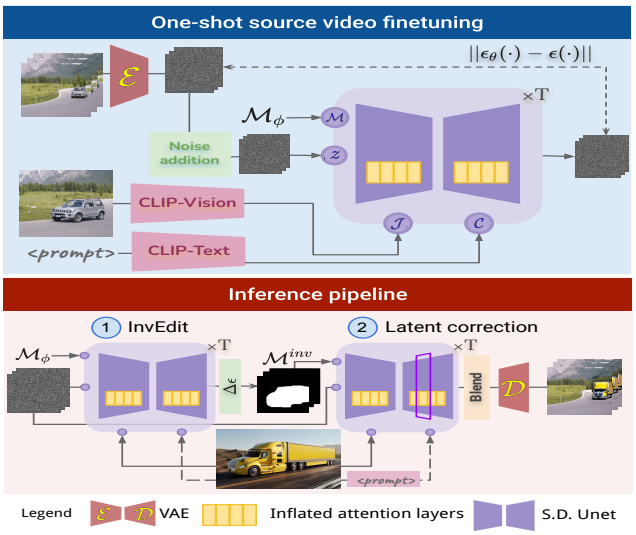

Figure 2. Overview of *GenVideo*. Inflated attention layers are fine-tuned during source video finetuning. During inference, *InvEdit* predicts a region to edit and *latent correction* uses that mask to improve the inter-frame temporal consistency. $\mathcal{M}_\phi$ - "no mask".

3.1. Fine-tuning on the source video

We fine-tune an inflated SD-unCLIP model on the source video as shown in Fig. 2. The finetuning procedure of the inflated pretrained T2I diffusion model is in line with the TAV method [43]. Unlike [43], we use the SD-unCLIP model [33] which conditions the generation on both target-image and the text prompt[46]. This model uses CLIP-vision branch to obtain an image embedding $\mathcal{J}^*$ for an input reference image and uses the CLIP-text branch to obtain the text embedding $\mathcal{C}^*$. As part of the inflation process, spatial self-attention is inflated into spatio-temporal attention (ST-attn), and an additional temporal attention (T-attn) is introduced after the ST-attn and cross-attention blocks. Please see the Supplementary Material (SM) for full details.

3.2. InvEdit mask generation

In this section, we describe the details of *InvEdit* – our novel zero shot, target-image and shape aware mask generation strategy using the finetuned diffusion model from

