AVID: <u>A</u>ny-Length <u>V</u>ideo <u>I</u>npainting with <u>D</u>iffusion Model

Zhixing Zhang[1]   Bichen Wu[2]   Xiaoyan Wang[2]   Yaqiao Luo[2]   Luxin Zhang[2]

Yinan Zhao [2]   Peter Vajda[2]   Dimitris Metaxas[1]   Licheng Yu[2]
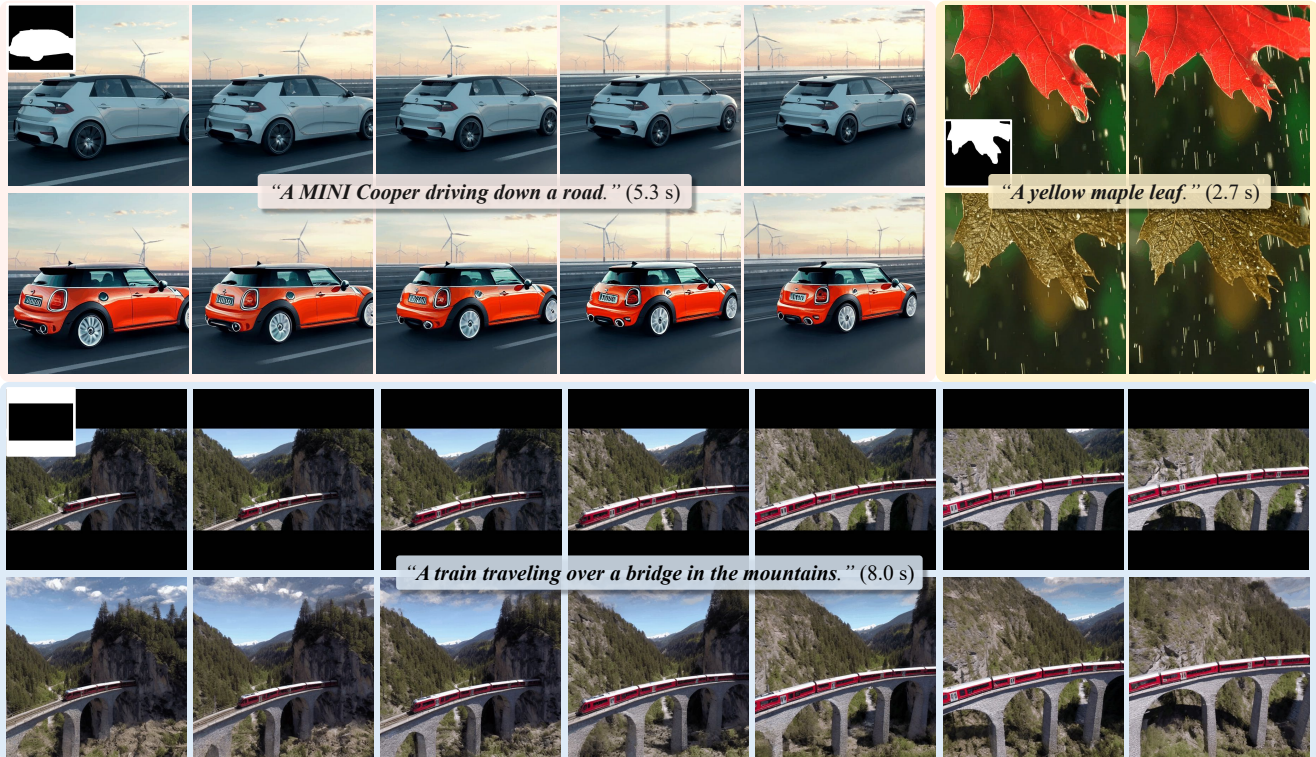
[1]Rutgers University   [2]GenAI, Meta

Figure 1. **Video inpainting.** We introduce a video inpainting method versatile across a spectrum of video durations and tasks. Displayed frames are uniformly selected from videos of different lengths. The first row in the figure contains the source videos and the target regions, while the bottom row shows the results. The caption in the middle represents the language guidance and duration for each video.

Abstract

Recent advances in diffusion models have successfully enabled text-guided image inpainting. While it seems straightforward to extend such editing capability into the video domain, there have been fewer works regarding text-guided video inpainting. Given a video, a masked region at its initial frame, and an editing prompt, it requires a model to do infilling at each frame following the editing guidance while keeping the out-of-mask region intact. There are three main challenges in text-guided video inpainting: (i) temporal consistency of the edited video, (ii) supporting different inpainting types at different structural fidelity levels, and (iii) dealing with variable video length. To address these challenges, we introduce Any-Length Video Inpainting with Diffusion Model, dubbed as AVID. At its core, our model is equipped with effective motion modules and adjustable structure guidance, for fixed-length video inpainting. Building on top of that, we propose a novel Temporal MultiDiffusion sampling pipeline with a middle-frame attention guidance mechanism, facilitating the generation of videos with any desired duration. Our comprehensive experiments show our model can robustly deal with various inpainting types at different video duration ranges, with high quality[1].

---

[1]More visualization results are made publicly available here.

1. Introduction

Past years have witnessed remarkable advancements in image inpainting [10, 38, 68], which infills a given masked region in an image. Owing to the recent explosive evolution of text-to-image (T2I) generative models, the content infilling becomes even more flexible by conditioning on a given textual description [46, 68]. On the other hand, text-to-video (T2V) generation is also developing rapidly [26]. This leads to a compelling inquiry: can we also harness this prowess for text-guided video inpainting?

Imagine being given a video of "a car traversing a road" and one wants to replace this car with a MINI Cooper. Perhaps one of the easiest editing approaches is the user just simply clicking on the car in the first frame and writing down a short textual prompt *"MINI Cooper"*, then an advanced model congests this information and generates a spatially-seamless and temporally consistent moving MINI Cooper in the scene while not changing any other video parts. We believe such interaction will become a fundamental and popular video editing approach, thus decide to address this task as its pioneering work.

