AVID: Any-Length Video Inpainting with Diffusion Model Supplementary Material

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

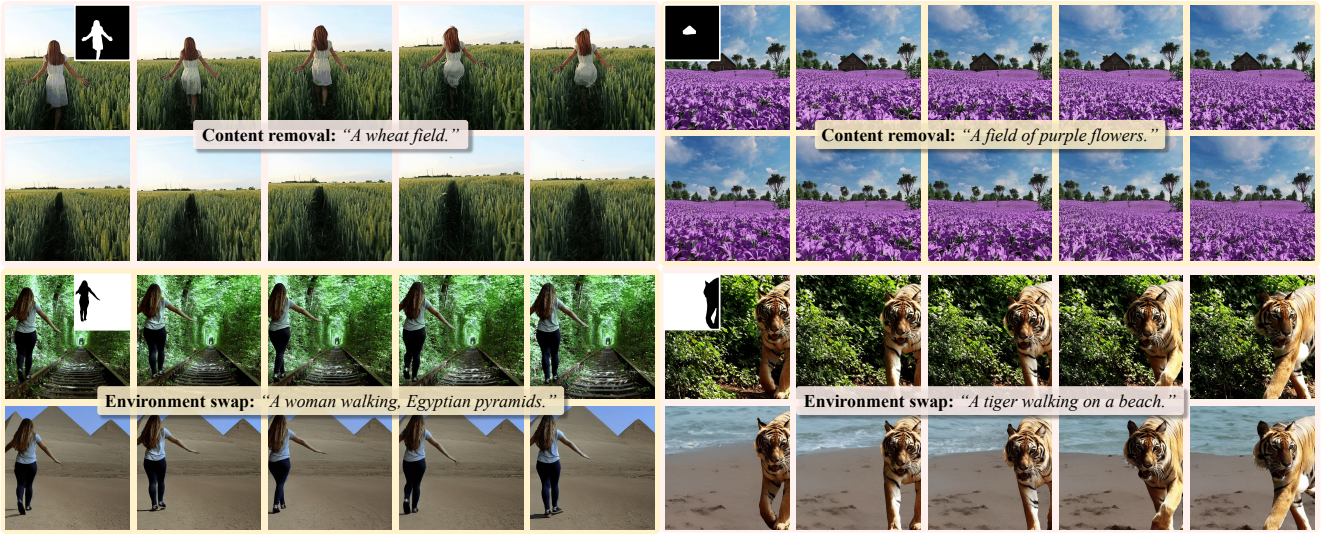

Figure A1. **Other applications.** We show how our approach can be applied to other video inpainting tasks, such as content removal and environment swap.

Overview

The supplementary material accompanying this paper provides additional insights and elaborations on various aspects of our proposed method. The contents are organized as follows:

- **Qualitative Results:** We showcase a broader range of qualitative results demonstrating the efficacy of every video inpainting type of AVID on videos of variable time duration. The results can be found in Appendix A of the supplementary material.
- **Application to Other Tasks:** Appendix B presents the application of AVID on other text-guided video inpainting types.
- **Test-Time Efficiency Analysis:** An in-depth analysis of the test-time efficiency of our method is provided in Appendix C.
- **More Comparative Analysis:** Additional comparative studies are detailed in Appendix D.
- **Ablation Study:** We extend the ablation analysis mentioned in the main paper (Appendix E).
- **Limitations:** Appendix F is dedicated to discussing the limitations and potential areas for improvement in our method.
- **Extension to Text-to-Video Generation:** We explore the application of our proposed sampling pipeline to the domain of any-length text-to-video generation. The results of this exploration are presented in Appendix G.

For a more immersive experience, we encourage readers to look at the results in video format, available here.

A. Qualitative Results

In this section, we present an extensive collection of qualitative results that demonstrate the capabilities of our proposed method, AVID. This includes both the examples showcased in the main paper and additional results, offering a comprehensive view of our method's performance in various scenarios.