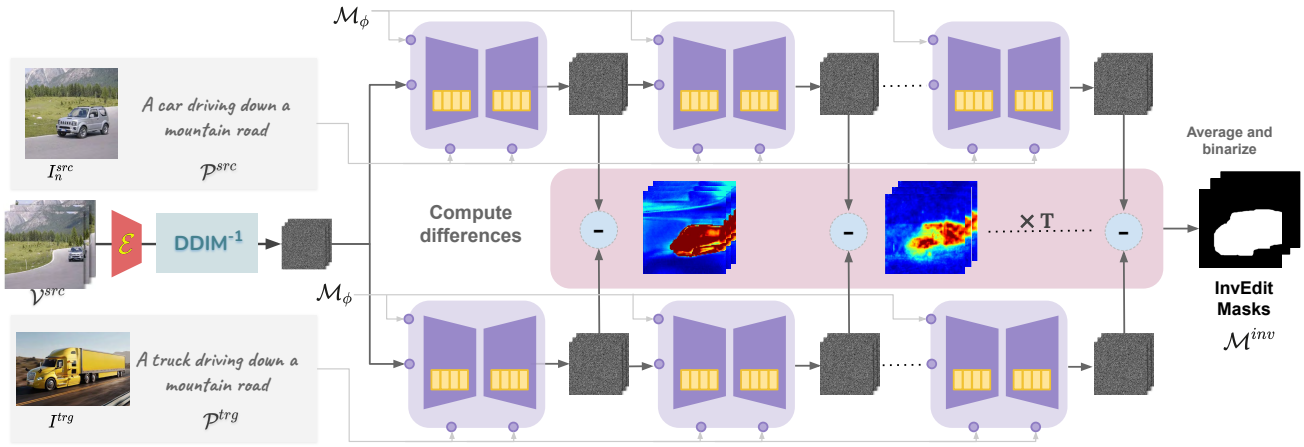

Figure 3. *InvEdit* approach - the mask is generated by first iteratively computing noise differences across multiple timesteps for the source denoising branch and target denoising branch. Then, these differences are averaged and binarized to obtain the *InvEdit* mask.

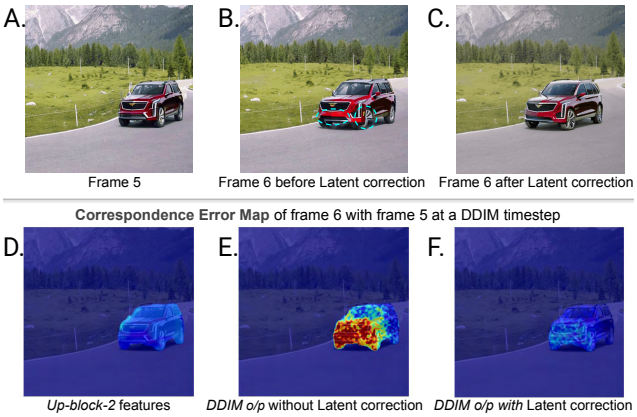

Figure 4. *Latent correction* strategy corrects the latent noise using UNet features from the previous and successive video frame. Correspondence Error (CE) map computed using the ground truth (source video correspondences) shows that the DDIM o/p of the model before correction has high CE (**E.**) which our *latent correction* strategy fixes (**F.**) using `Up-block-2` feature correspondences which have low CE (**D.**).

Sec. 3.1. Existing methods[31, 46] compute masks using the source video alone and hence have no shape awareness regarding the target object's relative shape and size (for example, changing a car to a bus). InvEdit is an adaptation of DiffEdit [6] for videos and includes target-image and shape awareness.

***InvEdit* steps:** First, we perform the DDIM inversion [38] to transform the source video into the corresponding random latent noise $\mathcal{Z}_T^{src} = [z_{T,1}^{src}, z_{T,2}^{src}, \cdots, z_{T,N}^{src}]$. Then, we denoise the latent code $\mathcal{Z}_T^{src}$ using the fine-tuned inflated SD-unCLIP model from Sec. 3.1 using deterministic DDIM sampling for both the source and target branches as shown in Fig. 3. For the source branch, we denoise $\mathcal{Z}_T^{src}$ using the source text prompt $\mathcal{P}^{src}$ and a randomly chosen frame I_n^{src} from $\mathcal{V}^{src}$ as conditioning inputs for the fine-tuned SD-unCLIP model. Similarly, in parallel, we repeat DDIM sampling using the target text prompt $\mathcal{P}^{trg}$ and the foreground object in the target-image I^{trg} as conditioning inputs. We segment the foreground object from the target-image using GroundedSAM [17, 20]. We compute the difference of the predicted noise by the denoising UNet model (denoted by ε_θ) obtained at each denoising timestep in the source and target branches. More formally, for every I_n^{src} in $\mathcal{V}^{src}$, we compute $\Delta\varepsilon_{t,n} = abs(\varepsilon_{t,n}^{src} - \varepsilon_{t,n}^{trg})$, where $\varepsilon_{t,n}^{u} = \varepsilon_\theta\left(z_{t,n}^{u}, t, \mathcal{C}^u, \mathcal{J}_n^u\right), u \in \{src, trg\}$. At each DDIM denoising step we obtain,

$$\mathcal{Z}_t^{src} = \text{DDIM}(\mathcal{Z}_{t+1}^{src}, \epsilon_{t+1}^{src}), \mathcal{Z}_t^{trg} = \text{DDIM}(\mathcal{Z}_{t+1}^{trg}, \epsilon_{t+1}^{trg}), \quad (1)$$

where $t \in \{T-1, \cdots, 1\}$ denotes the time step.