There are mainly three main challenges in video inpainting. First, the synthesized content must exhibit temporal consistency. If an entire object is rendered, its identity should persist throughout the video. For example, if the color of the car is changed to green, it is imperative that the hue remains consistent from start to finish, the car should remain the same shade of green, rather than transitioning from, say, neon green to a darker variant. Second, there are various editing types in video inpainting. For example, as in Fig. 1, one common inpainting type is object swap (*e.g.* "replace a car with a MINI Cooper", another type is re-texturing (*e.g.* "turn the leaf color from red to yellow"), and uncropping (*e.g.* "fill in the upper and lower region") is also popular. Different editing types require different levels of structural fidelity to the original video. For instance, in object swap, if the editing is to transform a masked person into a statue, the structure and motion of the original person indicated by the mask region should be maintained in the output video. In comparison, video uncropping entails filling in blank spaces to augment the view, with no guidance signal from the mask region. Third, an input video can be of variable length, and thus, we expect a good model should robustly handle any video duration.

In light of these challenges, we propose Any-Length Video Inpainting with Diffusion Model, dubbed as AVID, a unified framework tailored for video inpainting. Built upon a text-guided image inpainting framework, we integrate motion modules to ensure temporal coherence within the edited region. A structure guidance module is also incorporated, adaptable to the varying structural fidelity demands of different video inpainting tasks. Our innovation also includes a zero-shot generation pipeline, reinforced by

a middle-frame attention guidance mechanism, enabling the handling of videos across varying durations. With the proposed model, given a textual description for the desired modification, along with a mask on the initial frame of a video indicating the area to edit as in Fig. 1, we can adeptly modify the content of *variable* video length, aligning them with the desired narrative while conserving intricate details, maintaining temporal consistency, and leaving the region outside the mask unaltered.

To encapsulate our contributions: **(I)** We integrate motion modules [14] into a text-to-image inpainting model and optimize it on video sequences, thereby ensuring temporal consistency. **(II)** We propose a structure guidance module tailored for different sub-tasks, so that users can control different degrees of output's structural fidelity towards the input video depending on the task and editing need. **(III)** We incorporate a pioneering zero-shot generation technique, proficiently handling different video lengths without additional training. Concurrently, we introduce a middle-frame attention guidance methodology, ensuring temporal consistency even in elongated video sequences. **(IV)** Rigorous evaluation on various inpainting tasks, region areas, and video length demonstrates the robustness of our model.

2. Related Work

The success of diffusion models [23, 54, 56] have enabled advanced image generation [16, 42, 44, 45, 49]. These models leverage large-scale text-image datasets [50] to produce remarkable outcomes. Further, recent works utilized the pre-trained text-to-image models for image manipulation in response to natural language descriptions [4, 6, 12, 15, 18–20, 27, 30–33, 36, 37, 39, 47, 55, 57, 74]. Among them, customized image generation has particularly benefited from controlling the generated content via additional structure modules, *e.g.*, ControlNet [73], T2I-adapter [41], *etc*.

Such techniques are also influencing a traditional but popular image editing task – image inpainting. While generative adversarial networks [71, 72] have been mainly applied for this task, recent diffusion-based models [1, 48] are showing more impressive results. However, these models are limited in filling in the content using out-of-mask context only. A more flexible usage is adding text control [2, 3, 61, 68–70] that allow for text-guided image inpainting. Latent Blended Diffusion [3] proposed blending the generated and original image latents, Imagenator [61] and Diffusion-based Inpainting [46] fine-tune pre-trained text-to-image generation models with masked images as additional input, and SmartBrush [68] fine-tunes an additional mask prediction branch on object-centric datasets.

Extending the success of text-guided image inpainting to video domain presents unique challenges, especially in maintaining the temporal consistency in videos of arbitrary duration. The scarcity of high-quality, large-scale video

datasets [5, 8, 14, 17, 24–26, 53, 60, 64, 65] further complicates this task. Recent efforts have explored utilizing pre-trained image models for video editing, *e.g.* employing DDIM inversion [54] for consistent latents [7, 13, 43, 51, 62, 66]. Nevertheless, most approaches are proposed without considering an explicit mask input. Relying on the textual editing prompts only could easily change the undesired regions. As comparison, VideoComposer [63] takes the masked frames as input and addresses video inpainting as one of their editing tasks, yet its constraint of applying a uniform target region across all frames limits its flexibility and sacrifices the editing quality.

In this work, we introduce a simple and effective framework for text-guided video inpainting. Our approach incorporates motion modules into a pre-trained text-to-image diffusion model, ensuring ts temporal coherence, and integrates a structure guidance module to satisfy varying structural fidelity needs. Moreover, we present an inference pipeline capable of handling videos of any duration, paving the way for practical applications in real-world scenarios.

3. Methods

For a video of arbitrary duration, a mask region, and an editing prompt, our objective is to fill in the indicated region at each frame following the editing guidance, while keeping the out-of-mask video portion unchanged. We introduce our method, AVID, as depicted in Fig. 2. Our approach is built on top of a diffusion-based text-guided image inpainting model [46], then adapt it to video inpainting by inflating the base model and learning the motion modules (Sec. 3.2). We further add a structure guidance module to meet different inpainting types and structural fidelity requirements (Sec. 3.3). Additionally, we propose a zero-shot inference pipeline to deal with videos of varying lengths (Sec. 3.4).

3.1. Preliminaries

The diffusion model is defined to approximate the probability density of training data by reversing the Markovian Gaussian diffusion processes [23]. Consider an input image x_0, we conduct a forward Markov process described as:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(\sqrt{1 - \beta_t} x_{t-1}, \beta_t \mathrm{I}\right), \quad (1)$$

where $t = 1, \ldots, T$ indicates the number of diffusion steps, with β_t controlling the noise level at each step. A neural network ϵ_θ learns to reverse this process, approximating noise ϵ_t to restore x_{t-1} from x_t using the relation $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t\right)$, with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, as per [23]. For conditional diffusion, in our case, text-guided inpainting, we introduce conditions into ϵ_θ without altering the process. Our training objective can be formulated as:

$$\mathcal{L} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)}\left[\|\epsilon - \epsilon_\theta(x_t, t, \mathbf{c})\|_2^2\right], \quad (2)$$

where $\mathbf{c}$ denotes the conditional inputs. In our case, $\mathbf{c} = (x_m, m, \tau_\theta(y))$, where m is a binary mask indicating the region to modify, $x_m = x_0 \odot (1 - m)$ is the region to preserve, y represents the corresponding textual description, while $\tau_\theta(\cdot)$ embodies a text encoder that transposes the string into a sequence of vectors. Efficient sampling approaches, such as DDIM [54] or PNDM [34], and classifier-free guidance [22] can be applied during inference.