To facilitate a more interactive and illustrative experi-

ence, these qualitative results are provided in video format. Readers are recommended to check these results in the first section of our accompanying webpage. This visualization provides a more nuanced understanding of the temporal and visual qualities of our video inpainting results, as well as a deeper insight into the effectiveness of AVID in practical applications.

B. Exploring Additional Inpainting Tasks

This section delves into the adaptability of our AVID method to a broader spectrum of video inpainting applications, specifically focusing on content removal and environment swapping. Our experiments illustrate the versatility and effectiveness of AVID in handling diverse inpainting scenarios.

The experiments in this section are conducted using videos with $N' = 24$ frames, corresponding to a duration of 4 seconds. We set $\omega_s = 0.0$ in these experiments, meaning no structure guidance is applied. The results are visually represented in Fig. A1 of the supplementary material.

B.1. Content Removal

Video inpainting has been narrowly defined as content removal in previous literature [4, 10, 11]. However, with diffusion models, we can enable multiple new inpainting tasks as introduced, which traditional approaches cannot handle. This work focuses mainly on content generation/editing guided by a given prompt and mask. Nevertheless, our model does also support "content removal".

The primary goal in content removal is to eliminate a specific object or element from the video while maintaining seamless integration with the surrounding content. As demonstrated in the top block of Fig. A1, our method initiates this process by generating a mask sequence targeting the object to be removed. Subsequently, we input a prompt such as "A wheat field" that describes the desired background, omitting any mention of the target object. This strategy enables our model to effectively remove the object, replacing it with contextually coherent content that blends seamlessly with the surrounding area. We further qualitatively evaluate our model on the popular DAVIS [5] dataset to better illustrate the ability of our method, as shown in Fig. A2.

B.2. Environment Swap

The environment swap task involves altering the background or surrounding environment of a subject in the video. Our method showcases its capability in environment swapping in the bottom block of Fig. A1. By selecting the complement of the target region as the editing area, we can effectively modify the video's background. Through prompts describing the new environment, such as "Egyptian pyramids", our model can adeptly transform the surrounding setting, demonstrating its robustness in adapting to various inpainting contexts.

B.3. Multiple Regions Inpainting

Our method is not limited to inpainting one specific region in a video. Independent inpainting can be achieved sequentially for multiple objects. As shown in Fig. A3, we conduct re-texturing on two different regions, *i.e*. coat, and hair, further demonstrating the effectiveness of our method in real-world applications.

C. Test-time Efficiency

In this section, we extend the analysis to evaluate the test-time efficiency of our proposed Temporal MultiDiffusion pipeline. For simplicity, we bypass the structure guidance in this analysis. Building upon the foundation discussed in Sec. 3.2 of our main paper, our approach inflates an image inpainting diffusion model, inspired by AnimateDiff [3]. This is achieved by transforming 2D layers into pseudo-3D format, allowing independent processing of each frame. To capture temporal correlations, we incorporate motion modules, realized through pixel-wise temporal self-attention.

Considering a video sequence with N' frames, a direct inference approach using all N' frames simultaneously leads to a temporal complexity of $\mathcal{O}(N'^2)$. The spatial complexity for attention layers, both self and cross attention, is $\mathcal{O}((HW)^2)$, with H and W being the spatial dimensions. Thus, our base model exhibits a computational complexity of $\mathcal{O}((HW)^2 \times N'^2)$.

The Temporal MultiDiffusion pipeline, however, segments the video into n parts, each comprising N frames, where $n = \lceil \frac{N'-N}{\mathbf{o}} \rceil + 1$ and $\mathbf{o}$ represents the stride. This segmentation allows for independent calculation of each segment at every denoising step, reducing the temporal complexity to $\mathcal{O}(N^2 \times n)$. With the incorporation of our middle-frame attention guidance mechanism, the spatial self-attention calculation effectively doubles, leading to a total temporal complexity of $\mathcal{O}(2(HW)^2 \times (N^2 \times n))$. Notably, when $N' >> N$, the complexity of our approach approximates to $\mathcal{O}((HW)^2 \times N')$, significantly more efficient than direct inference with N' frames. Additionally, our pipeline necessitates the calculation of only the segment containing the middle frame for initial attention guidance, while other segments can be processed in parallel, leveraging multi-GPU setups to expedite the process and mitigate potential GPU memory overflows in practical applications.

