

Seen-to-Scene: Keep the Seen, Generate the Unseen for Video Outpainting

Inseok Jeon¹ Minhyeok Lee¹ Seunghoon Lee¹ Minseok Kang¹ Suhwan Cho² Sangyoun Lee¹

¹ Yonsei University ² GenGenAI

https://inseokjeon.github.io/seen_to_scene



Figure 1. **Seen-to-Scene** is a novel framework that unifies propagation-based and generation-based paradigms for video outpainting. It maintains strong spatio-temporal consistency across frames even under challenging conditions such as camera motion, dynamic objects and complex motion patterns. The official implementation is released at https://github.com/InSeokJeon/Seen_to_Scene.

Abstract

*Video outpainting aims to expand the visible content of a video beyond the original frame boundaries while preserving spatial fidelity and temporal coherence across frames. Existing methods primarily rely on large-scale generative models, such as diffusion models. However, generation-based approaches suffer from implicit temporal modeling and limited spatial context. These limitations lead to intra-frame and inter-frame inconsistencies, which become particularly pronounced in dynamic scenes and large outpainting scenarios. To overcome these challenges, we propose **Seen-to-Scene**, a novel framework that unifies propagation-based and generation-based paradigms for video outpainting. Specifically, *Seen-to-Scene* leverages flow-based propagation with a flow completion network pre-trained for video inpainting, which is fine-tuned in an end-to-end manner to bridge the domain gap and reconstruct coherent motion fields. To further improve the efficiency and reliability of propagation, we introduce a reference-guided latent propagation that effectively propagates source content across frames. Extensive experiments demonstrate that our method achieves superior temporal coherence and visual realism with efficient inference, surpassing even prior state-of-the-art methods that require input-specific adaptation.*

1. Introduction

As short-form and multi-platform video content grows rapidly, video production and consumption are expanding across a wide range of devices and displays. In practice, most videos are typically captured at fixed aspect ratios such as 16:9 and 9:16, while modern displays offer diverse aspect ratios and sizes. This mismatch often leads to black bars or undesired cropping, which remains a practical challenge in modern video editing and content creation.

Video outpainting, which generates visual content beyond the original frame boundaries across the video sequence, has emerged as a promising solution. Recently, a wide variety of methods have been proposed, and they can be categorized into two paradigms: (i) propagation-based approaches and (ii) generation-based approaches.

Propagation-based approaches complete the outpainting regions by warping neighboring frames using optical flow and propagating observable content into missing areas. Such propagation methods have been widely used in video inpainting [4, 10, 12, 24, 25, 28], where masked regions are restored by propagating reliable spatial and temporal information from neighboring frames. Building upon propagation-based approaches originally designed for video inpainting, Dehan [5] introduced a propagation-based

framework for video outpainting. However, applying these propagation strategies to video outpainting leads to several limitations. First, dense pixel-level warping and sequential flow composition across intermediate frames demand substantial computation, as the cost scales with both spatial resolution and the number of frames. Unlike video inpainting, which fills in missing regions within a fixed spatial resolution, video outpainting inherently requires spatial expansion beyond the original frame boundaries. As a result, propagation costs increase substantially in video outpainting. Second, existing methods rely on flow-completion networks trained for video inpainting, which introduces a domain gap when applied to video outpainting. These networks are optimized to estimate motion within localized masked regions and therefore struggle to generalize to the large spatial expansions required in outpainting. As the distance from the source region increases, the predicted flows become progressively unstable, resulting in flow misalignment and visible artifacts. Finally, propagation alone cannot synthesize non-observable regions and therefore necessitates generative models or auxiliary networks to fill the missing areas.

Motivated by these limitations and the success of generative models, recent studies [2, 6, 11, 13, 14, 19, 23, 27] have shifted toward generation-based approaches. Leveraging diffusion models [7] with powerful generative capability, they have achieved significant progress. However, even with these advances, generation-based approaches still face key challenges. First, generative models cannot ensure consistent preservation of source content over time, as diffusion-based models rely on implicit temporal correlations. Such modeling maintains source content between adjacent frames, but its temporal correlations break down over longer intervals, causing inter-frame inconsistencies and temporal discontinuities. Second, limited spatial cues restrict global scene understanding, forcing reliance on learned generative priors. These models often hallucinate content, resulting in intra-frame inconsistencies and structural artifacts. Overall, achieving spatio-temporal coherence remains an open challenge for video outpainting.