These differences $\Delta\varepsilon_{t,n}$, represented as heat maps in Fig. 3, are averaged over multiple denoising time steps and binarized to get the target-aware *InvEdit* mask for each frame of the video. In Fig. 3, the *InvEdit* mask is able to determine that edits are to be placed in a region shaped like a truck rather than the car, as the truck is much bigger than the car. We denote the masks for N frames by $\mathcal{M}^{inv} = [M_1, M_2, \cdots, M_N]$, where $M_n = binarize(mean_{t\in[t_i,t_j]}(\Delta\varepsilon_{t,n}))$. The *InvEdit* mask is used to identify the region in which the target-image embeddings are injected into the features of the ResNet blocks of the UNet along with the timestep embeddings [33].

***InvEdit* intuition:** Our intuition follows DiffEdit [6] for text driven image editing. We extend it for our target-image aware video editing use case. During DDIM denoising, the SD-unCLIP model will yield different noise estimates given different text and image conditioning. The noise estimates will be different in regions where the latent code will eventually decode different shapes, colors and textures depending on the conditioning. For the background, on the other hand, there is little change in the noise estimates. The difference between the noise estimates can thus be used to infer a mask that identifies the parts of each video frame need to be edited.

3.3. Latent noise correction via self-consistency

While *InvEdit* mask is able to accurately identify the regions to edit, it does not address the temporal consistency of the object within the region across generated frames. As an example, consider the edit where the "*silver car*" is edited to a "*red cadillac*" shown in Fig. 4. While the shape of the car generated by *InvEdit* mask in Frame 6 (see Fig. 4B) appears similar to Frame 5 (see Fig. 4A), it has a different stylistic appearance in the front and in the side. One trivial way to address this is by computing the optical flow of the source object in the video and then imposing this flow on the latent noise features $\mathcal{Z}$. However, this solution will struggle considerably even in typical situations where the target object is of a different shape. Thus, the issue of temporal inconsistency becomes challenging given the model has not seen the target-image before. We address it by introducing a *latent correction* strategy during inference. This correction is a blending strategy in the latent $\mathcal{Z}$ space to improve the inter-frame temporal consistency of the edited video. This is a three-step process (also see SM):

1. **Inter-frame latent field computation:** During each denoising time step t of inference, we utilize the features of the `Up-Block-2` of the UNet denoted as $[f_1^t, \cdots, f_N^t]$ for estimating latent correspondence maps/field using nearest neighbors between these features of the consecutive frames. First, we compute the nearest neighbors field $\mathcal{N}_{i\pm}(\cdot)$ defined as $\mathcal{N}_{i\pm}^t[p] := \text{argmax}_q d(f_i^t[p], f_{i\pm1}^t[q])$. This field is the mapping from the spatial locations p in the features of i^{th} frame to its nearest neighbor q (in terms of cosine similarity d) in the features of the $(i \pm 1)^{th}$ frame. This field is the mapping from the spatial locations p in the features of i^{th} frame to its nearest neighbor q in the features of the $(i \pm 1)^{th}$ frame.
2. **Blending using inter-frame latent field:** Starting from the DDIM inversion of the source video $\mathcal{Z}_T^{src}$, at each denoising timestep t during inference, we blend the latents $\mathcal{Z}_t = [z_{t,1}, z_{t,2}, \cdots, z_{t,N}]$ of adjacent frames inside the *InvEdit* mask region in the latent space of the decoder $\mathcal{D}$ of the VAE of SD-unCLIP. The blended latents $\tilde{\mathcal{Z}}_t = [\tilde{z}_{t,1}, \cdots, \tilde{z}_{t,N}]$ at timestep t are given by,