3.2. Text-guided Video Inpainting

Given a video $v_0 = \{x_0^i\}_{i=1}^{N}$ and a mask region at its first frame m^0, we seek to edit the target region at each frame, aligning its content with the given text prompt y. In order to predict a precise editing region at each frame, we first propagate m^0 to every following frame and obtain a mask sequence $m = \{m^i\}_{i=1}^{N}$. This is achieved by applying XMem [9] to track the masked region through the whole video, which we found simple and precise for editing types like object swap and re-texturing. For uncropping, we simply use the same to-fill region for all frames.

To pursue good temporal consistency across video frames, we follow AnimateDiff [14] to inflate an image diffusion model by converting its 2D layers to pseudo-3D and adding additional motion modules to learn the temporal correlations between frames. As described in Eq. (1) to obtain $v_t = \{x_t^i\}_{i=1}^{N}$, the video frames undergo a forward diffusion process individually. Then our objective for video inpainting becomes:

$$\mathcal{L} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)}\left[\|\epsilon - \epsilon_\theta(v_t, t, \mathbf{c})\|_2^2\right], \quad (3)$$

with $\mathbf{c} = (v_m, m, \tau_\theta(y))$ and $v_m = \{x_0^i \odot (1 - m^i)\}_{i=1}^{N}$. We optimize only the motion modules to retain the generative capabilities of the pre-trained model, as in Fig. 2 (a).

3.3. Structure Guidance in Video Inpainting

In practice, as in Fig. 1, video inpainting involves different editing types, with respect to different structural fidelity. For example, re-texturing requires the structure of the source video to be preserved, *e.g.*, converting the material of the person's coat to leather, as in Fig. 4, while uncropping has no such requirement. This inspires us to introduce an additional structure guidance module to our model. Following the design of ControlNet [73], we fix the parameters of ϵ_θ and proceed to train the structure-conditioned module, denoted as $\mathbf{s}_\theta$, as in Fig. 2 (b). We employ a structure information extractor $\mathcal{S}$ to obtain structure condition on each frame $s = \{s^i\}_{i=1}^{N}$, where $s^i = \mathcal{S}(x_0^i)$. The structure guidance module outputs $\mathbf{c}_s = \mathbf{s}_\theta(v_t, t, \mathbf{c}, s)$, composed of 13 feature maps $\{h_i\}_{i=1}^{13}$ at 4 different resolutions. These feature maps are then integrated into the skip connections and the output of the middle block of ϵ_θ, contributing to the generation process. Therefore, the training objective for
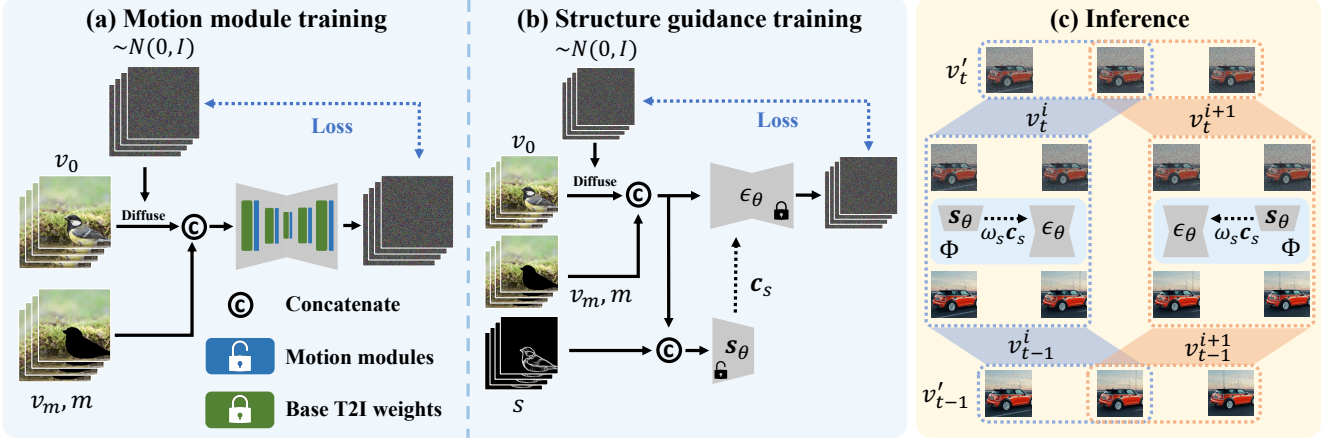
Figure 2. **Overview of our method.** In the training phase of our methodology, we employ a two-step approach. (a) Motion modules are integrated after each layer of the primary Text-to-Image (T2I) inpainting model, optimized for the video in-painting task via synthetic masks applied to the video data. (b) During the second training step, we fix the parameters in the UNet, ϵ_θ, and train a structure guidance module s_θ, leveraging a parameter copy from the UNet encoder. During inference, (c), for a video of length N', we construct a series of segments, each comprising N successive frames. Throughout each denoising step, results for every segment are computed and aggregated.

this phase is formulated as:

$$\mathcal{L} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)} \left[\|\epsilon - \epsilon_\theta \left(v_t, t, \mathbf{c}, \mathbf{c}_s \right)\|_2^2 \right], \quad (4)$$

with s_θ parameters being optimized.

During inference, the structural fidelity is modulated by a scaling factor ω_s applied to $\mathbf{c}_s$, with the formulation $\epsilon_\theta \left(v_t, t, \mathbf{c}, \mathbf{c}_s \cdot \omega_s \right)$. Here, $\mathbf{c}_s \cdot \omega_s = \{h_i \cdot \omega_s\}_{i=1}^{13}$, where the scaling factor is applied to each feature map individually. A higher ω_s leads to better structural fidelity.

3.4. Zero-shot Inference for Long Videos

Although motion modules (implemented as temporal self-attention layers) can take videos with varying numbers of frames theoretically, it suffers from drastic quality degradation when generating a longer video of more frames than training [14]. We show such quality degradation in *supplementary material*. Inspired by MultiDiffusion [4] generating a high-resolution image seamlessly consisting of multiple patches, we extend such idea into the temporal axis, making it work effectively to deal with a video of any length. We further propose a novel middle-frame attention guidance mechanism, to keep the identity unchanged throughout the video.

