Unboxed: Geometrically and Temporally Consistent Video Outpainting

Zhongrui Yu[1] Martina Megaro-Boldini[2] Robert W. Sumner[1,2] Abdelaziz Djelouah[2]

[1]ETH Zürich [2]DisneyResearch|Studios

zhonyu@ethz.ch abdelaziz.djelouah@disney.com

Figure 1. Video outpainting results on two different resolutions. In both case, the area outside the red box is outpainted. Our method achieves results of better quality and more temporally stable, compared to recent works such as Dehan [13], M3DDM [14] and MOTIA [48].

Abstract

Extending the field of view of video content beyond its original version has many applications: immersive viewing experience with VR devices, reformatting 4:3 legacy content to today's viewing conditions with wide screens, or simply extending vertically captured phone videos. Many existing works focus on synthesizing the video using generative models only. Despite promising results, this strategy seems at the moment limited in terms of quality. In this work, we address this problem using two key ideas: 3D supported outpainting for the static regions of the images, and leveraging pre-trained video diffusion model to ensure realistic and temporally coherent results, particularly for the dynamic parts. In the first stage, we iterate between image outpainting and updating the 3D scene representation - we use 3D Gaussian Splatting. Then we consider dynamic objects independently per frame and inpaint missing pixels. Finally, we propose a denoising scheme that allows to maintain known reliable regions and update the dynamic parts to obtain temporally realistic results. We achieve state-of-the-art video outpainting. This is validated quantitatively and through a user study. We are also able to extend the field of view largely beyond the limits reached by existing methods.

1. Introduction

Expanding video frames beyond their original field of view (FoV), known as **Video Outpainting**, is a technique with broad applications in modern media. One example is 4:3 legacy content re-targeting where video outpainting can be used to fill modern widescreen devices seamlessly while avoiding any unnatural stretching. A more forward-looking application is movie viewing with virtual reality (VR) headsets. Video outpainting can enhance the immersion by embedding the viewer in the scene, eliminating distractions from unrelated backgrounds. However, these applications require extremely high-quality results, which is particularly challenging since the synthesized images must integrate harmoniously with the original content and maintain strong temporal coherence across frames.

Although a wide variety of methods have been proposed, using a patch-based approach [1, 2, 12, 46], optical flow [13] or Generative Adversarial Networks (GAN) [24, 53], most recent methods have clearly identified Diffusion Models [14, 48] as the most promising direction. M3DDM [14] trains a 3D diffusion model on large-scale data for video outpainting with a mask modeling approach.

In contrast, MOTIA [48] proposes to fine-tune a pre-trained video diffusion model [19] on each test video to capture data-specific motion and content patterns. However, both of these methods have limitations. First, the temporal consistency of the outpainted regions is still unsatisfactory. We can observe duplications, distortions or objects appearing and disappearing. Second, relying entirely on video diffusion models has a strong impact in terms of memory and limits the methods to low resolution and small field of view extension. Some of these issues can be observed in Figure 1. Both M3DDM [14] and MOTIA [48] have strong visual artifacts and only the latter could be used to outpaint the video to a larger field of view.

To address these issues we propose two core ideas: 3D-supported outpainting for static regions and conditional video synthesis with a pre-trained video diffusion model to ensure realistic, temporally coherent results for dynamic areas. Following these key ideas, we decompose the video into static and dynamic regions and break down the video outpainting task into three stages. *First*, for static regions, we rely on 3D Gaussians [23] as a supporting 3D representation, and we alternate between outpainting key frames and updating the 3D representation. *Second*, for dynamic objects, we implement an object-wise inpainting. Here, among other things, we leverage object-tracking masks in each frame to define inpainting regions and use a fine-tuned diffusion model to inpaint. *Lastly*, using the resulting frames as a starting point, we synthesize temporally consistent results for all frames using a pre-trained video diffusion model [7].

Figure 1 demonstrates the effectiveness of our method. We achieve better results than prior methods both in terms of video quality and temporal consistency. This is confirmed by both quantitative and qualitative experiments. Furthermore, by using GPU-efficient Gaussian Splatting for static region outpainting and avoiding fine-tuning of the video diffusion model, we can go from an input 480×480 video (FoV$= 31°$) to an output resolution of 2560×720 (FoV$= 120°$) with a GPU memory usage of less than 16GB for the entire pipeline.

In summary, our contributions are as follows:

- a novel video outpainting method that combines a 3D representation with a pre-trained video diffusion model;
- an optimization strategy with an additional loss to enable seamless integration of the outpainting with the original static content;
- demonstrating that a pre-trained video diffusion model can function as a temporal filter, making discontinuous frames temporally coherent and simplifying the inpainting task for the dynamic parts.
- State-of-the-art results for video outpainting

2. Related Work

Immersive Scene Generation. To create immersive experiences for users in VR applications, several methods are proposed to generate realistic 3D or 4D scenes that allow for arbitrary viewpoint exploration. LucidDreamer [10] and the following works [16, 35, 43, 54, 55] lift a single image to 3D space, project it from different viewpoints, and inpaint them using diffusion models to create a 3D scene. However, directly applying their method to consecutive video frames would cause severe geometric distortion between frames due to the simple depth alignment and optimization techniques. Dreamscene360 [63] and 4D4KGen [30] create panoramic 3D and 4D scenes separately from a text-generated panoramic image. VividDream [27] generates multi-view videos from a single image and reconstructs a 4D scene with 4D Gaussians. These methods are primarily designed for static or simply animated images and cannot handle videos with complex camera and object motion. Methods such as [9, 28, 29, 49, 50] reconstruct 4D scenes from casual videos, but they are unable to complete regions beyond what is captured in the video.