$$\begin{aligned}\tilde{z}_{t,i}[p] = w_{-1}(M_i \odot z_{t,(i-1)}[\hat{\mathcal{N}}_{i-}^t(p)]) + w_0(M_i \odot z_{t,i}) \\ + w_1(M_i \odot z_{t,(i+1)}[\hat{\mathcal{N}}_{i+}^t(p)]) \\ + (1 - M_i) \odot z_{t,i},\end{aligned}$$

   where w_{-1}, w_0, w_{+1} are non-negative weight hyperparameters which add up to 1 and $\hat{\mathcal{N}}_{i\pm}^t[p]$ is the nearest neighbors field above upsampled to match the dimension of the z_t. This blending happens at each inference timestep t for $t \geq T - 5$.
3. **Background preservation:** We also correct the latent noise corresponding to background regions using the inverse *InvEdit* mask, *i.e.,* $(1 - M_i)$ and denoise only the masked area [39]. This is achieved by retaining the clean latent corresponding to the source video frame outside the masked region as given by $\tilde{z}_{t,i} = (1 - M_i) \odot \mathcal{E}(I_i^{src}) + (M_i) \odot \tilde{z}_{t,i}$, where $\mathcal{E}$ is the encoder of VAE. We skip this step when the background of the target video is expected to be different from the source.

Why Up-block-2? We use the features of `Up-block-2` since it demonstrates a low correspondence error (CE) than the CE in the latent noise obtained after DDIM step. Continuing the example of the car in Fig. 4, we first compute the feature correspondences in consecutive frames using RAFT optical flow [40] on the source video frames. This serves as a ground truth correspondence for this example since the edited object has the same shape as the source object. Then we compute the correspondences of the features of `Up-block-2` in consecutive frames and find that these features have a low CE rate as shown by the heatmap in Fig. 4D. On the other hand, the CE rate is higher for latent noise computed after the DDIM Step (Fig. 4E). With our *latent correction* strategy, the CE is lowered since the proposed blending strategy improves the consistency of the latent noise features across consecutive frames as demonstrated by Fig. 4C and Fig. 4F. We show in our experiments that this process improves the temporal consistency of the edited target video.

4. Experiments

4.1. Implementation details

For all of our experiments, we use the SD-unCLIP version of Stable Diffusion v2.1 T2I model [33]. We inflate the model with temporal layers as in [43]. See SM for more details on model inflation. During source finetuning, we train the inflated layers using 16 source video frames with a learning rate of 1×10^{-5} for 400 iterations. We use DDIM sampler during inference to obtain the *InvEdit* mask on the DDIM-inverted source video by computing the mean of $\Delta\epsilon_t$ across $t \in [0.8 \times T, T]$ where $T = 50$ and use a difference threshold of 0.6 to binarize the heatmap. During the *latent correction*, we restrict the nearest neighbour search to a 4×4 window of pixels and set $w_{-1} = 0.1$, $w_0 = 0.8$ and $w_1 = 0.1$ across the experiments. We also fix classifier-free guidance scale to 12.5 during training and inference.

4.2. Applications

Shape-aware video object editing : Our approach can be used to do target-image and shape aware video object editing. Fig. 5, along with the second series in Fig. 1, showcase the outcomes of temporally consistent video object editing using *GenVideo*. This task is difficult as the mask needs to be target-image and shape aware to effectively identify the region of interest for localized edits. For instance, in the