To overcome the limitations of these two paradigms, we propose **Seen-to-Scene**, a novel propagation-based video diffusion framework for video outpainting. At its core, Seen-to-Scene unifies propagation and generation by combining the source content preservation of propagation with the generative capability of diffusion models. The integration of latent propagation into pre-trained video diffusion models [1] contributes in three key aspects. (i) **Implicit Latent Diversity**: Latent propagation introduces variations in spatial structure and temporal dynamics across frames. Such variations inherently expose the model to diverse spatial layouts and motion patterns, improving its generalization across a wide range of visual scenes and spatial expansion ratios. (ii) **Enhancing Spatial Reasoning**: Propagated latents provide additional spatial cues beyond frame boundaries. Such guidance promotes global structural awareness and coherent spatial layout reasoning, miti-

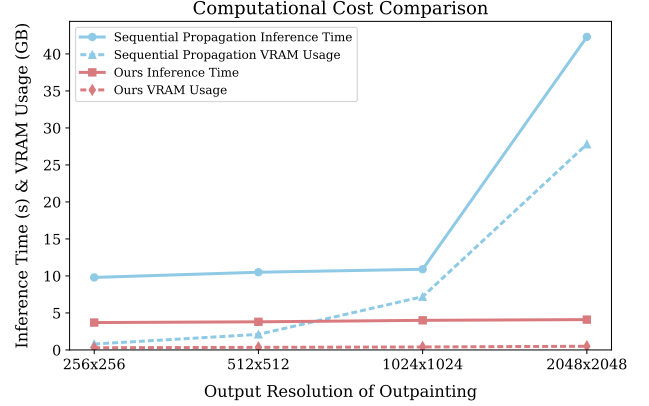


Figure 2. Computational cost comparison with a SOTA propagation-based model [4]. Dense and sequential propagation incurs significantly higher inference time and VRAM usage.

gating hallucinated structures and improving spatial coherence. (iii) **Improving Temporal Consistency**: Explicit propagation across frames encourages the model to preserve observable regions while generating non-observable content. This mechanism further strengthens temporal coherence and structural consistency throughout the sequence.

To effectively integrate propagation into a generative model, several practical challenges must be addressed. Conventional propagation schemes incur significant computational overhead due to dense warping in pixel space and sequential flow composition across frames. We alleviate this issue by introducing a reference-guided latent propagation that identifies content-rich reference frames via inter-frame correlation within a sliding temporal window and performs propagation directly in the latent space. Since propagation is directly integrated into the generation process, misaligned propagated latents can adversely affect subsequent generation. We also design a lightweight refinement module to mitigate propagation artifacts and reduce flow dependency. Furthermore, we observe that existing flow completion networks trained for video inpainting lead to a non-trivial domain gap when applied to video outpainting. We therefore adapt the flow completion network to video outpainting, yielding robust motion estimates for propagation.

Our main contributions can be summarized as follows:

- We propose Seen-to-Scene, a novel propagation-based video diffusion model that integrates propagation and generation for video outpainting to preserve observed content while naturally generating unseen regions.
- We introduce an efficient reference-guided latent propagation tailored for video outpainting that retrieves informative frames based on cross-frame similarity, enabling propagation without sequential frame-wise composition.
- To improve the alignment of propagated latents, we design a refinement module and conduct the first analysis of the domain gap in flow completion for video outpainting.
- Despite not requiring input-specific adaptation, Seen-to-Scene achieves superior visual fidelity and temporal coherence with efficient inference on standard outpainting benchmarks composed of diverse real-world videos.



Figure 3. **Key limitations of existing video outpainting approaches.** (a) *Flow completion domain gap*: Flow completion networks trained for video inpainting struggle to generalize to video outpainting due to the substantially larger missing regions involved. Consequently, the predicted flow often suffers from bleeding around dynamic objects and fails to provide reliable motion estimates for regions far from the original content. (b) *Source content preservation failure*: Generation-based approaches often fail to preserve the original source content across frames when spatially extending video frames. As a result, visual content may disappear or fail to appear consistently in the generated regions, leading to temporal inconsistency throughout the video sequence. (c) *Contextual hallucination*: Due to the limited contextual cues available for video outpainting, generation-based methods may hallucinate semantically irrelevant or unrealistic content in the outpainting regions. Such hallucinations introduce visual artifacts and intra-frame inconsistency.

2. Related Work

Flow-based propagation. Videos inherently exhibit strong spatio-temporal continuity, as adjacent frames share highly correlated structures and motion patterns. This motivates the use of optical flow to explicitly capture pixel-wise motion across frames. To preserve source content, flow-based propagation typically consists of three stages: flow estimation, flow completion, and flow-guided warping. Prior methods have improved these components individually or jointly. DFVI [22] completes optical flow fields using a Deep Flow Completion Network (DFC-Net) and propagates known pixels along the flow to fill missing areas. E2FGVI [12] jointly learns flow completion and propagation. ProPainter [28] enhances propagation through dual-domain integration. RGVI [4] introduces a one-shot pixel pulling strategy to mitigate error accumulation. Although these approaches have achieved notable success in video inpainting, they are inherently limited when extended to video outpainting due to the lack of observable context and the large spatial extent of missing regions. In addition, their pixel-level warping relies on sequential flow composition across intermediate frames, resulting in higher computational cost as spatial resolution and video length increase. These limitations are illustrated in Fig. 2 and Fig. 3 (a).