Temporal MultiDiffusion. We denote our model learned from Sec. 3.2 and Sec. 3.3 as $\Phi_{\{\epsilon_\theta, s_\theta\}}$, such that $v_{t-1} = \Phi(v_t)$. Given a video with N' frames longer than training video length N, we first segment such longer video v_t' into overlapping clips, achieved by applying a sliding window of N frames with a stride of $\mathbf{o}$. This process partitions v_t' into a series of clips $\{v_t^i\}_{i=1}^n$, where each v_t^i contains N frames, and $n = \lceil \frac{N'-N}{\mathbf{o}} \rceil + 1$ is the total number of clips. Applying our model once on each clip, we can get one-step denoising $\{v_{t-1}^i\}_{i=1}^n$, where $v_{t-1}^i = \Phi(v_t^i)$.
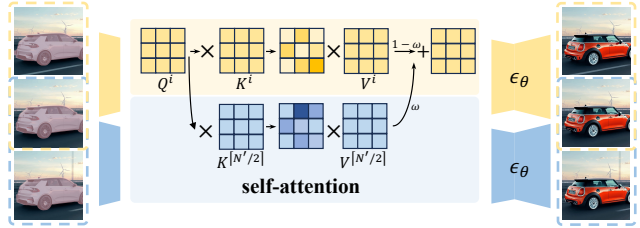

Figure 3. **Middle-frame attention guidance.** At inference, during each denoising step and within every self-attention layer, we retain the $K^{\lceil N'/2 \rceil}$ and $V^{\lceil N'/2 \rceil}$ values from the frame in the middle of the video. For the video's i^{th} frame, we utilize its pixel queries, denoted as Q^i, to compute an auxiliary attention feature map. This is subsequently fused with the existing self-attention feature map within the same layer.

For any given frame $v_t'[k]$ within v_t', we identify the set of clip indices $\mathbb{S}_k$ that contain this frame. Subsequently, for each index i in $\mathbb{S}_k$, the corresponding frame from $v_t'[k]$ is mapped to the j-th frame in v_t^i, denoted as $v_t^i[j]$. Similar to [4], we determine the frame $v_{t-1}'[k]$ in the output video by averaging the results from all relevant clips:

$$v_{t-1}'[k] = \frac{1}{\|\mathbb{S}_k\|} \sum_{i \in \mathbb{S}_k} v_{t-1}^i[j], \quad (5)$$

where $v_{t-1}^i[j]$ is the processed frame from v_t^i corresponding to $v_t'[k]$, and $\|\mathbb{S}_k\|$ is the cardinality of the set $\mathbb{S}_k$, indicating the number of times frame k is processed across all clips.

Middle-frame attention guidance. With Temporal Multi-Diffusion, we show the smoothness of our long video can be hugely improved Fig. 8. However, another critical issue arises – identity gradually changes from the initial frame to the last, and such issue could become more and more severe as longer video, as shown in Fig. 9.

Source ($N' = 32$, 5.3 s)   **Source** ($N' = 44$, 7.3 s)

Object swap: *"A young boy wearing a cowboy hat standing in a wheat field."*   **Re-texturing:** *"A women with blond hair walking through a wheat field."*

Source ($N' = 24$, 4.0 s)   **Source** ($N' = 48$, 8.0 s)

Uncropping: *"A Tiger walking through a jungle."*   **Re-texturing:** *"A women in glossy leather red coat walking through a greenhouse."*

Source ($N' = 36$, 6.0 s)   **Source** ($N' = 32$, 5.3 s)

Object swap: *"A Porsche car driving down a road."*   **Re-texturing:** *"A white duck swimming in a lake."*

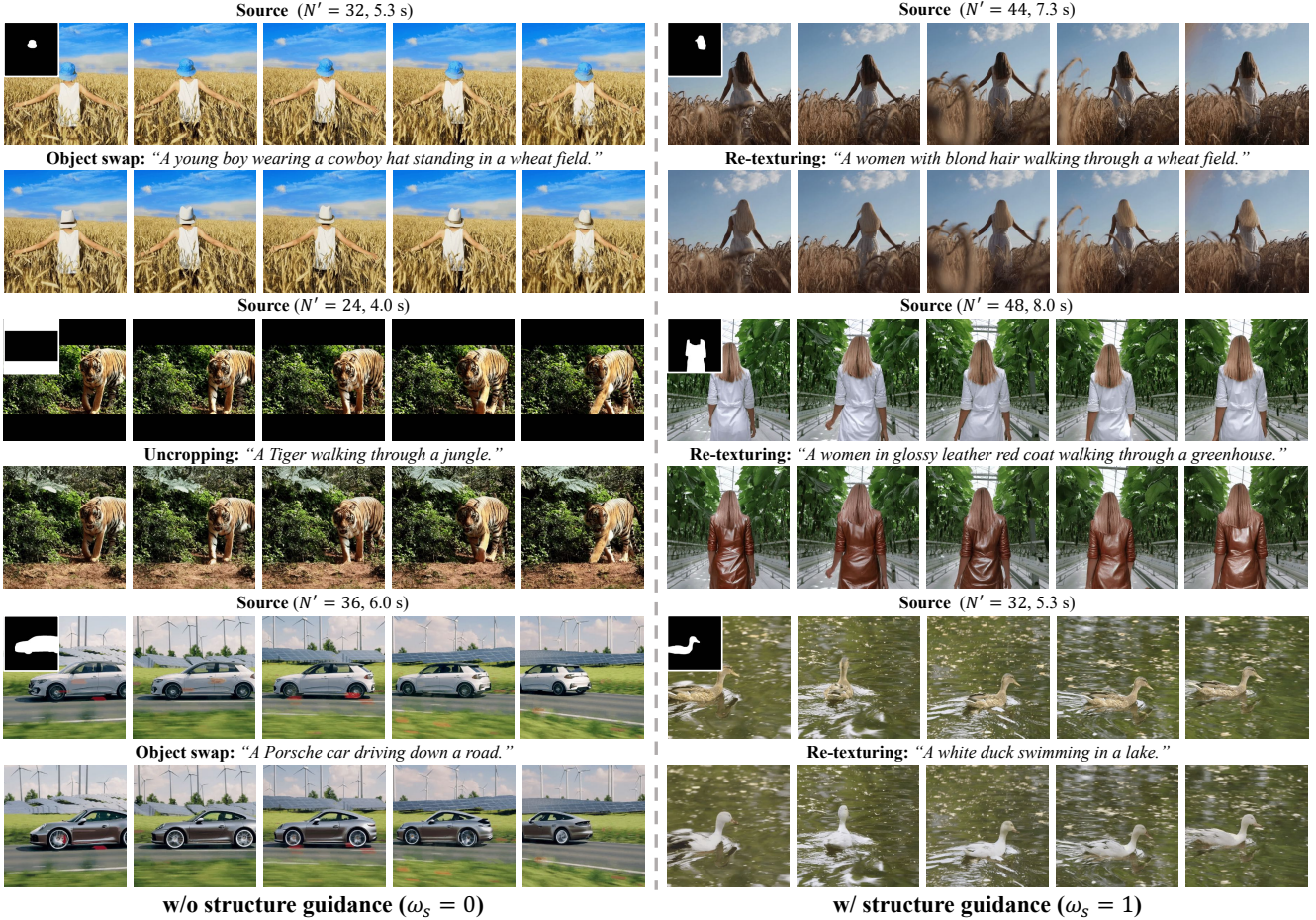**w/o structure guidance ($\omega_s = 0$)**   **w/ structure guidance ($\omega_s = 1$)**