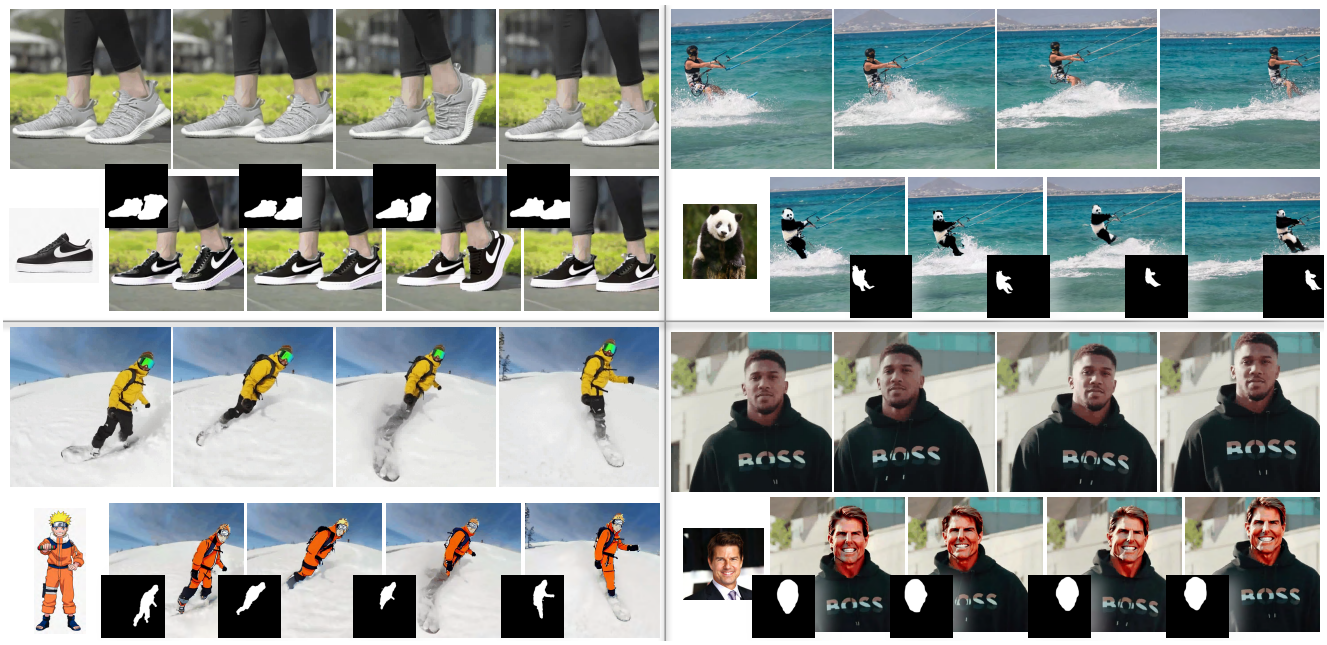

Figure 5. Source videos, Target images, *InvEdit* mask results and the *GenVideo* results. Top: *"A person wearing a ~~gray~~ black shoe."*, and *"A ~~man~~ panda rides a kite surfboard in deep waters."* Bottom: *"~~man~~ naruto skiing on snow"*, and *" ~~man~~ Tom Cruise walking down the street"*. We recommend zooming in for best viewing.

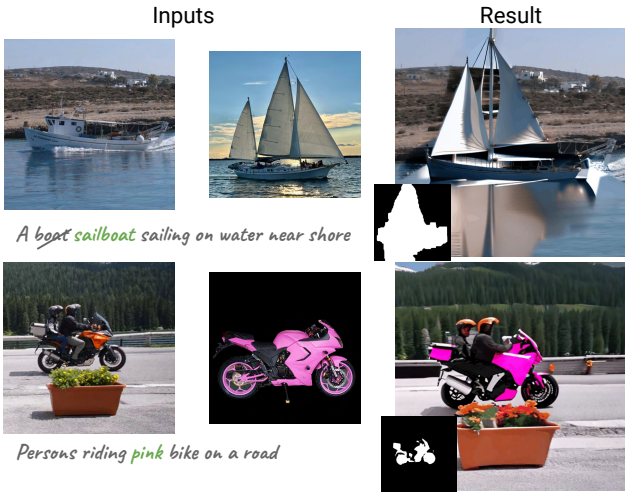

Figure 6. Results of *InvEdit* on zero-shot image editing showing its capability for target-objects of varying shape and size.

man $\longrightarrow$ *naruto*, *man* $\longrightarrow$ *panda* and *silver car* $\longrightarrow$ *white bus* examples, *InvEdit* mask is able to accurately identify the mask for *naruto*, *panda*, and the *white bus* respectively. Furthermore, our *latent correction* strategy is able to correct the UNet latents on the fly during inference.