Video outpainting. Unlike video inpainting, which fills in missing regions within a fixed resolution, video outpainting expands the spatial canvas by generating new content beyond the original frame boundaries. Previous works on video outpainting can be categorized into zero-shot and one-shot methods based on inference strategies.

Zero-shot methods generate spatially extended content directly from the input video without any fine-tuning on

the specific video. Dehan [5] estimates optical flow to warp background content from adjacent frames and fills the remaining regions using an image completion network. M3DDM [6] trains a 3D diffusion model on large-scale video datasets with mask modeling. Follow-Your-Canvas [2] distributes the generation process across spatial windows to overcome GPU memory constraints. However, these approaches often suffer from limited visual fidelity, temporal inconsistency, and blurry results. In particular, Li et al. [11] propose a flow completion network for video outpainting, yet their method directly conditions the generation process on completed flow, leading to flow dependency.

In contrast, one-shot methods perform fine-tuning on each specific input video before generating new content. MOTIA [19] performs input-specific adaptation to capture sequence-specific appearance and motion patterns. Unboxed [23] expands static scenes via 3D Gaussian Splatting and utilizes diffusion models for dynamic-region synthesis and temporal coherence. Although these methods achieve strong performance, they incur substantial inference overhead from sequence-specific adaptation and often produce repetitive or irrelevant content in the outpainting regions. Fig. 3 illustrates several limitations of existing methods.

Unlike prior approaches that handle propagation and generation separately, we propose a unified framework that jointly exploits their complementary strengths within a single end-to-end framework. By integrating propagation-driven source preservation with diffusion-based content generation, our method preserves original structures while providing additional spatial cues during generation. Despite operating in a zero-shot manner, our framework outperforms existing one-shot approaches while achieving faster inference with reduced computational resources.

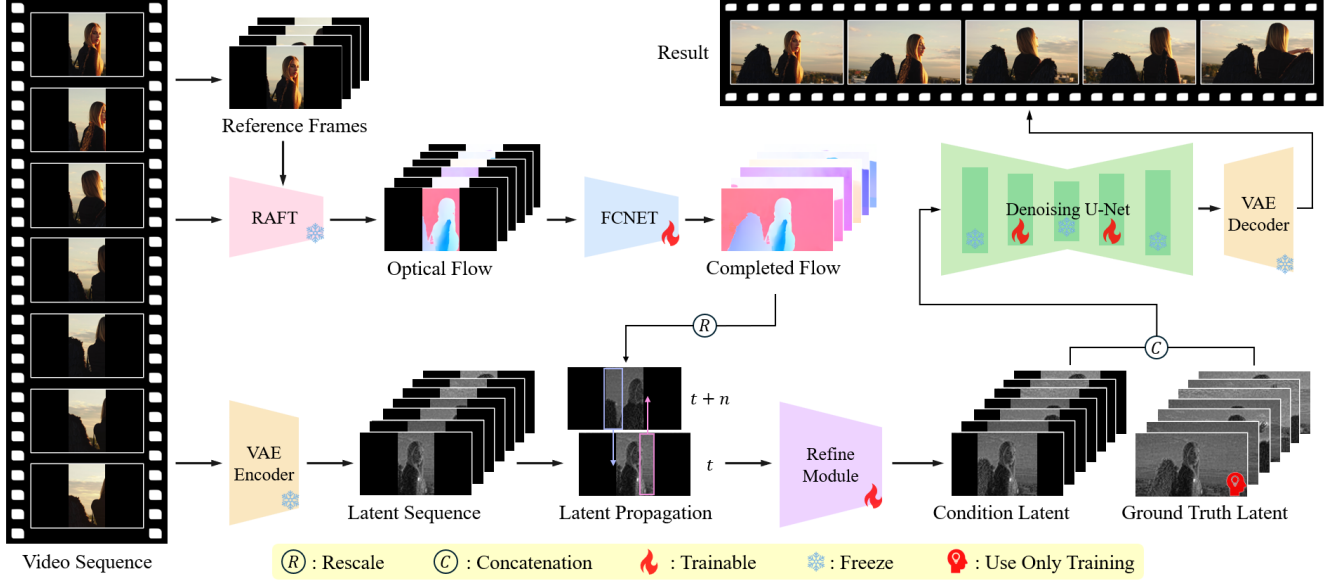


Figure 4. **Overview of the proposed Seen-to-Scene framework.** Seen-to-Scene unifies propagation and generation for coherent video outpainting. Given an input video, reference frames are selected and optical flow is estimated using RAFT [17], followed by flow completion with FCNet [28]. The input frames are encoded into latent representations via a pre-trained VAE encoder. Using the completed flow, we perform reference-guided latent propagation to preserve source content across frames. A lightweight refinement module is then applied to mitigate propagation artifacts. The refined latents serve as conditions for a diffusion-based generation process, where a denoising U-Net synthesizes the outpainted content in latent space. The final video frames are reconstructed using a VAE decoder.