D. More Comparison.

In this section, we engage in comparative experiments with TokenFlow [2], a state-of-the-art video editing method, to demonstrate the effectiveness of our proposed approach. We are following the methodology outlined in Sec. 4.2 in

Figure A2. **Content removal on DAVIS [5] dataset.** We apply our method for content removal on different videos in the DAVIS [5] dataset. All frames of each video are passed to our model. Frames shown in the figure are evenly distributed in each video. We use prompts "a field", "a grassland", and "a park" respectively for these videos.

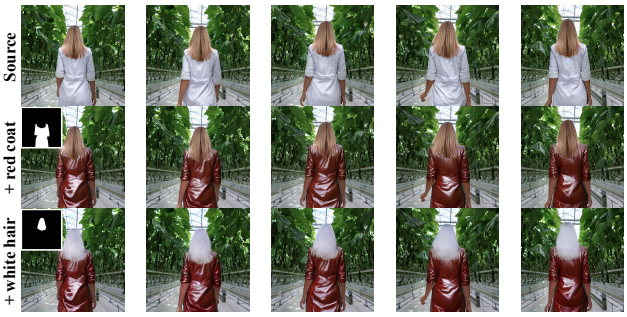

Figure A3. **Multiple objects inpainting.** We show how our approach can be applied to inpaint multiple objects in a video independently.

our main paper, we assess the performance of TokenFlow against our method, particularly focusing on tasks such as re-texturing and object swapping. Our evaluation utilizes the same set of videos and automatic metrics detailed in Sec. 4.2. This comparative study aims to provide an objective and quantifiable measure of each method's capabilities.

Despite TokenFlow's advanced editing capabilities, our experiments reveal a significant shortfall in its background preservation ability. Specifically, in the context of object swapping, TokenFlow scores 93.3 compared to our method's 41.1. Similarly, in re-texturing tasks, TokenFlow scores 90.8 versus our 40.7. This disparity can be attributed to TokenFlow's reliance on language-based guidance for determining the editing region, rather than using an explicit mask sequence. This approach undermines the method's suitability for precise video inpainting tasks, where maintaining contextual consistency is paramount.

An additional consideration is TokenFlow's use of DDIM inversion [6] for temporal consistent latent initialization. In contrast, our method employs initialization from a standard Gaussian distribution. This fundamental difference in initialization strategy highlights TokenFlow's limitations in tasks where no guidance can be obtained from the source video in the target region, such as video uncropping.

Task	Object swap			Re-texturing*		
Metric	BP	TA	TC	BP	TA	TC
TF	93.3	31.5	**97.5**	90.8	**32.2**	**97.8**
Ours	**41.1**	**31.5**	96.5	**40.7**	32.0	96.3

Table A1. **Quantitative results.** We compare our method against TokenFlow (TF) [2] on different video generative fill sub-tasks and evaluate generated results using different metrics, including background preservation (BP $\times 10^{-3}$, $\downarrow$ better), text-video alignment (TA, $\uparrow$ better), and temporal consistency (TC, $\uparrow$ better). * indicates structure guidance is applied for our approach.

E. More Ablation Analysis

E.1. Temporal MultiDiffusion

This section aims to evaluate the efficacy of our Temporal Multi-Diffusion sampling pipeline, especially in handling videos of varying durations. As discussed in Sec. 3.4 of the main paper, our model, while versatile, faces challenges in maintaining quality when dealing with frame counts different from those used in training. We address this issue by comparing the performance of our model using the Temporal Multi-Diffusion pipeline against its direct application on videos of different lengths.

Following the framework of AnimateDiff [3], our model incorporates sinusoidal position encoding [7] within each temporal self-attention motion module. This encoding is pivotal in making the network aware of the temporal positioning of frames within a video clip. During training, we set the maximum length of this encoding to 24 frames.