Zero-shot image editing: Moreover, our approach is also capable of zero-shot image editing for objects of varying shapes and sizes. This can be observed in Fig. 6, which demonstrates object shape changes (like *boat* $\longrightarrow$ *sailboat*) in images. Additional examples and results are provided in the SM we have made available.

Table 1. Quantitative comparison of the state-of-the-art video editing methods. For the user study, *Text*, *Image*, and *Visual* refer to the average rank in target text alignment, target-image alignment, and visual quality of the edited video respectively.

Method	Model scoring metrics (↑)			User Study (↓)		
	CLIP-T	DINO	Temp	Text	Image	Visual
TAV [43]	0.238	0.236	0.957	3.6	3.3	4.2
StableVideo [5]	0.234	0.189	0.980	4.3	4.3	3.7
TokenFlow [9]	0.231	0.216	**0.985**	3.3	3.8	**2.1**
FateZero [31]	0.235	0.262	0.951	3.9	3.6	3.4
Make-a-Pro [46]	0.234	0.195	0.949	4.0	4.1	5.0
Ours *GenVideo*	**0.241**	**0.374**	0.967	**1.7**	**1.8**	2.3

4.3. Comparison with baselines

We do qualitative and quantitative comparisons with the following five existing approaches: 1) Tune-A-Video (TAV) [43], 2) StableVideo [5], 3) TokenFlow [9], 4) FateZero [31] and 5) Make-A-Protagonist (MAP) [46]. We use target reference image only with our method and [46].

Qualitative Results: Fig. 7 shows a comparison of our *GenVideo* approach with the above baselines. The edits in TAV[43] are not target-image aware and cannot be localized due to lack of mask. Hence, it doesn't guarantee fidelity to the target-image and also ends up modifying the background substantially. StableVideo is not target-image aware and hence, cannot generate indescribable features of the target-image. While it provides temporal consistency using NLA, these atlas are computed only using the source video and have no awareness of target-object. Hence, they

Figure 7. Qualitative comparison of *GenVideo* against SOTA. Shape aware methods demonstrate the ability to edit the shape of object. Target-aware methods condition the diffusion model on both target text prompt and target-image. *GenVideo* combines both target awareness and shape awareness, thus allowing it to infer shape and appearance based on the target-image and target text. This offers much superior alignment to the target text and target-image. First two columns show large shape variation while the the third shows results for same shape edits.

fail when the target-object has a different shape and size. TokenFlow also does not provide temporal consistency for target objects of varying shapes and sizes since the nearest neighbor field they use is based solely on the source video frames. Besides both TokenFlow and StableVideo are not target-image aware and hence, do not align well with the target-image. FateZero can change shapes to some extent but since their mask is again source video-based, they cannot create large variations in the shape (like car to bus). MAP also offers source video based mask guidance and hence, cannot substantially change shapes. Our results perform temporally consistent video editing when the target-object has a substantially different shape and size while maintaining fidelity to the target-image.

Quantitative Results: Table 1 shows a quantitative comparison between our approach and the state-of-the-art (SOTA) methods. We use trained CLIP [32] model to compute the average similarity of generated video frames to target text prompt which is denoted as CLIP-T in Table 1. To evaluate temporal consistency, we use CLIP model to compute average inter-frame similarity score for pairs of consecutive frames of the edited video. We denote this score as *Temp* in Table 1. To measure alignment with target-image, we use trained DINO [3] to compute the similarity between the video frames and the target-image. We also conduct a user study to evaluate the edited videos on three metrics: target text alignment, target-image alignment, and visual quality of the edited video. We asked 20 subjects