3. Preliminary

Diffusion Model. A typical diffusion model consists of three components: a pre-trained VAE encoder [9] that maps each video frame into a latent code, a denoising network (e.g., a 3D-UNet or Transformer [16]) that models the diffusion process in the latent space, and a VAE decoder that reconstructs video frames from the latent representations.

Given a video consisting of N frames, the encoder maps each frame into a latent code $\mathbf{z}_i$, forming the latent sequence $\mathbf{Z}_0 = \{\mathbf{z}_i\}_{i=1}^N$. In the forward diffusion process, Gaussian noise is progressively added to $\mathbf{Z}_0$ over T time steps. With a pre-defined noise schedule $\{\beta_t\}_{t=1}^T$, we define $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$, where α_t controls the noise level at step t , and $\bar{\alpha}_t$ represents the cumulative noise scaling factor that determines the relative contribution of the original signal in $\mathbf{Z}_t$. The noisy latent $\mathbf{Z}_t$ at time step t can be written as

$$\mathbf{Z}_t = \sqrt{\bar{\alpha}_t} \mathbf{Z}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (1)$$

where ϵ denotes Gaussian noise sampled from a standard normal distribution. The denoising network ϵ_θ is trained to predict the injected noise given $(\mathbf{Z}_t, t)$ by minimizing the standard diffusion training objective:

$$\mathcal{L} = \mathbb{E}_{t, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \|\epsilon - \epsilon_\theta(\mathbf{Z}_t, t)\|^2. \quad (2)$$

In the inference stage, the denoising network iteratively removes noise from $\mathbf{Z}_T$ to obtain the clean latent sequence $\mathbf{Z}_0$. The resulting latent sequence is then decoded by the VAE decoder to reconstruct the final video frames.

4. Method

We present **Seen-to-Scene**, a novel video outpainting framework that integrates propagation and generation within a unified architecture. We first formulate the problem in Sec. 4.1. We then describe the reference-guided latent propagation mechanism in Sec. 4.2. Finally, we describe the diffusion-based generation framework in Sec. 4.3.

4.1. Problem Formulation

Given a video sequence $\{I_i\}_{i=1}^N$ consisting of N frames, where each frame $I_i \in \mathbb{R}^{3 \times h \times w}$, our goal is to expand each frame beyond its original spatial canvas. Specifically, we aim to generate an expanded sequence $\{\hat{I}_i\}_{i=1}^N$, where $\hat{I}_i \in \mathbb{R}^{3 \times H \times W}$ with $H > h$ and $W > w$, while preserving both intra-frame and inter-frame consistency. The overall pipeline of the proposed framework is illustrated in Fig. 4.

4.2. Latent Propagation

Preserving source content across frames is essential for maintaining both intra-frame and inter-frame consistency. Observable content can be propagated across frames to provide reliable spatial cues that partially extend the outpainting regions. This insight motivates our latent propagation, which propagates visual information in latent space to guide the generation of coherent outpainting regions. We now describe our reference-guided latent propagation in detail.

Reference Frame Selection. Video sequences inherently exhibit strong temporal continuity, where adjacent frames share highly overlapping visual content. To avoid redundant

information while providing informative context for the outpainting regions, we select reference frames based on structural correlation. For an input sequence $\{I_i\}_{i=1}^N$, we define a temporal window of size m and construct a global reference chain $\mathcal{R} = [r_1, r_2, \dots, r_L]$, initialized with the first frame $r_1 = I_1$. At step l , given the current reference index r_l , we define the candidate index set within the temporal window as $\mathcal{C}(r_{l+1}) = \{r_l + 1, r_l + 2, \dots, \min(r_l + m, N)\}$. For each candidate frame I_j with $j \in \mathcal{C}(r_{l+1})$, we measure the structural correlation between I_{r_l} and I_j using the structure component of the Structural Similarity Index (SSIM) [20]. To suppress illumination and color variations and focus purely on structural similarity, we compute the structure term of SSIM on gray-scale images as follows:

$$s(r_l, j) = \text{SSIM}_{\text{structure}}(\text{gray}(I_{r_l}), \text{gray}(I_j)), \quad (3)$$

where $s(r_l, j)$ denotes the structural correlation score between frames I_{r_l} and I_j . We select the next reference frame by minimizing the score within the candidate set:

$$r_{l+1} = \arg \min_{j \in \mathcal{C}(r_{l+1})} s(r_l, j). \quad (4)$$

Intuitively, selecting the frame with the lowest spatial similarity encourages the reference chain to include frames that contain rich contextual information while avoiding redundant content from temporally adjacent frames. After selecting r_{l+1} , the temporal window is advanced to the new reference index and the same procedure is iteratively applied to extend the reference chain. When the number of candidates becomes smaller than the window size m , the final frame I_N is appended to $\mathcal{R}$ to complete the reference sequence.