Image completion. Image completion [6, 39], including image inpainting and outpainting, is a task that restores missing part of an image, and it serves as a basis for video outpainting. Traditional methods [6, 18, 21] use low-level features from the incomplete image to propagate information in known regions to fill missing areas. Early deep learning based methods [32, 42, 44] introduced end-to-end frameworks that learned mappings from corrupted to complete images. Recent methods have increasingly leveraged generative models, including GAN [17, 56, 57, 62], and Diffusion Models [4, 11, 22, 34, 40, 41]. Reference-based image completion methods utilize reference images as generation conditions [52] or fine-tune models on a set of reference images [45] to produce completions that are more faithful to the references. However, even fine-tuned on all input video frames, frame-by-frame completion introduces substantial temporal inconsistency in video outpainting.

Video Outpainting. In the field of video completion, video outpainting is challenging but less explored. Early patch-based methods [1, 2] select the nearest patch from the spatio-temporal domain of the video to expand each frame. Lee *et al*. [26] employ 3D information to warp neighboring frames to the current frame. Dehan *et al*. [13] utilize estimated optical flow to warp the background from adjacent frames and use an image completion network to fill any remaining uncompleted areas. While Zhang *et al.* [61] target video in-painting motion module and structure guidance. M3DDM [14] extends the image diffusion model framework to 3D and trains it on extensive video data for the outpainting task. A global frame is introduced as a conditioning input to improve global information perception. MOTIA [48] proposes a method that fine-tunes a pre-trained

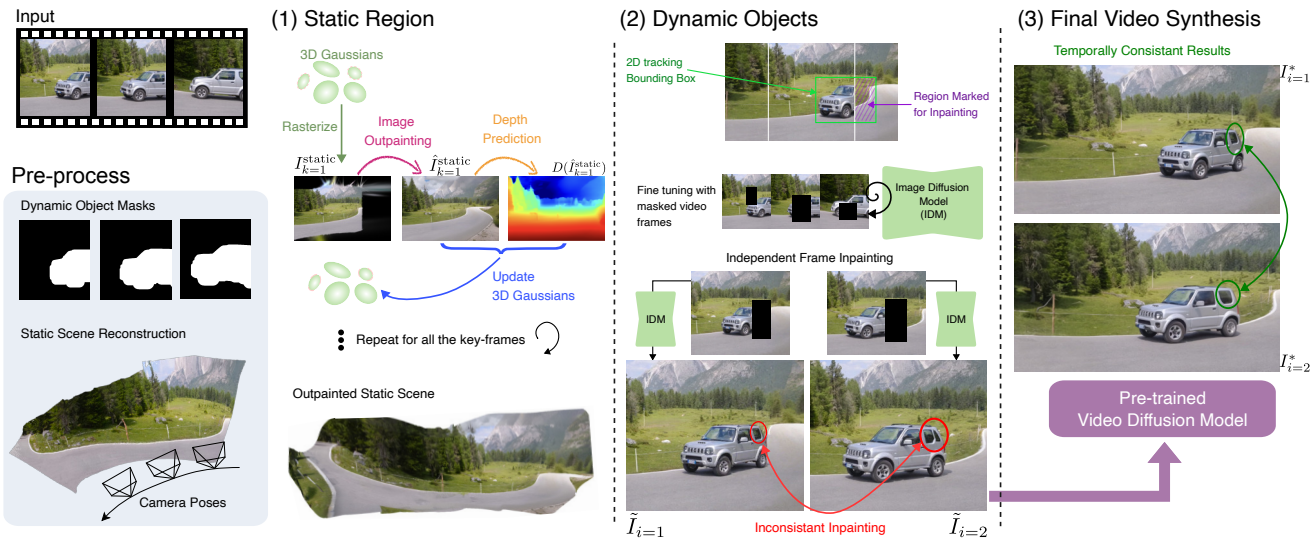

Figure 2. Overview of the proposed method. Starting from the input video sequence, dynamic objects are separated from the static parts of the image, and a static scene reconstruction is made using Gaussian Splatting (GS). In the first stage of our method, the static part is outpainted by iterating between image outpainting, depth prediction and updating the GS model. After this stage, we identify parts of the dynamic object that may require inpainting. Here, an image diffusion model is finetuned on the input frames and used for inpainting each frame independently, which may result in temporal inconsistencies as seen of the car windows. In the last step, a pre-trained video diffusion model is used to fix temporal inconsistencies. Details about each step are provided in the text.

video diffusion model on test samples to learn the intrinsic patterns of a video. In general, diffusion-based approaches rely on priors learned by temporal modules to maintain continuity in the outpainted regions, which is not sufficient. We demonstrate that a pipeline integrating constraints from a 3D representation achieves results of much higher quality.

3. Method

Given a sequence of video frames $\{I_i \in \mathbb{R}^{h\times w\times c}\}_{i=1}^{N}$, our goal is to expand each frame beyond its original field of view to $\{\tilde{I}_i \in \mathbb{R}^{H\times W\times c}\}_{i=1}^{N}$, while maintaining both intra-frame and cross-frame consistency. Although the problem is more challenging than image outpainting, due to the temporal consistency requirements, the presence of multiple frames offers additional observations that can be leveraged. More specifically, it is possible to build a 3D representation of the scene using 3D Gaussians.