Figure 4. **Editing on videos of different durations.** We employ our method on various videos and edit them for different tasks. We show the wide range of edits our approach can be used with different region sizes and video durations. Above each video, we note the number of frames, N', and the video duration.

To address this, we introduce a novel attention guidance mechanism, enforcing identity consistency across clips. As in Sec. 3.2, we inflate each self-attention layer to pseudo-3D self-attention layers. We denote the input to a self-attention layer of a video frame as ψ^i and W^Q, W^K, W^V as the attention weights, then $Q^i = W^Q \psi^i$, $K^i = W^K \psi^i$, and $V^i = W^V \psi^i$. For each self-attention layer, we use the features from the middle frame as guidance, $\psi^{\lceil N'/2 \rceil}$. We chose the middle frame due to the overall distance between it and the other frames being the smallest. Thus different clips can be more easily connected when their attention is regularized to the same reference frame, mitigating the identity shift issue. A qualitative study of the key frame selection is provided in *supplementary material*. The attention formulation for each frame is thus:

$$\text{Attention}(\psi^i) = \text{softmax}\left(\frac{Q^i K^{i^T}}{\sqrt{d}}\right) V^i \cdot (1 - \omega) + \\ \text{softmax}\left(\frac{Q^i K^{\lceil N'/2 \rceil^T}}{\sqrt{d}}\right) V^{\lceil N'/2 \rceil} \cdot \omega, \quad (6)$$

where ω is a parameter controlling the strength of the guidance signal. We provide a visual illustration of the attention guidance in Fig. 3.

4. Experiments

Implementation details. Our implementation is built upon a large-scale pre-trained inpainting Latent Diffusion Model (LDM) [46]. For training, we use the watermark-removed Shutterstock video dataset [52], with motion modules being trained using 16 frames at a 512×512 resolution using synthetic random mask. Subsequently, the parameters from the UNet encoder are transferred to the control module, which is then trained using the same dataset. We use Holistically-Nested Edge Detection (HED) [67] within the synthetic region as the structure guidance for the control module training. During the structure guidance module training, all parameters in the control module are optimized.

We evaluate our method on 125 videos (unseen during training). Objects designated for editing are pinpointed

Re-texturing: *"A purple car driving down a road."* **Object swap:** *"A flamingo swimming in a lake."*
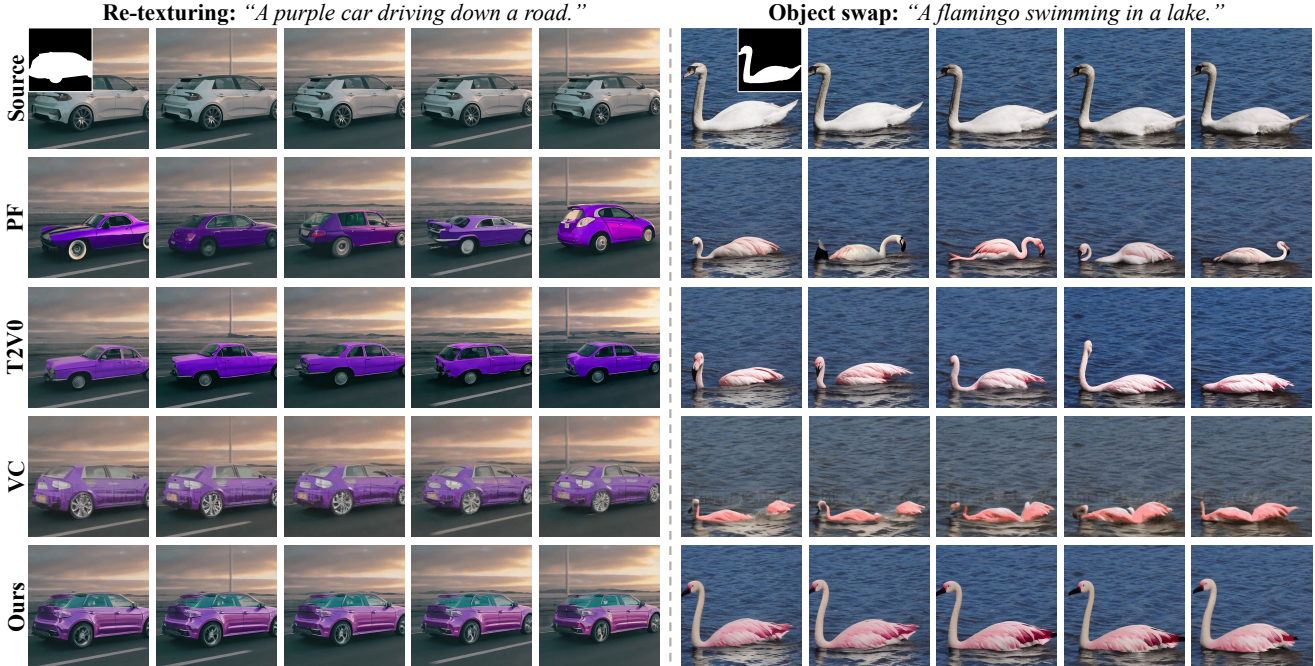
Figure 5. **Comparison of various methods.** We compare our method against several approaches, including per-frame inpainting using *text-to-image LDM inpainting* (PF) [46], *Text2Video-Zero* (T2V0) [28], *VideoComposer* (VC) [63]. All methods are evaluated using their default hyper-parameters as specified in either their corresponding publications or source codes. Each video in our experiments consists of 16 frames. Our proposed approach successfully edits the videos as intended while retaining the details outside the designated target region. Moreover, our method upholds the editing capabilities of the image in-painting model we utilized. Notably, our results demonstrate remarkable consistency, outperforming other methods in our comparison.

in these videos. On the initial frame of each video, object regions are identified using Grounding-DINO [35] and Segment-Anything (SAM) [29]. Subsequent segmentation for the entirety of the video is achieved using XMem [9]. We further prepare three editing types for evaluation: object swap, re-texturing, and uncropping. Llama [58] is then tasked with generating multiple editing prompts for every object within each video, tailored to diverse tasks. The FPS for both training and inference are set to 6.