Flow Estimation and Completion. To preserve source content via flow-based propagation, we estimate optical flow using RAFT [17] between each frame I_i and its nearest past $r^-(i)$ and future $r^+(i)$ reference frames, defined as:

$$\begin{aligned} F_{i \rightarrow r^-(i)} &= \text{RAFT}(I_i, I_{r^-(i)}) \quad r^-(i) = \max\{r_l: r_l \leq i\}, \\ F_{i \rightarrow r^+(i)} &= \text{RAFT}(I_i, I_{r^+(i)}) \quad r^+(i) = \min\{r_l: r_l \geq i\}. \end{aligned} \quad (5)$$

On the expanded canvas, we define a binary outpainting mask $\{M_i\}_{i=1}^N$ with $M_i \in \mathbb{R}^{1 \times H \times W}$, where $M_i(x) = 1$ denotes the outpainting region outside the original frame boundary. Each estimated flow is mapped onto the expanded canvas. The regions indicated by the outpainting mask are treated as missing areas to be completed. To complete the missing flows in the outpainting regions, we adopt a pre-trained flow completion network [28]. However, existing flow completion models for video inpainting are typically specialized for small and spatially contiguous missing regions. When applied to large outpainting masks, these models often exhibit a domain gap, resulting in spatially diffuse or inconsistent flow estimates. To mitigate this domain gap, we jointly fine-tune the pre-trained flow completion network within our end-to-end pipeline using outpainting masks as supervision for the missing regions. This joint

optimization allows the network to adapt to the large and open-ended missing regions in outpainting, thereby improving flow reliability beyond the original frame boundaries. As a result, we obtain completed optical flows $\{\hat{F}_i\}_{i=1}^{N-1}$ between each frame I_i and its reference frames $I_{r(i)}$.

Reference Guided Latent Propagation. Each frame I_i is encoded by the VAE encoder into a latent code $Z_i \in \mathbb{R}^{c' \times h/s \times w/s}$, where s denotes the spatial downsampling factor. Since the VAE encoder preserves spatial correspondence, the completed flows $\{\hat{F}_i\}_{i=1}^{N-1}$ in pixel space and the outpainting masks $\{M_i\}_{i=1}^N$ are downsampled by the same factor s . Each latent Z_i is placed on the expanded latent canvas according to the mask. To fill in the outpainting regions, we propagate latent codes from reference frames using the completed flows and one-shot pulling strategy [4]. For each target latent Z_i , latent codes are aggregated only along the pre-defined reference chain instead of across all frames. Specifically, starting from the nearest past and future reference $r^-(i)$ and $r^+(i)$, the corresponding reference latents are first propagated to the target latent Z_i via grid-based warping using the completed flows $F_{i \rightarrow r^-(i)}$ and $F_{i \rightarrow r^+(i)}$. We further extend the propagation range by recursively composing reference-to-reference flows, allowing temporally distant references to contribute without requiring direct flow estimation between all frame pairs. For references r_t ordered in time, the accumulated flows toward the target latent Z_i are defined as follows:

$$\begin{aligned} F_{i \rightarrow r_{t+1}^+}(i) &= F_{i \rightarrow r_t^+}(i) + \mathcal{W}(F_{r_t^+}(i) \rightarrow r_{t+1}^+(i); F_{i \rightarrow r_t^+}(i)), \\ F_{i \rightarrow r_{t-1}^-}(i) &= F_{i \rightarrow r_t^-}(i) + \mathcal{W}(F_{r_t^-}(i) \rightarrow r_{t-1}^-(i); F_{i \rightarrow r_t^-}(i)). \end{aligned} \quad (6)$$

Using the accumulated reference flows, we warp reference latents to the target frame under the corresponding source masks $\dot{M}_i = 1 - M_i$. This enables direct long-range propagation along the reference chain instead of sequential propagation, reducing the accumulation of warping errors.

Latent Alignment. Through latent propagation, each latent aggregates globally observable visual content across the video sequence. To further reduce the dependency on completed flow and mitigate local misalignments, we introduce a lightweight refinement module that refines latent alignment using appearance cues rather than motion information. The refinement module takes the propagated latent, the original latent, the outpainting mask, and the propagated mask as inputs. It refines locally misaligned structures by predicting residual sampling offsets and adaptive modulation weights that selectively adjust latent alignment within uncertain regions. Alignment is performed bidirectionally, and the two direction-specific latents are fused using a shallow convolution layer to produce the final refined latent representation. Additional details, including the architecture of the refinement module and pseudo code for the latent propagation procedure, are provided in Appendix C.