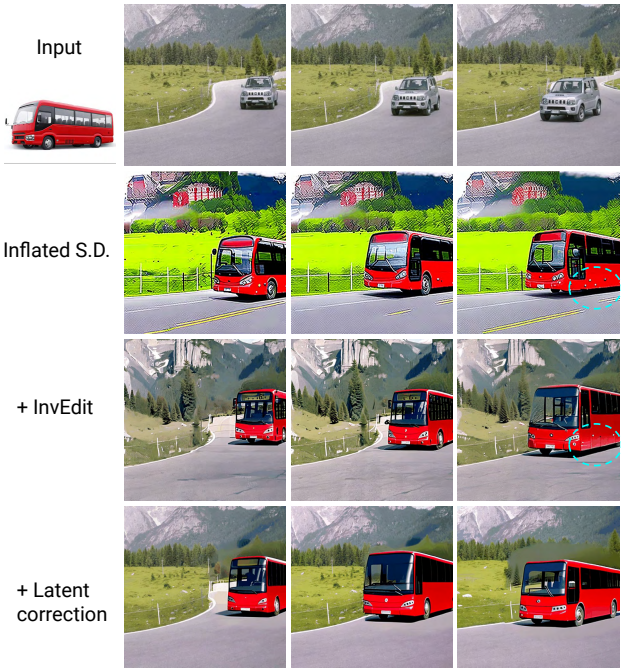

Figure 8. Ablation study of the *GenVideo* method. Prompt: *Driving a ~~silver car~~ red bus down a mountain road.*

to rank the edited videos for the six approaches (*i.e.,* ours and the five baselines mentioned above) on the above three metrics by asking each subject 9 sets of comparisons. The results are tabulated in Table 1. These results show our proposed approach outperforms existing approaches in both target text and target-image alignment. This is reflected in higher CLIP-T and DINO scores as well as in the user study where *GenVideo* achieves the best average rank for text and image alignment. *TokenFlow* and *StableVideo* achieved a marginally better score in inter-frame temporal consistency (measured by *Temp*) but this is because they use the source video ground truth for enforcing temporal consistency. This works well for consistency but the output videos do not align well with the target-image as reflected in much lower DINO scores. In the user study, *TokenFlow* achieved the best rank for visual quality outperforming our approach by a small margin. Again, this is because of the source video's flow-based feature propagation which leads to good quality videos even though they rank lower in the text and image alignment. Our results show that our approach outperforms the other approaches in text and image alignment while being close to the state-of-the-art in temporal consistency and visual quality. Note that to be fair we do not expect to be the best in temporal consistency and quality because we are creating novel views of the target object from a single target reference image. Rather, the strength of our method lies in being able to create target-image and shape aware masks for localized video edits and enforce temporal consistency in masks in a training-free fashion while ensuring image and

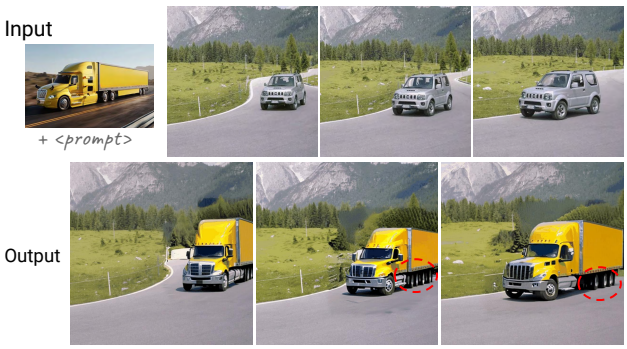

Figure 9. Limitation of *GenVideo* in fine-grained editing of videos. Prompt: *Driving a ~~silver car~~ yellow truck down the mountain road.*

text alignment.