For our comparative analysis, we standardized the video length to 24 frames. This approach allows for a balanced evaluation of our method against the baseline model. Notably, in these tests, we disabled the middle-frame attention guidance to ensure fairness in comparison.

As depicted in Fig. A4, we observed that direct inference with 24 frames resulted in a significant decline in generation

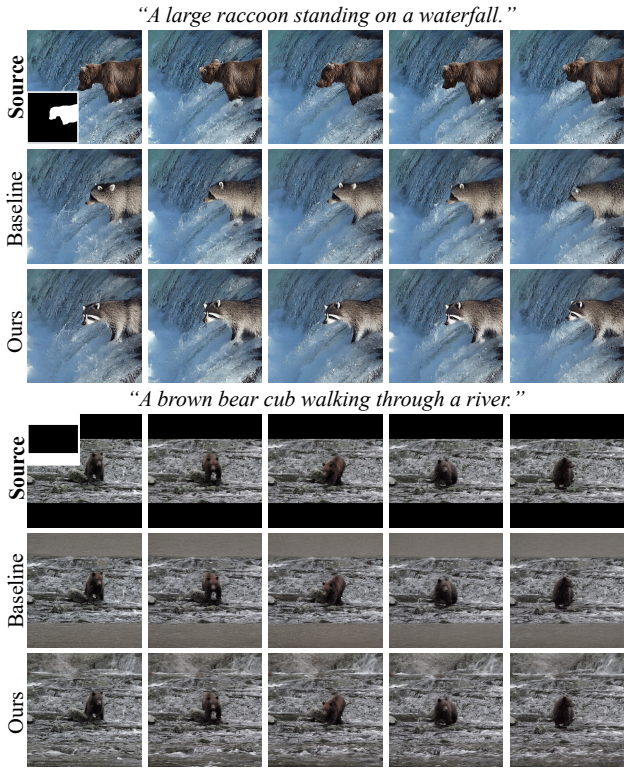

Figure A4. **Ablation analysis of temporal multi-diffusion.** When we directly apply our model to video generative fill tasks of longer durations (specifically, 24 frames), it does not produce out-of-distribution results (row 2 and row 5). However, there's a noticeable decline in the quality of the filled content when the length of the inference video differs from the training setup, the ear of the generated raccoon in the first case (row 2). In the second case (row 5), the model fails to fill-in the target region with content that can seamlessly blend in with the rest area. In contrast, our method (row 3 and row 6) effectively addresses this issue, synthesizing high-quality content even for extended-duration videos.

quality. In stark contrast, the adoption of our Temporal MultiDiffusion pipeline markedly improved performance. This pipeline effectively preserved the model's generative quality, showcasing its robustness and adaptability to different video durations without compromising the visual fidelity of the generated content.

E.2. Middle-frame Attention Guidance

In this section, we conduct an ablation study to underscore the efficacy of the middle-frame attention guidance mechanism introduced in our method. This study is pivotal in demonstrating how our approach enhances temporal coherence in video inpainting tasks, a challenge extensively explored in recent works [8, 9].

Attention mechanism: Tune-A-Video [9] proposes the use of Sparse-Casual Attention (SC_Attn), which calculates the