4.3. Video Diffusion Model

After latent propagation, the outpainting regions in the latents are partially filled with content propagated from source regions across frames. This propagated content serves as conditioning for the subsequent diffusion process.

Training Stage. For diffusion training, we encode the ground-truth video into latent codes $\{\dot{Z}_i\}_{i=1}^N$ using a pre-trained VAE encoder. The propagated latents $\{\tilde{Z}_i\}_{i=1}^N$, which contain source content partially propagated through reference-guided latent propagation, are concatenated with $\{\dot{Z}_i\}_{i=1}^N$ along the channel dimension to form the input to the denoising network. Gaussian noise is then added to the concatenated latents according to the diffusion timestep to obtain the noisy latents for training. The noisy latents are fused using a 1×1 convolution layer and fed into the 3D U-Net backbone. To leverage the strong spatio-temporal priors learned from large-scale video data, we fine-tune a pre-trained video diffusion model [1] for the video outpainting task. Specifically, all convolutional and spatial attention layers are frozen to preserve the strong spatial priors learned during large-scale pre-training, while only the temporal transformer blocks remain trainable. The temporal blocks leverage motion-aware propagation-conditioned latents to aggregate temporal context and perform temporal reasoning across frames. The denoising network is then optimized for outpainting using the objective defined in Eq. 2.

Inference Stage. During inference, the ground-truth latent used during training is replaced with Gaussian noise sampled from a standard normal distribution. This noise serves as the initial noise sample for the generation process. The propagated latents obtained from the reference-guided latent propagation are concatenated with the noise latent along the channel dimension. This conditioning provides contextual guidance that enables the diffusion model to generate coherent and plausible outpainting content. Unlike existing state-of-the-art approaches [2, 19], our approach does not require text prompts or additional conditioning signals during inference. The concatenated latents are then fed into the 3D U-Net to predict the outpainted latents. Finally, the VAE decoder reconstructs the final video frames. As a result, our method performs inference in a zero-shot manner.

5. Experiments

5.1. Experiment Setup

Implementation Details. For latent propagation, we set the window size to $m = 4$. For diffusion training, we construct 25-frame video clips. The model is trained using Adam [8] with an initial learning rate of 1×10^{-5} for 100K iterations, with a batch size of 1 on two NVIDIA RTX A6000 GPUs.

Training Datasets. Previous diffusion-based video outpainting methods are trained on large-scale or private datasets containing millions of video samples [3]. In contrast, our model is trained using only 100K video samples from the publicly available YouTube-VOS training set [21].

Evaluation Datasets. To align with existing benchmarks, we evaluate our method on the DAVIS [15] and YouTube-VOS [21] datasets. For DAVIS 2017, we test on 90 video sequences. The evaluation protocol for YouTube-VOS is not explicitly standardized in prior works [2, 23], which often rely on selected high-quality subsets. To avoid potential selection bias, we randomly sample 60 videos from the official test set. Following the standard protocol, we report horizontal outpainting results under mask ratios of 0.25 and 0.66 for all test videos. The complete list of sampled videos is provided in Appendix E to ensure reproducibility.

Evaluation Metrics. We evaluate our method using standard quantitative metrics, including PSNR, SSIM [20], LPIPS [26], and FVD [18]. Fréchet Video Distance (FVD) serves as the principal metric for video outpainting, as it captures both perceptual realism and temporal consistency by measuring the distributional distance between the generated and original videos. All evaluation metrics are computed following the same protocol as prior works [2, 19].

5.2. Quantitative Evaluation

Table 1 reports the quantitative comparison with existing video outpainting methods on the DAVIS and YouTube-VOS datasets. Our method outperforms prior approaches across all evaluation metrics on both datasets. In particular, on DAVIS, our method achieves the best performance with a PSNR of 21.95, LPIPS of 0.141, SSIM of 0.783, and FVD of 218.8, surpassing previous methods by a clear margin. Similarly, on YouTube-VOS, our method outperforms all baselines across all four metrics, achieving improvements of 6.7% in FVD and 19.7% in LPIPS, indicating superior perceptual quality and temporal stability. These results demonstrate the effectiveness of our reference-guided latent propagation in preserving source content while maintaining coherent spatio-temporal structures during video outpainting. Notably, all results are achieved without any scene-specific adaptation, highlighting the strong generalization capability of Seen-to-Scene across diverse video scenarios.