In Sec. 3.3, we propose to use a scaling factor to flexibly adjust the structural fidelity. Specifically, object swap and uncropping requires no structural information from the original video, and thus we set the scaling factor for structural fidelity, $\omega_s = 0$. Conversely, re-texturing, leverages structural data for superior generative outcomes, we empirically set $\omega_s = 1$. For all long video inference, we set the middle-frame attention guidance scale ω to 0.3 and stride for temporal window $\mathbf{o} = 4$ unless otherwise noted.

4.1. Qualitative Results

To comprehensively evaluate the capabilities of our method, we test it on videos of various durations across different inpainting types. As shown in Fig. 4, our zero-shot inference approach as in Sec. 3.4 is capable of performing diverse editing types, catering to a wide range of mask sizes,

and dealing with variable video durations. Our method adeptly modifies the specified region without affecting the surrounding content. Additionally, it keeps the identity (color, structure, *etc*.) of the generated content consistent across video frames.

4.2. Comparisons

Task	Uncropping			Object swap			Re-texturing*		
Metric	BP	TA	TC	BP	TA	TC	BP	TA	TC
PF	43.1	31.3	93.6	41.4	31.1	92.5	41.4	31.2	92.4
T2V0	49.0	**31.4**	96.5	47.3	30.1	94.9	47.9	30.6	95.0
VC	55.7	31.2	96.4	71.0	**31.5**	96.5	64.5	**32.1**	95.5
Ours	**42.3**	31.3	**97.2**	**41.1**	31.5	**96.5**	**40.7**	32.0	**96.3**

Table 1. **Quantitative results.** We compare our method against several approaches, including per-frame in-painting (PF) using *Stable Diffusion In-painting* [46], *Text2Video-Zero* (T2V0) [28], and *VideoComposer* (VC) [63] on different video inpainting sub-tasks and evaluate generated results using different metrics, including background preservation (BP $\times 10^{-3}$, ↓ better), text-video alignment (TA, ↑ better), and temporal consistency (TC, ↑ better). * indicates structure guidance is applied for VC and our approach.

Qualitative comparisons. We present a comprehensive evaluation of our method against other diffusion-based video inpainting techniques, notably the per-frame inpaint-
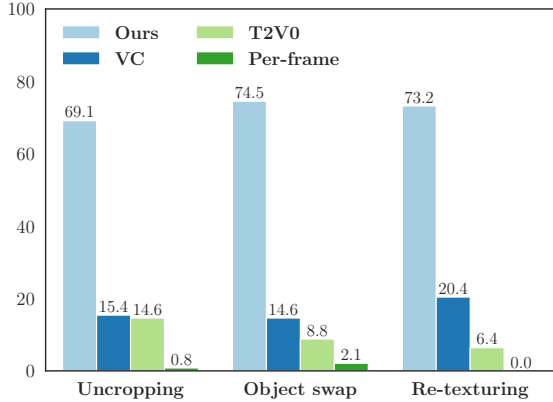
Figure 6. **User study.** In our user preference studies, we juxtaposed our method against per-frame in-painting techniques by evaluating prominent models such as *Diffusion-based Image Inpainting*[46], *Text2Video-Zero* (T2V0)[28], and *VideoComposer* (VC)[63], assessing their performances across various tasks.

ing techniques using inpainting LDM [46] and VideoComposer [63]. We apply our model on the same videos as in [59], which have provided results of VideoComposer [63]. Recognizing that video inpainting can be construed as a text-to-video generation with set boundary conditions, we also evaluate the training-free Text2Video-Zero [28] model on top of inpainting LDM as another comparison. To highlight the robustness of our foundational model, we limit our evaluations to videos spanning 16 frames, intentionally omitting the zero-shot long video inferences introduced in Sec. 3.4 for fair comparison.

Fig. 5 (left) compares the performance on re-texturing. Note both our model and VideoComposer [63] can apply the structural guidance. We let VideoComposer [63] integrate more cues (including the sketch maps, depth maps, and motion vectors) for better results, while our model only applies HED. Fig. 5 (right) proceeds object swap without structural guidance for all methods. Specifically, the comparison between VideoComposer [63] (row 4) and our model (row 5) reveals an extraneous color shift in the former, whereas the latter demonstrates impeccable background preservation. Furthermore, compared with the other two zero-shot methodologies, per-frame editing (row 2) and Text2Video-Zero (row 3) [28], our method exhibits far better temporal consistency. For example, our model can consistently keep the synthesized purple car shape in Fig. 5, while Text2Video-Zero [28] presents temporal inconsistencies.

Quantitative comparisons. *(a) Automatic Metric Evaluation.* Our model's performance is further quantified using three automatic evaluation metrics. Background preservation is measured using the L1 distance between the original and the edited videos within unaltered regions. The alignment of the generated video with the text description is evaluated using the CLIP-score [11, 21]. Temporal consistency
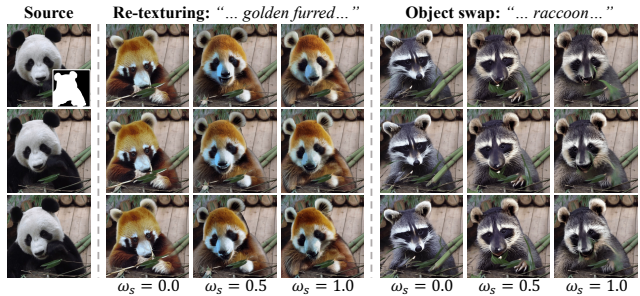

Figure 7. **Analysis of structure guidance.** We choose three evenly spaced frames from a video to highlight the impact of our structure guidance. To optimize results, the scale of our structure guidance must be tailored to each specific editing sub-task.

is assessed by computing the cosine similarity between consecutive frames in the CLIP-Image feature space, as per [11]. As shown in Tab. 1, our model exhibits excellent temporal consistency without compromising per-frame quality, as indicated by the text-video alignment scores. However, we acknowledge that CLIP score may not always correlate with human perception [40, 62], thus further conducted a user study. *(b) User study.* We evaluated our model's effectiveness via user study. Specifically, the annotators are presented with videos processed on three inpainting types: uncropping, object swap, and re-texturing. They were asked to judge the overall quality of the video inpainting based on temporal consistency, text-video alignment, and background preservation. Results illustrated in Fig. 6 indicate that our model was greatly favored in producing the best outcomes in all types – uncropping (69.1%), object swap (74.5%), and re-texturing (73.2%), demonstrating a consistent advantage over competing approaches.