As illustrated in Figure 2, our proposed method can be divided into 3 distinct stages. In the first stage, we focus on the static parts of the scene and alternate between image outpainting and updating the 3D scene representation. In the second stage, we focus on the dynamic objects in the scene. Here, we fine-tune an image diffusion model on the available images and use it to inpaint the missing parts of the object in each frame. At this point, all the frames are outpainted, but temporal inconsistencies remain. These frames serve as a starting point for a video synthesis scheme leveraging pre-trained stable video diffusion to synthesize a temporally consistent video sequence. Next, we detail each of these main steps.

3.1. Pre-processing and Static Scene Outpainting

From the static parts of the video, it is possible to estimate a 3D reconstruction of the scene and use the different views across time to provide information about the regions to outpaint. In the first step, we identify the dynamic parts of the scene. A wide variety of methods can be employed here, but we simply rely on object segmentation and, in particular, object tracking models [33]. Figure 2 shows an example of the binary segmentation masks. These masks are expanded to avoid including incorrectly labeled pixels into the background. We also compute a 2D bounding box around the object to help keep track of the object size in case it partially or completely leaves the field of view. The static parts are all the pixels that are not part of any dynamic object.

Static Scene Reconstruction. We propose to use Gaussian Splatting (GS) [23] to reconstruct the static part of the video. This approach allows regions visible in other frames to be rendered into the current frame. For the input video frames $\{I_i\}_{i=1}^{N}$, we first estimate the corresponding camera poses $\{\mathbf{p}_i\}_{i=1}^{N}$ and intrinsic parameters $\{\mathbf{K}_i\}_{i=1}^{N}$ using an off-the-shelf camera pose estimator [60]. We compute a depth map for each frame using parallax and project the background pixels into 3D space, yielding an initial point-cloud for GS training. The training process is supervised by the image reconstruction loss and the depth loss. The image reconstruction loss is the combination of the L1 loss and

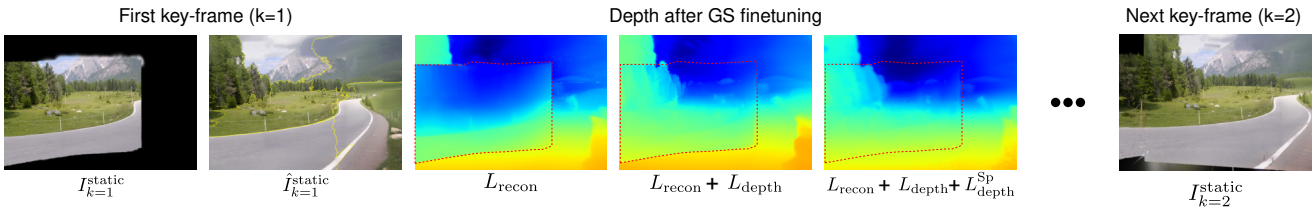

Figure 3. Effect of the loss terms on the GS fine-tuning. Starting from the rasterization of the 3D Gaussians for the first key-frame $k = 1$, we use image outpainting to obtain $\hat{I}^{\text{static}}_k$. This image is subdivided into super-pixels as illustrated here. During the optimization to update the GS model, we use a combination of losses. We can see that all are needed to avoid discontinuities and obtain a realistic smooth depth map. After this update, the process is repeated with the next key-frame using the updated 3D Gaussians.

the structural similarity loss (SSIM). The depth loss acts as a regularization and uses Pearson Correlation [64] between the rendered depth and the depth estimated from a monocular depth estimation model [37]. Apart from the dynamic object masking, this is similar to the optimization proposed by Zhu *et al.* [64].

Single Image Outpainting. Starting from the GS model GS_{static} representing the static parts of the scene, we uniformly sample $k \in \{1, \dots, K\}$ key frames. Our objective is to outpaint them one by one, while we progressively update the GS model. Here we first detail outpainting for a single image. Considering frame I^{static}_k, we adjust its field of view to obtain new intrinsic parameters $\hat{\mathbf{K}}_k$, and render GS_{static} based on the new parameters to get the image $\hat{I}^{\text{static}}_k$. The mask $\hat{\mathbf{m}}^{\text{static}}_k$ for outpainting is defined by checking whether each pixel is covered by the projection of Gaussian points. Any image outpainting method can be used at this level, but we find that the Stable Diffusion XL (SDXL) image outpainting model [38] performs well for our task. This can be summarized as:

$$I^{\text{static}}_k = \phi_{GS_{\text{static}}}(\hat{\mathbf{K}}_k, \mathbf{p}_k) \tag{1}$$

$$\hat{I}^{\text{static}}_k = \phi_{\text{SDXL}}(I^{\text{static}}_k, \hat{\mathbf{m}}^{\text{static}}_k) \tag{2}$$

Updating the GS model. To ensure the consistency of the outpainting across frames, we update the GS model each time we outpaint a key-frame. Denote by K' the number of key-frames that have been outpainted so far. First, we estimate the depth for the newly outpainted image using a single image depth model [37]. Then we initialize new Gaussian points for the outpainted pixel and fine-tune the GS model with this additional data. We use an image reconstruction loss

$$\begin{aligned} L_{\text{recon}} = \sum_{k=1}^{K'} & \lambda_1 L_1(\hat{I}^{\text{static}}_k, \phi_{\text{GS}}(\hat{K}_k, p_k)) \\ & + \lambda_2 L_{\text{SSIM}}(\hat{I}^{\text{static}}_k, \phi_{\text{GS}}(\hat{K}_k, p_k)) \end{aligned} \tag{3}$$

between the rasterization from the model $\phi_{\text{GS}}(\hat{K}_k, p_k)$ and the outpainting result. We also use 2 depth based losses, the first one on the full depth map:

$$L_{\text{depth}} = \sum_{k=1}^{K'} \frac{\text{Cov}(D(\hat{I}^{\text{static}}_k), D_{\text{GS}}(\hat{K}_k, p_k))}{\sqrt{\text{Var}(D(\hat{I}^{\text{static}}_k))\text{Var}(D_{\text{GS}}(\hat{K}_k, p_k))}} \tag{4}$$

where $D(\hat{I}^{\text{static}}_k)$ is the depth predicted by the single image depth model [37], and $D_{\text{GS}}(\hat{K}_k, p_k)$ is the depth from GS model that we optimize. For every image, we extract S super-pixels, and use the second depth loss at this level:

$$L^{\text{Sp}}_{\text{depth}} = \sum_{k=1}^{K'} \sum_{s=1}^{S} \frac{\text{Cov}_s(D(\hat{I}^{\text{static}}_k), D_{\text{GS}}(\hat{K}_k, p_k))}{\sqrt{\text{Var}_s(D(\hat{I}^{\text{static}}_k))\text{Var}_s(D_{\text{GS}}(\hat{K}_k, p_k))}} \tag{5}$$

Figure 3 demonstrate this process and shows the importance of the different loss terms for the GS update. Starting from the rasterization of the initial GS model, we can see that the 3D model does not cover the new field of view. The frame is outpainted and superpixels are illustrated in yellow. Using only the reconstruction loss L_{recon} leads to strong discontinuities between the original GS model and the new outpainted areas. Adding the depth loss L_{depth} helps, but the bottom part of the image still has important discontinuities that would be revealed when projecting to other frames (i.e other views). Using all the proposed loss terms provides the best results, and the process can now be iterated by adding new views.

3.2. Initial Inpainting for the Dynamic Object(s)

After the static scene outpainting, we can render any frame with $\phi_{\text{GS}}(\hat{K}_i, p_i)$. As illustrated in Figure 2, we miss part of the object when it is partially outside the filed of view. We need to inpaint this missing part. We use the 2D tracking bounding box to identify the areas that require inpainting for each frame of the video.

For this inpainting problem we use a similar strategy as ReaFill [45]. An image diffusion model is fine-tuned for the inpainting task on the input frames (in our case SDXL). It is then applied independently on each frame. Thanks to the fine-tuning, the resulting inpainting is very similar across frames. However temporal inconsistencies remain as it is

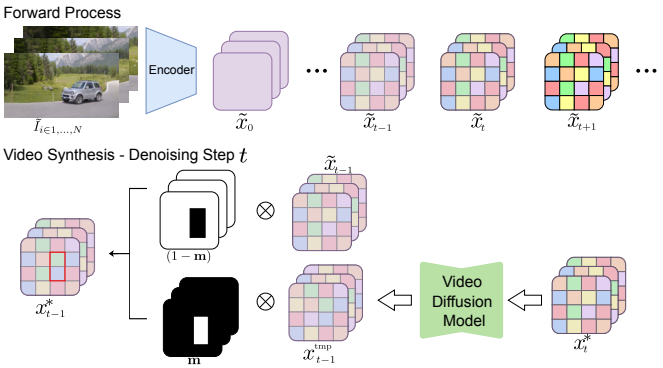

Figure 4. Details of the video latent denoising strategy. Starting from the independently outpainted frames $\tilde{I}_{i\in1,\dots,N}$, we obtained the corresponding latents in the forward process ($\tilde{x}_{i\in1,\dots,N}$). In the denoising process we blend *known* latent with denoising latents according to the mask $\mathbf{m}$. The mask and the time step t_{end} respectively control which region of the frames get updated by the video diffusion model and how large the changes are.

clearly visible on the inpainting results in Figure 2. Next we describe how a Video Diffusion Model [7, 8] is used to fix all these temporal artifacts.

3.3. Guided Video Synthesis

If we consider the original input sequences of frames $I_{i\in1,\dots,N}$, after the previous stages we obtain fully outpainted frames $\tilde{I}_{i\in1,\dots,N}$. Static regions are consistent thanks to the supporting GS model, while the dynamic objects are independently inpainted for each frame. Starting from this data, our objective is to synthesize a realistic sequence of frames $I^*_{i\in1,\dots,N}$. Not only the temporal inconsistencies on the dynamic objects need to be fixed, but the entire outpainted regions need to appear realistic. Indeed, real-world videos often contain subtle background movements, such as leaves swaying in the wind or water rippling.

We leverage pre-trained video diffusion models for this video synthesis. The idea is that video diffusion models are trained for generating real video sequences, with consistent content and realistic motion. We propose using the previously generate frames in a denoising scheme that modifies the different frame areas based on the expected inconsistency level. On the dynamic objects, we expect more changes (as the inpainting is independent), while less change should be allowed on the rest of the image as it is supported by the 3D consistent GS model.

More specifically, if we consider a latent video diffusion model, a sequence of input frames I_i is encoded into the corresponding sequence of latent values x_i. The forward process can be applied to the sequence of frames $\tilde{I}_{i\in1,\dots,N}$ to obtain *known* latents

$$\tilde{x}_{t-1} \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}\tilde{x}_0, \sqrt{1-\bar{\alpha}_t}\mathbf{1}). \tag{6}$$

For simplicity, we dropped the frame index i.