5.3. Qualitative Evaluation

Fig. 7 illustrates the qualitative comparison between our method and existing approaches. Although M3DDM [6] introduces masked modeling for data augmentation, it often produces blurry content or fails to generalize when generating regions distant from the observed area. MOTIA [19], which performs test-time adaptation to capture input-specific patterns, frequently leads to repetitive objects or background artifacts. Follow-Your-Canvas [2] occasionally generates hallucinated structures unrelated to the original video content. Fundamentally, these methods fail to preserve the source content observed in the original sequence. In contrast, our method faithfully preserves the source content while maintaining strong spatio-temporal consistency, producing coherent outpainting results even in challenging scenarios. Additional results are provided in Appendix E.

Table 1. **Quantitative evaluation on DAVIS and YouTube-VOS.** Following prior benchmarks, the source videos with different aspect ratios are outpainted to 256×256 resolution. Gray rows indicate methods requiring input-specific adaptation.

Method	Publication	DAVIS				YouTube-VOS			
		PSNR↑	LPIPS↓	SSIM↑	FVD↓	PSNR↑	LPIPS↓	SSIM↑	FVD↓
Dehan [5]	CVPRW'22	17.96	0.233	0.627	363.1	-	-	-	-
M3DDM [6]	ACMMM'23	20.26	0.204	0.708	300.0	20.93	0.186	<u>0.704</u>	266.2
MOTIA [19]	ECCV'24	20.36	<u>0.160</u>	<u>0.758</u>	286.3	19.91	0.205	0.64	313.3
Follow-Your-Canvas [2]	AAAI'25	20.80	<u>0.160</u>	0.726	242.8	<u>21.01</u>	<u>0.178</u>	0.699	<u>260.1</u>
Li et al. [11]	Neurips'25	<u>20.81</u>	0.184	0.725	<u>234.7</u>	-	-	-	-
OutDreamer [27]	Arxiv'25	20.30	0.174	0.757	268.9	-	-	-	-
GlobalPaint [14]	Arxiv'26	20.91	0.154	0.762	227.8	-	-	-	-
Ours		21.95	0.141	0.783	218.8	21.89	0.143	0.783	242.8

Table 2. Ablation on model components. (a). Video diffusion with fine-tuning only; (b). (a) + inpainting completion network with latent propagation; (c). (a) + flow completion network fine-tuning; (d). (c) + refinement module (our complete model).

Method	PSNR↑	LPIPS↓	SSIM↑	FVD↓
(a)	19.42	0.254	0.556	586.3
(b)	20.92	0.195	0.687	355.7
(c)	21.61	0.144	0.742	220.7
(d)	21.95	0.141	0.783	218.8

5.4. Ablation Study

To validate the design of Seen-to-Scene, we conduct an ablation study on the DAVIS dataset, as summarized in Table 2. Each component progressively improves spatio-temporal coherence and overall visual quality. The baseline, which fine-tunes only the pre-trained video diffusion model, maintains temporal continuity but struggles to preserve the source content, resulting in inconsistent scene layouts in the outpainting regions. Introducing latent propagation provides an explicit mechanism for preserving source content by propagating structural cues from the observed regions. Even when using a pre-trained flow completion network trained for video inpainting, this propagation significantly improves spatial consistency and scene layout stability across frames. When the flow completion network is further fine-tuned for video outpainting, the performance improves substantially, highlighting the impact of the domain gap on flow completion quality as illustrated in Fig. 5. Finally, the refinement module yields additional performance gains by suppressing minor artifacts using appearance cues.

Analysis of Domain Gap. To directly verify the necessity of fine-tuning the flow completion network for the video outpainting, we conduct a qualitative comparison using pixel-level warping results as illustrated in Fig. 6. When using a pre-trained flow completion network trained for video inpainting, the completed flow often exhibits noticeable distortions in the outpainted regions, leading to unrealistic geometric deformations and inconsistent object boundaries. In contrast, after fine-tuning the flow completion network for video outpainting, our method produces more reliable flow fields, effectively preserving scene geometry and maintaining spatial consistency in the generated regions.



Figure 5. Comparison of flow completion results between an inpainting pre-trained model and our outpainting fine-tuned model.



Figure 6. Visualization of warping results using an inpainting pre-trained and an outpainting fine-tuned flow completion network.

Analysis of Efficiency. Seen-to-Scene also demonstrates strong computational efficiency. As illustrated in Fig. 2 and Table 3, our latent-level propagation with a reference strategy significantly reduces both memory and computational overhead compared to pixel-level or sequential flow composition, resulting in faster inference.

Table 3. Inference time and peak memory usage.

Method	Inference time (s)↓	Peak memory (GB)↓
M3DDM	43.127	9.63
MOTIA	373.19	11.98
Ours	12.109	8.81

Analysis of Temporal Window Size. We study the impact of the temporal window size m . As shown in Fig. 8, we set



Figure 7. Qualitative comparison between Seen-to-Scene and State-of-the-art methods.