4.3. Ablation Analysis

Effect of structure guidance. In Fig. 7, we exhibit the effects of varying the structure guidance scale, ω_s, during the editing of a video of 16 frames across different tasks. We highlight the first, middle, and last frames to demonstrate how structure guidance impacts the editing outcomes. For the re-texturing task, where the objective is to transform the color of the panda's fur to golden, a higher ω_s ensures that more motion and structural details from the original video are retained. Conversely, with $\omega_s = 0$, the motion of the generated golden-furred panda bears no correlation to the original video. In the object swap, the goal is to replace the panda with a raccoon. Yet, when a high structure guidance scale is employed, the generated raccoon retains the shape of the panda, leading to a mismatch with the textual instruction. In essence, the appropriate scale for structure guidance should be judiciously chosen based on the specific inpainting type the model is addressing.

Temporal MultiDiffusion. We conduct longer video experiments on 32-frame 6-FPS videos, while our base model
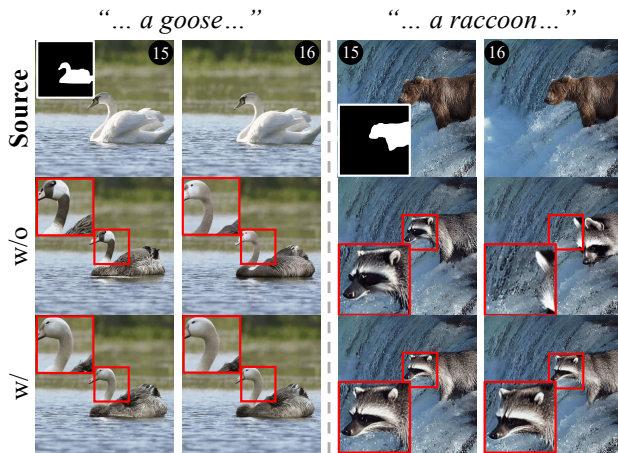
"... a goose..."   "... a raccoon..."

Figure 8. **Analysis of Temporal MultiDiffusion.** While direct sampling of two video clips (w/o) results in inconsistent content within the target area, temporal multi-diffusion sampling (w/) ensures the synthesis of longer videos with seamless transitions. We highlight part of the target region in each video to better illustrate the effectiveness of our approach.

is trained on video clips consisting of 16 frames. As in 3.4, we can directly apply our model to the longer video due to the flexibility of our motion modules, but it will result in drastic quality degradation Another naive approach is to apply our model twice on each non-overlapped 16 frames. However, this approach significantly changes the identity of the generated content within the target region between the two segments. In Fig. 8, we demonstrate the advantages of the Temporal MultiDiffusion pipeline we propose. We set attention guidance $\omega = 0$ in this experiment to better illustrate the effect of Temporal MultiDiffusion sampling. For illustration, we highlight the 15th frame (the end of the first segment) and the 16th frame (the beginning of the second segment). As observed, there is a sudden change between the segments using the naive approach, *e.g.*, there is a sudden shift in the texture pattern on the generated goose's neck and the position of the generated raccoon. In comparison, the content transitions smoothly between any two consecutive frames, with the Temporal MultiDiffusion module.

Effect of middle-frame attention guidance. We delve into the effect of the attention guidance weight, denoted as ω, as illustrated in Fig. 9. Our experimental videos are with 32 frames and 6 FPS. Setting $\omega = 0$ bypasses our proposed self-attention guidance mechanism. Although Temporal MultiDiffusion sampling ensures a smoother transition of content within the generated area, a noticeable identity shift persists throughout the video. For instance, with $\omega = 0$, the rendered MINI Cooper gradually transitions from red to dark red. Our proposed attention guidance significantly mitigates this issue, ensuring a consistent identity from the first frame to the last. Conversely, setting a too large value for $\omega = 1$ may introduce artifacts, as the keys and values in each self-attention layer Eq. (6) are dominated
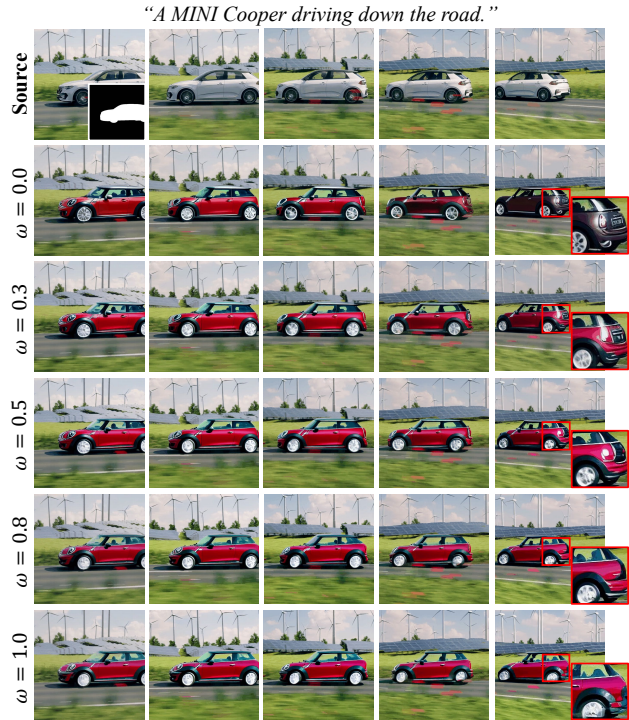

"A MINI Cooper driving down the road."

Figure 9. **Analysis of middle-frame attention guidance.** Adjusting ω yields editing results of differing quality.

by the middle frame signal.