We use the reverse process to synthesize the final video sequence. At each time step, we denoise previous latents using the latent video diffusion model with parameters θ, to obtain temporary latents

$$x^{\text{tmp}}_{t-1} \sim \mathcal{N}(\mu_\theta(x^*_t, t), \Sigma_\theta(x^*_t, t)) \tag{7}$$

To obtain the final latents for this time step, we blend in the known latents according to

$$x^*_{t-1} = (1-\mathbf{m}) \odot \tilde{x}_{t-1} + \mathbf{m} \odot x^{\text{tmp}}_{t-1} \tag{8}$$

where the binary mask $\mathbf{m}$ is defined as

$$\mathbf{m} = \begin{cases} \mathbf{m}^{\text{dynamic}} & \text{if } t \geq t_{\text{end}} \\ \mathbf{m}^{\text{dynamic}} + \mathbf{m}^{\text{static}} & \text{if } t < t_{\text{end}} \end{cases} \tag{9}$$

The mask, $\mathbf{m}^{\text{dynamic}}$, corresponds to the dynamic objects in the scene, and the mask $\mathbf{m}^{\text{static}}$ corresponds to the *static* parts in the outpainted regions. This scheme is similar to ideas already explored in the context of image diffusion models [3, 34], but applied to video diffusion models. If we observe the changes in Figure 2, we can see that the pre-trained video diffusion model can function as a temporal filter, making discontinuous frames temporally coherent. This simplifies the inpainting task since we do not rely on expensive training of the video model and simply use it at inference time.

4. Experiments

4.1. Implementation details

Static Scene. In pre-processing, we estimate the camera poses of the source video [31, 60]. Our GS training process consists of $1+K+1$ rounds in total, where K is the number of key frames we sample from the video. In the first training round, we train the GS model only on the static part of the video for 3000 iterations. Then we iterate between outpainting the K key frames and K rounds of GS training to update the outpainted regions into the GS model. To avoid overfitting, we employ a sliding window training strategy: following outpainting of the i-th frame, we train only on frames within the window $[i-5, i+5]$ for 500 iterations. This window shifts progressively as outpainting and training continues. After outpainting all K key frames, we conduct a final round of training for refinement on all outpainted and original background frames for 1000 iterations. All training and outpainting can be completed on a single 4090 GPU.

Initial Object Inpainting. Before initial inpainting on the missing regions of dynamic objects, we fine-tune the SDXL model on the masked source video. The approach is similar to the one proposed by Tang *et al.* [45]. Masks are randomly

Figure 5. Qualitative Comparison with State-of-the-art methods on the DAVIS dataset [36]. The area outside of the red lines (only shown for the first frame) is outpainted. We show the results for 3 frames from the each video sequence. Compared to MOTIA [48] and M3DDM[14], our results are of better quality and temporally stable. Please pay attention to the changes of the outpainted area between the frames of the same sequence.

generated for each input frame, with random distributed positions and mask ratios varying from 0.2 to 0.8. For each object, a prompt "`a photo of [v]`" is applied. During fine-tuning, a Low-Rank Adaption (LoRA) [20] with rank 8 is trained for both the text-encoder and UNet with a batch size of 4 and gradient accumulation steps of 4 for 2000 steps.

Video Diffusion Model. We use a pretrained image-to-video model [7] to complete object motion. The number of denoising steps is set to 25, with the level of noise added to the reconstructed frames generally set to 0.8 in our experiments. We apply a resampling strategy similar to Repaint [34] five times after each denoising step to improve the quality and stability of the objects. For videos where the object fully appears in the middle or last frame, we employ the Time-Reversal strategy from [15], conditioning the generation on additional frames to better retain the object identity.

4.2. Experiment Setup

Competitors. We compare our method with the following methods. (1) **Dehan *et al.*** [13] separate foreground and background components. They use estimated optical flow to warp the background, with missing regions filled in using an image completion network. (2) **M3DDM** [14] proposes a masked 3D Diffusion Model architecture, which is trained on WebVid [5] and a self-collected 5M e-commerce dataset for video outpaint tasks. (3) **MOTIA** [48] is the current state-of-the-art method for video outpainting. It is based on the ControlNet [58] inpainting model and the temporal module from AnimateDiff [19]. It trains a LoRA [20] for

each test sample to learn the intrinsic patterns.

Datasets. Following M3DDM [14] and MOTIA [48], we evaluate our method on the DAVIS [36] and YouTube-VOS [51] datasets. For the YouTube-VOS [51] dataset, due to its large number of videos and the lack of clarity in prior works on which videos were tested, we selected a high-quality subset and evaluated all methods on it. Consistent with prior works, we compare the horizontal outpainting results for each test video at mask ratios of 0.25 and 0.66.

Evaluation Metrics. The quality of outpainting is primarily evaluated in terms of visual quality, realism, and temporal consistency. For this purpose, we employ two well-established metrics: Fréchet Video Distance (FVD) [47] and flow warping error (E_{warp}) [25]. FVD assesses the similarity between the distribution of the outpainted video and the original video, while flow warping error evaluates temporal consistency by measuring the error when outpainted frames are warped to adjacent frames using estimated optical flow. In addition, to align with previous works, we also report Peak Signal to Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) [59] between the outpainted and ground-truth videos, although we consider these metrics limited in capturing the quality of outpainting. We further conducted a user study to gather subjective evaluations of outpainting quality, which we detail in Sec. 4.3.

4.3. Evaluation & Results

Qualitative Results. Fig. 5 and Fig. 6 illustrate the qualitative comparison of our method against other approaches on the DAVIS and YouTube-VOS datasets. M3DDM [14] frequently suffers from generation failures and produces blurred content, as shown in the first samples of Fig. 5 and Fig. 6. MOTIA [48], while fine-tuning on each test sample to capture input-specific patterns, often generates repetitive objects and background patterns. More importantly, both methods struggle to accurately capture contextual information across frames, resulting in inconsistency when outpainting elements seen in adjacent frames. In contrast, our method overcomes these issues, achieving better cross-frame consistency. Furthermore, our approach's object-specific completion approach prevents objects from duplicating or disappearing beyond the frame.