4.4. Ablations

We show the ablations for our *InvEdit* and *latent correction* strategies in Fig. 8. The first row shows the inputs, *i.e.,* target-image of a red bus and the source video of a silver car driving down a mountain road. The second row shows the edited video frames with a finetuned inflated SD-unCLIP model with target-image guidance without the *InvEdit* mask. Note that without mask localization, the fidelity of the edited bus is of low quality and the background gets substantially modified. In third row, using *InvEdit* mask localization, we notice that the fidelity of the bus to the target-image is improved and the background is not substantially modified. However, we notice that in the third column, the wheel of the bus disappears. Our *latent correction* strategy is able to correct such inconsistencies by imposing temporal self-consistency between frames and preserving the background as shown in the fourth row.

5. Conclusion

We introduce *GenVideo*, a pipeline for target-image and shape aware video editing using image diffusion models. The proposed pipeline allows localized edits with target-objects using the proposed *InvEdit* mask and enforces temporal consistency between the frames using *latent correction* strategy. Results show that *GenVideo* outperforms existing methods both qualitatively and quantitatively on the video editing task. **Limitations and Future Work:** The underlying SD-unCLIP model may have limitations regarding the quality and diversity of the generated content, thus affecting the edit quality. The *latent correction* approach for inter-frame temporal consistency may not completely eliminate fine-grained inconsistencies, especially for complex objects like a truck with several wheels as shown in Fig. 9. Furthermore, similar to previous other methods [31, 43], our method cannot generate entirely new motion like changing motion from driving to flying. As part of further work, we aim to test our approach with image conditioned video diffusion models like [10].

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, 2022. 1

[2] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *ECCV*, 2022. 1, 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 7

[4] Duygu Ceylan, Chun-Hao P. Huang, and Niloy J. Mitra. Pix2video: Video editing using image diffusion. In *ICCV*, 2023. 1, 2

[5] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *ICCV*, 2023. 1, 2, 3, 6

[6] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *ICLR*, 2023. 1, 2, 4

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021. 1

[8] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *ICCV*, 2023. 1, 2

[9] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing, 2023. 1, 2, 3, 6

[10] 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. 2023. 8

[11] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control, 2022. 1, 2

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1, 2

[13] 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, 2022. 2, 3

[14] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers, 2022. 2

[15] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. *CVPR*, 2023. 2

[16] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv*, 2023. 2

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv*, 2023. 4

[18] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *ICCV*, 2023. 1

[19] Yao-Chih Lee, Ji-Ze Genevieve Jang Jang, Yi-Ting Chen, Elizabeth Qiu, and Jia-Bin Huang. Shape-aware text-driven layered video editing demo. *arXiv preprint arXiv:2301.13173*, 2023. 3

[20] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv*, 2023. 4

[21] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control, 2023. 1, 2

[22] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jinren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *CVPR*, 2023. 2

[23] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. 2

[24] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors, 2023. 1, 2

[25] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 1

[26] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022. 1

[27] Aaron Van Den Oord et al. The frobnicatable foo filter, 2017. Neural discrete representation learning. In NeurIPS. 2

[28] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *SIGGRAPH*, 2023. 1

[29] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021. 2

[30] 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*, 2023. 1

[31] 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 *ICCV*, 2023. 1, 2, 4, 6, 8

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 7

[33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 1, 2, 3, 4, 5

[34] A. Revanur, D. Basu, S. Agrawal, D. Agarwal, and D. Pai. Coralstyleclip: Co-optimized region and layer selection for image editing. In *CVPR*, 2023. 2

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3

[36] 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. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 1, 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, 2022. 1, 2

[38] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2023. 2, 4

[39] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *CVPR*, 2023. 5

[40] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 5

[41] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023. 1

[42] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 2

[43] 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 *ICCV*, 2023. 1, 2, 3, 5, 6, 8

[44] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv*, 2022. 2

[45] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2

[46] Yuyang Zhao, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. Make-a-protagonist: Generic video editing with an ensemble of experts, 2023. 1, 2, 3, 4, 6

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