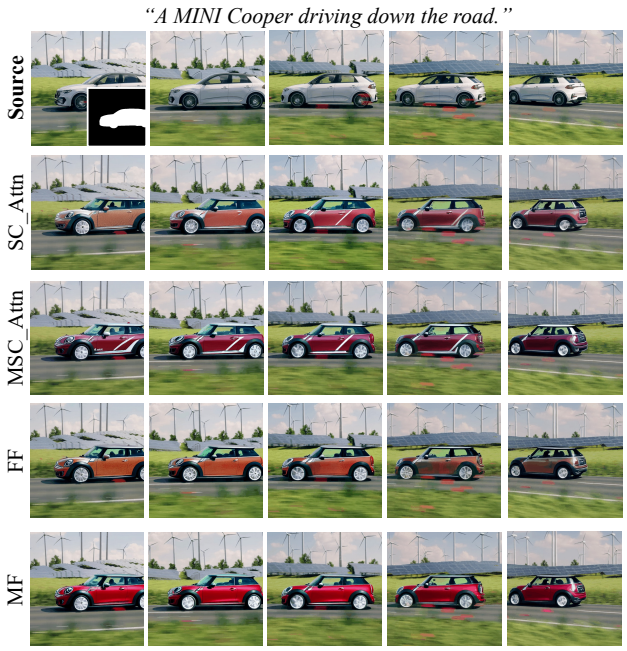

Figure A5. **Ablation analysis of attention guidance.** We compare our middle-frame attention guidance approach (MF) with other temporal correlation modeling method variants, including Sparse-Casual Attention (SC_Attn), Middle-frame Sparse-Casual Attention (MSC_Attn), and First-frame attention guidance (FF).

attention matrix between the current frame ψ^i and two previous frames (ψ^1 and ψ^{i-1}), as described in the following equation:

$$\text{Attention}(\psi^i) = \text{softmax}\left(\frac{Q^i {K^i}^T}{\sqrt{d}}\right) V^i, \tag{1}$$

where $Q^i = W^Q \psi^i$, $K^i = W^K \left[\psi^1, \psi^{i-1}\right]$, and $V^i = W^V \left[\psi^1, \psi^{i-1}\right]$. A similar technique is also adopted in Pix2Video [1]. We adapt Sparse-Casual Attention within each segment of our Temporal MultiDiffusion pipeline.

SC_Attn can be further extended to Middle-frame Sparse-Casual Attention (MSC_Attn) by changing the anchor frame from the first frame within each segment to the middle frame in the whole video, $\psi^{\lceil N'/2 \rceil}$.

Key frame selection: Additionally, we experiment with using the first frame of the video as the guidance frame, modifying our self-attention computation as per Equ. 6 in the main paper:

$$\begin{aligned}\text{Attention}(\psi^i) = &\text{softmax}\left(\frac{Q^i {K^i}^T}{\sqrt{d}}\right) V^i \cdot (1-\omega) + \\ &\text{softmax}\left(\frac{Q^i {K^1}^T}{\sqrt{d}}\right) V^1 \cdot \omega.\end{aligned} \tag{2}$$

We employ an attention guidance weight of $\omega = 0.3$ for this variant.

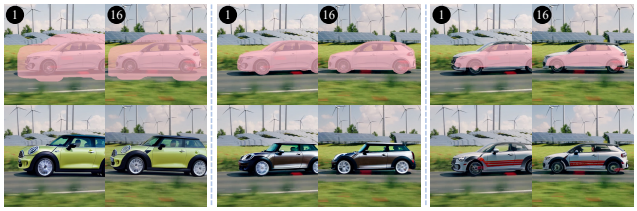

Figure A6. **Ablation analysis of mask accuracy.** We explore the robustness of our method using different mask regions. Here we show 3 examples using "inaccurately" expanded, accurate, and "inaccurately" eroded masks with the same prompt of "Mini Cooper".

Results and discussion: Our experiments, visualized in Fig. A5, demonstrate the varying degrees of success in addressing identity shift issues. The Sparse-Casual Attention (row 2) struggles to prevent identity shifts due to using different key-frames within each segment. Middle-frame Sparse-Casual Attention (row 3) yields better identity preservation, although inconsistencies in the generated patterns can still be observed. The approach using the first frame as guidance (row 4), while maintaining pattern stability, still exhibits significant color variance between the first and last frames.

In contrast, our proposed middle-frame attention guidance mechanism (row 5) excelled in preserving both the color and pattern on the car consistently throughout the video. This result not only highlights the superiority of our method in maintaining temporal coherence but also emphasizes the critical role of strategic frame selection in attention guidance mechanisms for video inpainting tasks.