Discussion. Our model performance is limited by the learned motion module quality. There are scenarios in which our model cannot generate content well especially when the editing prompt involves complicated actions, *e.g.*, "transforming the head of a horse from left to right" (in *supplementary material*). We believe a stronger motion module or a better text-to-video foundation model could further help improve the inpainting performance. Another promising direction for future exploration is to turn the hyperparameter of the structure guidance scale into a learnable parameter, controlled by the editing prompt.

5. Conclusion

We present AVID, a novel approach to address text-guided video inpainting. Our model incorporates motion modules for better temporal consistency, and a structure guidance module for better structural fidelity to the original video. At the sampling phase, we introduce a zero-shot inference pipeline, enabling our model to deal with videos with extended lengths. Additionally, we propose a simple middle-frame attention guidance mechanism, greatly improving the identity consistency across video frames. Rigorous experiments show the effectiveness and robustness of AVID.

References

[1] Titas Anciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J Mitra, and Paul Guerrero. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *CVPR*, pages 12608–12618, 2023. 2

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, pages 18208–18218, 2022. 2

[3] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 42 (4):1–11, 2023. 2

[4] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *ICML*, pages 1737–1752. PMLR, 2023. 2, 4

[5] 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 *CVPR*, pages 22563–22575, 2023. 3

[6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, pages 18392–18402, 2023. 2

[7] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *ICCV*, pages 23206–23217, 2023. 3

[8] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation, 2023. 3

[9] Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, 2022. 3, 6

[10] Yen-Chi Cheng, Chieh Hubert Lin, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, and Ming-Hsuan Yang. Inout: Diverse image outpainting via gan inversion. In *CVPR*, pages 11431–11440, 2022. 2

[11] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *ICCV*, pages 7346–7356, 2023. 7

[12] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 2

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

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

[15] Ligong Han, Ruijiang Gao, Mun Kim, Xin Tao, Bo Liu, and Dimitris Metaxas. Robust conditional gan from uncertainty-aware pairwise comparisons. In *AAAI*, pages 10909–10916, 2020. 2

[16] Ligong Han, Martin Renqiang Min, Anastasis Stathopoulos, Yu Tian, Ruijiang Gao, Asim Kadav, and Dimitris N Metaxas. Dual projection generative adversarial networks for conditional image generation. In *ICCV*, pages 14438–14447, 2021. 2

[17] Ligong Han, Jian Ren, Hsin-Ying Lee, Francesco Barbieri, Kyle Olszewski, Shervin Minaee, Dimitris Metaxas, and Sergey Tulyakov. Show me what and tell me how: Video synthesis via multimodal conditioning. In *CVPR*, pages 3615–3625, 2022. 3

[18] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023. 2

[19] Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Anastasis Stathopoulos, Xiaoxiao He, Yuxiao Chen, et al. Proxedit: Improving tuning-free real image editing with proximal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4291–4301, 2024.

[20] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[21] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 7

[22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 2, 3

[24] 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. 3

[25] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022.

[26] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2, 3

[27] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 2

[28] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 6, 7

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 6

[30] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, pages 1931–1941, 2023. 2

[31] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022.

[32] Di Liu, Xiang Yu, Meng Ye, Qilong Zhangli, Zhuowei Li, Zhixing Zhang, and Dimitris N Metaxas. Deformer: Integrating transformers with deformable models for 3d shape abstraction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14236–14246, 2023.

[33] Di Liu, Anastasis Stathopoulos, Qilong Zhangli, Yunhe Gao, and Dimitris Metaxas. Lepard: Learning explicit part discovery for 3d articulated shape reconstruction. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[34] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022. 3

[35] 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 preprint arXiv:2303.05499*, 2023. 6

[36] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. *arXiv preprint arXiv:2112.05744*, 2021. 2

[37] Zhi-Song Liu, Li-Wen Wang, Wan-Chi Siu, and Vicky Kalogeiton. Name your style: An arbitrary artist-aware image style transfer. *arXiv preprint arXiv:2202.13562*, 2022. 2

[38] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, pages 11461–11471, 2022. 2

[39] Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. *arXiv preprint arXiv:2305.16807*, 2023. 2

[40] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 7

[41] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2

[42] 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*, pages 16784–16804. PMLR, 2022. 2

[43] 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. *arXiv preprint arXiv:2303.09535*, 2023. 3

[44] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831. PMLR, 2021. 2

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

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 3, 5, 6, 7

[47] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 2

[48] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2

[49] 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. *NeurIPS*, 35:36479–36494, 2022. 2

[50] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2

[51] Chaehun Shin, Heeseung Kim, Che Hyun Lee, Sang-gil Lee, and Sungroh Yoon. Edit-a-video: Single video editing with object-aware consistency. *arXiv preprint arXiv:2303.07945*, 2023. 3

[52] Shutterstock. Stock images, photos, vectors, video, and music — shutterstock (https://www.shutterstock.com/). 5

[53] 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. 3

[54] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*. OpenReview.net, 2021. 2, 3

[55] Kunpeng Song, Ligong Han, Bingchen Liu, Dimitris Metaxas, and Ahmed Elgammal. Diffusion guided domain adaptation of image generators. *arXiv preprint arXiv:2212.04473*, 2022. 2

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

[57] Anastasis Stathopoulos, Ligong Han, and Dimitris Metaxas. Score-guided diffusion for 3d human recovery. In *CVPR*, 2024. 2

[58] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov,

Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 6

[59] John Vats. Exploring video-inpainting: A comparative analysis of videocomposer, 2023. 7

[60] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 3

[61] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *CVPR*, pages 18359–18369, 2023. 2

[62] Wen Wang, 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. 3, 7

[63] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023. 3, 6, 7

[64] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 3

[65] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pretraining for neural visual world creation. In *ECCV*, pages 720–736. Springer, 2022. 3

[66] 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*, pages 7623–7633, 2023. 3

[67] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, pages 1395–1403, 2015. 5

[68] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *CVPR*, pages 22428–22437, 2023. 2

[69] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *CVPR*, pages 18381–18391, 2023.

[70] Shiyuan Yang, Xiaodong Chen, and Jing Liao. Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In *ACMMM*, pages 3190–3199, 2023. 2

[71] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *CVPR*, pages 5505–5514, 2018. 2

[72] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *ICCV*, pages 4471–4480, 2019. 2

[73] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 2, 3

[74] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *CVPR*, pages 6027–6037, 2023. 2