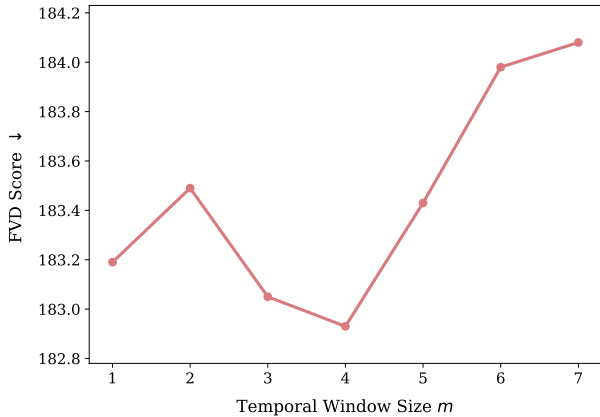


Figure 8. Effect of the temporal window size m on FVD.

$m = 4$ in our method, which effectively balances temporal consistency and computational efficiency.

6. Conclusion and Future Work

In this work, we present Seen-to-Scene, a novel end-to-end framework that unifies propagation-based and diffusion-based paradigms for video outpainting. We further introduce an efficient and effective reference-guided latent propagation scheme. By combining propagation-based and generation-based paradigms, our approach achieves spatially faithful and temporally consistent results across diverse scenarios. Extensive qualitative and quantitative results demonstrate that Seen-to-Scene significantly pushes the limits of zero-shot video outpainting. However, our method still has several limitations, including potential detail loss introduced by latent-space propagation and the reliance on implicitly fine-tuning the flow completion network. We hope our framework encourages future research toward more robust and scalable video outpainting.

Acknowledgement. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(No. RS-2024-00340745) and Korea Institute of Science and Technology (KIST) Institutional Program (26E0062).

References

- [1] 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, 6
- [2] Qihua Chen, Yue Ma, Hongfa Wang, Junkun Yuan, Wenzhe Zhao, Qi Tian, Hongmei Wang, Shaobo Min, Qifeng Chen, and Wei Liu. Follow-your-canvas: Higher-resolution video outpainting with extensive content generation. *arXiv preprint arXiv:2409.01055*, 2024. 2, 3, 6, 7
- [3] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024. 6
- [4] Suhwan Cho, Seoung Wug Oh, Sangyoun Lee, and Joon-Young Lee. Elevating flow-guided video inpainting with reference generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2527–2535, 2025. 1, 2, 3, 5
- [5] 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, 3, 7
- [6] 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. 2, 3, 6, 7
- [7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2
- [8] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [9] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4
- [10] Minhyeok Lee, Suhwan Cho, Chajin Shin, Jungho Lee, Sunghun Yang, and Sangyoun Lee. Video diffusion models are strong video inpainter. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4526–4533, 2025. 1
- [11] Ruilin Li, Hang Yu, and Jiayan Qiu. Dynamic shadow unveils invisible semantics for video outpainting. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 2, 3, 7
- [12] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17562–17571, 2022. 1, 3
- [13] Takuya Murakawa, Takumi Fukuzawa, Ning Ding, and Toru Tamaki. M3ddm+: An improved video outpainting by a modified masking strategy. *arXiv preprint arXiv:2601.11048*, 2026. 2
- [14] Yueming Pan, Ruoyu Feng, Jianmin Bao, Chong Luo, and Nanning Zheng. Globalpaint: Spatiotemporal coherent video outpainting with global feature guidance. *arXiv preprint arXiv:2601.06413*, 2026. 2, 7
- [15] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 6
- [16] 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. 4
- [17] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 4, 5
- [18] 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. 6
- [19] 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. In *European Conference on Computer Vision*, pages 153–168. Springer, 2024. 2, 3, 6, 7
- [20] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5, 6
- [21] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 585–601, 2018. 6
- [22] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3723–3732, 2019. 3
- [23] Zhongrui Yu, Martina Megaro-Boldini, Robert W Sumner, and Abdelaziz Djelouah. Unboxed: Geometrically and temporally consistent video outpainting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7309–7319, 2025. 2, 3, 6
- [24] Kaidong Zhang, Jingjing Fu, and Dong Liu. Flow-guided transformer for video inpainting. In *European conference on computer vision*, pages 74–90. Springer, 2022. 1
- [25] Kaidong Zhang, Jingjing Fu, and Dong Liu. Inertia-guided flow completion and style fusion for video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5982–5991, 2022. 1
- [26] 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. [6](#)
- [27] Linhao Zhong, Fan Li, Yi Huang, Jianzhuang Liu, Renjing Pei, and Fenglong Song. Outdreamer: Video outpainting with a diffusion transformer. *arXiv preprint arXiv:2506.22298*, 2025. [2](#), [7](#)
- [28] Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10477–10486, 2023. [1](#), [3](#), [4](#), [5](#)