Generalization and multiple objects. Our strategy, differentiating between static and dynamic elements, can naturally extend to multiple dynamic objects by decomposing the frame into object layers based on their depth. Each object is processed separately using our pipeline and then seamlessly blended to produce the final result. Figure 7 shows that our method still performs best on such complex scenarios.

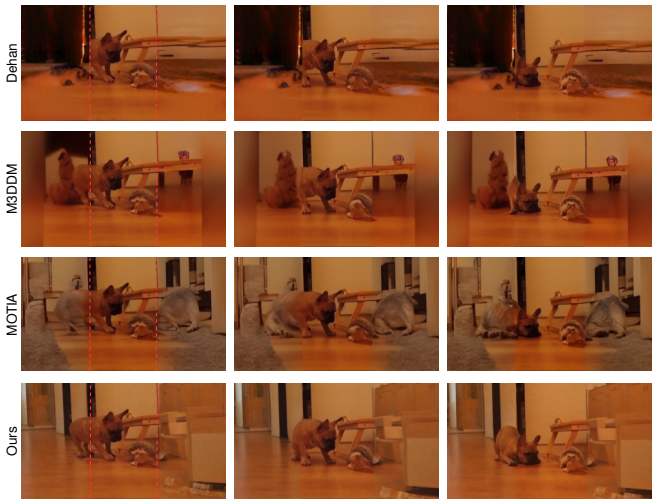

Figure 6. Qualitative Comparison with State-of-the-art methods on Youtube-VOS dataset [51]. The area outside of the red lines (only shown for the first frame) is outpainted. We show the results for 3 frames from the each video sequence. Compared to Dehan [13], MOTIA [48] and M3DDM[14], our results are of better quality and temporally stable. Please pay attention to the changes of the outpainted area between the frames of the same sequence.

Figure 7. Qualitative Comparison with State-of-the-art methods on challenging scenarios (multiple dynamic objects, complex static regions, etc.).

Quantitative Results. Tab. 1 presents a quantitative comparison of our method with other approaches. Our method outperforms the current state-of-the-art in both FVD and E_{warp}, achieving improvements of 18.8% and 11.7% on the DAVIS dataset and 26.9% and 33.3% on the YouTube-VOS dataset, respectively. We also achieve the highest PSNR and SSIM scores on both datasets. For LPIPS, while our method is the second behind the number reported in MOTIA [48] on the DAVIS dataset, it surpasses the score obtained from using their public implementation and provided model.

Ablation Study We conducted ablation studies for the different components of our method. Quantitative evalua-

	DAVIS					YouTube-VOS				
	PSNR↑	SSIM↑	LPIPS↓	FVD↓	E_{warp}↓	PSNR↑	SSIM↑	LPIPS↓	FVD↓	E_{warp}↓
Dehan [13]	17.96	0.627	0.233	363.1	0.030	18.34	0.761	0.287	353.6	0.043
M3DDM [14]	20.26	0.708	0.204	300.0	0.017	18.33	0.725	0.424	283.1	0.024
MOTIA (model) [48]	18.25	0.699	0.282	294.8	0.021	18.54	0.790	0.250	156.4	0.027
MOTIA (paper) [48]	20.36	0.758	**0.160**	286.3	-	-	-	-	-	-
Ours	**22.62**	**0.806**	0.200	**232.5**	**0.015**	**19.88**	**0.832**	**0.236**	**114.2**	**0.016**

Table 1. Quantitative Comparison with State-of-the-art methods on DAVIS dataset [36] and YouTube-VOS [51]. Our method achieves the best performance overall. In the case of MOTIA [48], we have 2 rows as we report the evaluation using the model and code available online(*model*) and the numbers from the original paper (*paper*). See the text for details.

	PSNR↑	SSIM↑	LPIPS↓	FVD↓	E_{warp}↓
(a)	18.15	0.694	0.248	448.8	0.032
(b)	18.05	0.716	0.274	323.2	0.022
(c)	20.68	0.772	0.218	293.3	0.020
(d)	20.05	0.776	0.223	319.8	0.017
(e)	**22.62**	**0.806**	**0.200**	**232.5**	**0.015**

Table 2. Quantitative ablation results. (a). Outpainting with fine-tuned Image Model; (b). (a) + Guided Video Synthesis; (c). Our model without guided video synthesis; (d). Our model without the initialization from object inpainting; (e). Our complete model.

tion is shown in Table 2, while the visual results are provided in supplementary material. When only using single image outpainting (SDXL fine-tuned on content [45]), the images achieve good realism but lack temporal continuity. When applying video diffusion guided synthesis to these outpainted frames directly, the temporal consistency within the outpainted areas improves individually. However, coherence between the outpainted regions and the original video content remains weak. This issue is difficult to resolve solely with Video Diffusion Model priors. When using the static region outpainting (assisted with GS) and either the image inpainting or video diffusion alone are applied for object completion, the objects suffer from significant temporal variations in appearance. Using all the components achieves the best results.

User Study. Given the subjective nature of evaluating outpainting quality, we conducted a user study to compare our method with competing approaches. Users were asked to evaluate each method across several dimensions, including realism (such as whether the outpainted results look natural and harmonious), temporal consistency (including background consistency, object motion consistency, etc.), and overall visual quality (including color fidelity, smoothness of boundaries, blurriness etc.). The results, shown in Fig. 8, indicate that our method was clearly preferred by users compared to others, achieving more than 80% of the votes in all dimensions. Further details of the user study can be found in the supplementary materials.