E.3. Test-time Masks Accuracy

Due to using random synthetic masks during training, our model is very robust to inaccurate masks. As shown in Fig. A6, our method can successfully inpaint the video following the given text prompt when the mask region is significantly larger than the region size. However, when the mask area can not cover the whole to-be-replaced object, our method will fail to modify the shape of the object due to the preservation of out-of-region details.

F. Limitations

In this section, we delve into specific instances where our method can not yield the desired results, as illustrated in Fig. A7. These failure cases, particularly in scenarios involving complex actions, offer crucial insights into the limitations of our current approach and highlight areas for future improvement.

As shown in the first case, in an attempt to generate a horse moving its head from left to right, our method fails to generate plausible results. Instead of showing a smooth

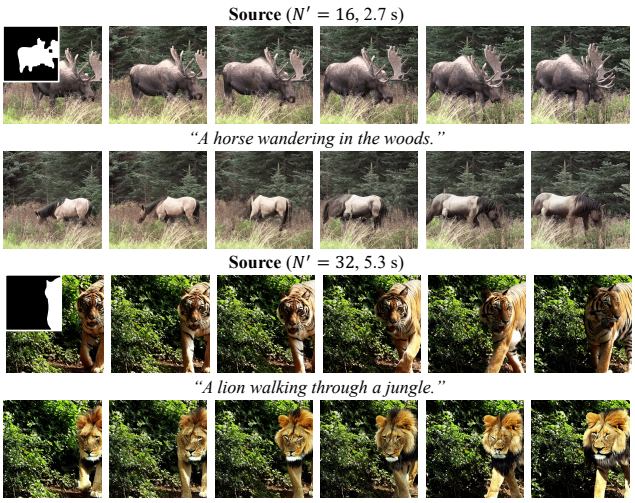

Figure A7. **Failure cases.** We showcase where our method fails to generate results with high fidelity. In the first case, the head of the horse first disappears and then reappears, while in the second case the left foot of the generated lion moving forward, it goes through the right foot of the lion. Please refer to the video results for a better illustration.

head movement, the generated video exhibits the head of the horse disappearing and reappearing on the right side. Concurrently, the body of the horse undergoes an unnatural morphing, transitioning from facing left to right with only minor shape changes. Another challenging scenario involves a lion walking forward. The generated video inaccurately shows the left foot of the lion moving through its right foot, an evident deviation from natural movement. For both cases, we recommend viewing the video results for a more comprehensive understanding of these issues.

As noted in the main paper, these limitations are perhaps due to that our current foundation text-to-video model lacks high-quality motion generation capability. We believe that enhancing the model with more advanced capabilities, especially in interpreting and rendering complex actions, can further improve the quality. A stronger foundation model may also offer better comprehension of intricate movements and interactions, thereby producing more accurate and realistic video content.

Besides the limitations discussed above, we admit our model fails at handling discontinuity, especially objects moving out and back to the video. Such an issue could be mitigated with a more deliberate cross-clip attention injection mechanism, which is a critical direction to further improve the robustness.

G. Any-length Text-to-Video Generation

In this section, we explore the application of our proposed inference pipeline to existing text-to-video generation frameworks, such as AnimateDiff [3], demonstrating

its potential in facilitating any-length text-to-video generation. This exploration serves as a testament to the versatility and adaptability of our method in broader video generation contexts. We have included preliminary results of this extension on the accompanying webpage.

A promising direction for future research lies in the realm of sequential storytelling through video. This involves the idea of performing inference with a series of text prompts, effectively guiding the attention mechanism to evolve in tandem with the narrative. Such an approach could revolutionize how stories are visually narrated, aligning the generated video content with a progressive textual storyline.

References

[1] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *ICCV*, pages 23206–23217, 2023. 4

[2] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 2, 3

[3] 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, 5

[4] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep video inpainting. In *CVPR*, pages 5792–5801, 2019. 2

[5] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 2, 3

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

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 3

[8] 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. 4

[9] 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. 4

[10] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In *CVPR*, pages 3723–3732, 2019. 2

[11] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *ECCV*, pages 528–543. Springer, 2020. 2