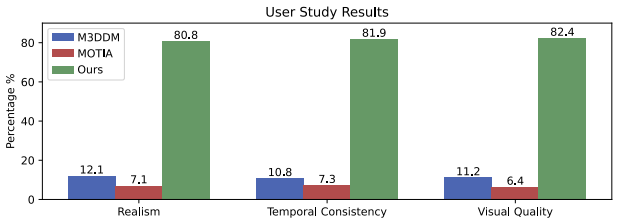

Figure 8. The result of the user study includes 619 votes from 37 participants. Our method is preferred by more than 80% of the votes in all dimensions.

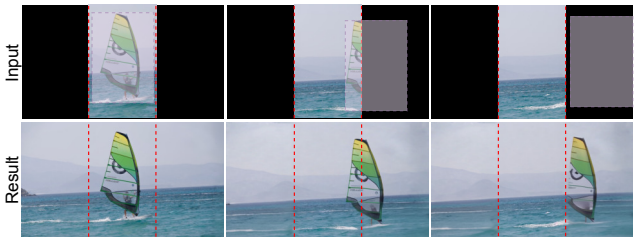

Figure 9. Challenges with dynamic objects moving in and out of the original frame: Our current solution of tracking and extrapolating the trajectory of the bounding box can be improved.

5. Discussions

Building on Gaussian Splatting and a Video Diffusion Model, we propose a new state-of-the-art method for video outpainting. Despite our strong results, some issues remain, especially around objects moving in and out of the original frame. An example is shown in Fig. 9. Currently tracking and extrapolating the trajectory of the bounding box offers a solution, but limitations in the single image inpainting negatively impact the quality of the final result. Another area of improvement is the outpainting involving multiple occluded dynamic objects. Our current solution of inpainting each object separately and blending them based on the estimated depth is limited. Exploring the use of more advanced 4D dynamic Gaussian representations jointly with video models offers a promising future direction towards higher and higher quality video ouptainting.

References

[1] Amit Aides, Tamar Avraham, and Yoav Y Schechner. Multiscale ultrawide foveated video extrapolation. In *2011 IEEE International Conference on Computational Photography (ICCP)*, pages 1–8. IEEE, 2011. 1, 2

[2] Tamar Avraham and Yoav Y Schechner. Ultrawide foveated video extrapolation. *IEEE Journal of Selected Topics in Signal Processing*, 5(2):321–334, 2010. 1, 2

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18208–18218, 2022. 5

[4] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18208–18218, 2022. 2

[5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 6

[6] M Bertalmio. Image inpainting, 2000. 2

[7] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 5, 6

[8] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 5

[9] Wen-Hsuan Chu, Lei Ke, and Katerina Fragkiadaki. Dreamscene4d: Dynamic multi-object scene generation from monocular videos. *arXiv preprint arXiv:2405.02280*, 2024. 2

[10] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023. 2

[11] Ciprian Corneanu, Raghudeep Gadde, and Aleix M Martinez. Latentpaint: Image inpainting in latent space with diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4334–4343, 2024. 2

[12] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on image processing*, 13(9): 1200–1212, 2004. 1

[13] Loïc Dehan, Wiebe Van Ranst, Patrick Vandewalle, and Toon Goedemé. Complete and temporally consistent video outpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 687–695, 2022. 1, 2, 6, 7, 8

[14] Fanda Fan, Chaoxu Guo, Litong Gong, Biao Wang, Tiezheng Ge, Yuning Jiang, Chunjie Luo, and Jianfeng Zhan. Hierarchical masked 3d diffusion model for video outpainting. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7890–7900, 2023. 1, 2, 6, 7, 8

[15] Haiwen Feng, Zheng Ding, Zhihao Xia, Simon Niklaus, Victoria Abrevaya, Michael J Black, and Xuaner Zhang. Explorative inbetweening of time and space. In *European Conference on Computer Vision*, pages 378–395. Springer, 2025. 6

[16] Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. Scenescape: Text-driven consistent scene generation. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[18] Qiang Guo, Shanshan Gao, Xiaofeng Zhang, Yilong Yin, and Caiming Zhang. Patch-based image inpainting via two-stage low rank approximation. *IEEE transactions on visualization and computer graphics*, 24(6):2023–2036, 2017. 2

[19] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2, 6

[20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 6

[21] Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf. Image completion using planar structure guidance. *ACM Transactions on graphics (TOG)*, 33(4):1–10, 2014. 2

[22] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. *arXiv preprint arXiv:2403.06976*, 2024. 2

[23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 2, 3

[24] Naoki Kimura and Jun Rekimoto. Extvision: augmentation of visual experiences with generation of context images for a peripheral vision using deep neural network. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–10, 2018. 1

[25] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *European Conference on Computer Vision*, 2018. 7

[26] Minhyeok Lee, Suhwan Cho, Chajin Shin, Jungho Lee, Sunghun Yang, and Sangyoun Lee. Video diffusion models are strong video inpainter. *arXiv preprint arXiv:2408.11402*, 2024. 2

[27] Yao-Chih Lee, Yi-Ting Chen, Andrew Wang, Ting-Hsuan Liao, Brandon Y Feng, and Jia-Bin Huang. Vividdream: Generating 3d scene with ambient dynamics. *arXiv preprint arXiv:2405.20334*, 2024. 2

[28] Yao-Chih Lee, Zhoutong Zhang, Kevin Blackburn-Matzen, Simon Niklaus, Jianming Zhang, Jia-Bin Huang, and Feng Liu. Fast view synthesis of casual videos with soup-of-planes. In *European Conference on Computer Vision*, pages 278–296. Springer, 2025. 2

[29] Jiahui Lei, Yijia Weng, Adam Harley, Leonidas Guibas, and Kostas Daniilidis. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. *arXiv preprint arXiv:2405.17421*, 2024. 2

[30] Renjie Li, Panwang Pan, Bangbang Yang, Dejia Xu, Shijie Zhou, Xuanyang Zhang, Zeming Li, Achuta Kadambi, Zhangyang Wang, and Zhiwen Fan. 4k4dgen: Panoramic 4d generation at 4k resolution. *arXiv preprint arXiv:2406.13527*, 2024. 2

[31] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast, and robust structure and motion from casual dynamic videos, 2024. 5

[32] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 85–100, 2018. 2

[33] 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. 3

[34] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022. 2, 5, 6

[35] Hao Ouyang, Kathryn Heal, Stephen Lombardi, and Tiancheng Sun. Text2immersion: Generative immersive scene with 3d gaussians. *arXiv preprint arXiv:2312.09242*, 2023. 2

[36] 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 *Computer Vision and Pattern Recognition*, 2016. 6, 7, 8

[37] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 4

[38] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 4

[39] Weize Quan, Jiaxi Chen, Yanli Liu, Dong-Ming Yan, and Peter Wonka. Deep learning-based image and video inpainting: A survey. *International Journal of Computer Vision*, 132(7): 2367–2400, 2024. 2

[40] 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

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[42] Iizuka Satoshi, Simo-Serra Edgar, and Ishikawa Hiroshi. Globally and locally consistent image completion. *ACM Trans. Graph*, 36(4):3073659, 2017. 2

[43] Jaidev Shriram, Alex Trevithick, Lingjie Liu, and Ravi Ramamoorthi. Realmdreamer: Text-driven 3d scene generation with inpainting and depth diffusion. *arXiv preprint arXiv:2404.07199*, 2024. 2

[44] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 2

[45] Luming Tang, Nataniel Ruiz, Qinghao Chu, Yuanzhen Li, Aleksander Holynski, David E Jacobs, Bharath Hariharan, Yael Pritch, Neal Wadhwa, Kfir Aberman, et al. Realfill: Reference-driven generation for authentic image completion. *ACM Transactions on Graphics (TOG)*, 43(4):1–12, 2024. 2, 4, 5, 8

[46] Laura Turban, Fabrice Urban, and Philippe Guillotel. Extrafoveal video extension for an immersive viewing experience. *IEEE transactions on visualization and computer graphics*, 23(5):1520–1533, 2016. 1

[47] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 7

[48] Fu-Yun Wang, Xiaoshi Wu, Zhaoyang Huang, Xiaoyu Shi, Dazhong Shen, Guanglu Song, Yu Liu, and Hongsheng Li. Be-your-outpainter: Mastering video outpainting through input-specific adaptation. *arXiv preprint arXiv:2403.13745*, 2024. 1, 2, 6, 7, 8

[49] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. *arXiv preprint arXiv:2407.13764*, 2024. 2

[50] Shizun Wang, Xingyi Yang, Qiuhong Shen, Zhenxiang Jiang, and Xinchao Wang. Gflow: Recovering 4d world from monocular video. *arXiv preprint arXiv:2405.18426*, 2024. 2

[51] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas S. Huang. Youtube-vos: A large-scale video object segmentation benchmark. *CoRR*, abs/1809.03327, 2018. 7, 8

[52] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 2

[53] Zhenqiang Ying and Alan Bovik. 180-degree outpainting from a single image. *arXiv preprint arXiv:2001.04568*, 2020. 1

[54] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. *arXiv preprint arXiv:2406.09394*, 2024. 2

[55] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6658–6667, 2024. 2

[56] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018. 2

[57] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Aggregated contextual transformations for high-resolution image inpainting. *IEEE Transactions on Visualization and Computer Graphics*, 29(7):3266–3280, 2022. 2

[58] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 6

[59] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[60] Zhoutong Zhang, Forrester Cole, Zhengqi Li, Michael Rubinstein, Noah Snavely, and William T Freeman. Structure and motion from casual videos. In *European Conference on Computer Vision*, pages 20–37. Springer, 2022. 3, 5

[61] Zhixing Zhang, Bichen Wu, Xiaoyan Wang, Yaqiao Luo, Luxin Zhang, Yinan Zhao, Peter Vajda, Dimitris Metaxas, and Licheng Yu. Avid: Any-length video inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7162–7172, 2024. 2

[62] Haitian Zheng, Zhe Lin, Jingwan Lu, Scott Cohen, Eli Shechtman, Connelly Barnes, Jianming Zhang, Ning Xu, Sohrab Amirghodsi, and Jiebo Luo. Image inpainting with cascaded modulation gan and object-aware training. In *European Conference on Computer Vision*, pages 277–296. Springer, 2022. 2

[63] Shijie Zhou, Zhiwen Fan, Dejia Xu, Haoran Chang, Pradyumna Chari, Suya Bharadwaj, Tejas You, Zhangyang Wang, and Achuta Kadambi. Dreamscene360: Unconstrained text-to-3d scene generation with panoramic gaussian splatting. *arXiv preprint arXiv:2404.06903*, 2024. 2

[64] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting. *arXiv preprint arXiv:2312.00451*, 2023. 4