

Appendix

A. Video Inpainting vs. Video Outpainting

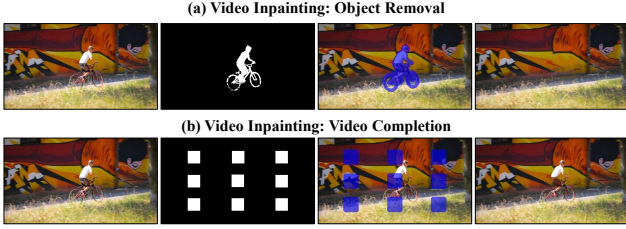


Figure 1. Two settings of video inpainting: (a) Object Removal eliminates dynamic objects and fills background regions; (b) Completion reconstructs locally missing areas within the visible frame.



Figure 2. Conceptual illustration of video outpainting, which generates new regions beyond the original frame.

Conceptual Difference between VI and VO. As illustrated in Fig. 1, video inpainting typically falls into two main settings: object removal and completion. The object removal setting aims to eliminate dynamic foreground objects such as humans or vehicles and fill the resulting holes using motion-consistent background propagation. In contrast, the completion setting assumes corrupted or missing regions caused by occlusions, sensor noise, or editing, and seeks to reconstruct them based on surrounding spatio-temporal context. Both settings share a common characteristic: they operate strictly within the original frame boundaries, where abundant spatial and temporal information is available from the observed regions. As a result, the task primarily involves filling missing areas using nearby context rather than generating unseen content. Moreover, since the reconstruction is constrained to the given resolution, the computational complexity remains stable regardless of frame size, making video inpainting relatively contained in both scope and difficulty.

As illustrated in Fig. 2, video outpainting fundamentally differs from inpainting in both objective and scope. Instead of restoring corrupted pixels within a given frame, it aims to extend the scene beyond the original field of view, synthesizing unseen structures and motion trajectories that remain unobserved in the input sequence. This requires the model to infer plausible geometry, depth, and dynamics of the environment while maintaining consistency with the visible

Table 1. Quantitative evaluation of ProPainter under the video outpainting setting.

| Method | PSNR \uparrow | LPIPS \downarrow | SSIM \uparrow | FVD \downarrow |
|----------------|-----------------|--------------------|-----------------|------------------|
| ProPainter [6] | 18.55 | 0.278 | 0.673 | 555.1 |

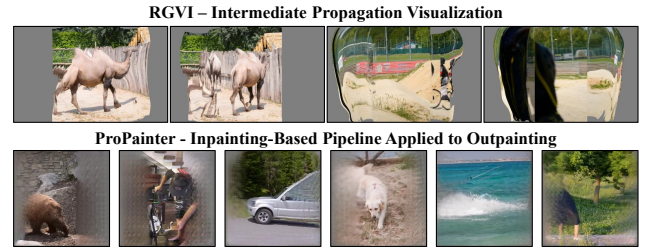


Figure 3. Limitations of Inpainting-Based Methods in Video Outpainting.

content. In other words, video outpainting transforms the problem from reconstruction to open-ended generation, demanding reasoning beyond directly observed data. Moreover, unlike inpainting, the computational cost of outpainting scales with the size of the expanded regions, as the model must process increasingly larger spatio-temporal domains during generation.

Experimental Comparison Between VI and VO. To validate the conceptual distinction between video inpainting and outpainting, we evaluate two representative inpainting-oriented propagation frameworks, RGVI [2] and ProPainter [6], under the outpainting setting. RGVI represents the current state of the art in flow-based propagation, while ProPainter is a widely adopted video inpainting model that integrates flow completion and content reconstruction into a unified pipeline. These models serve as strong baselines to analyze how inpainting-specific assumptions fail to generalize to the outpainting domain. As shown in Fig. 3, we visualize the internal pixel propagation stage of RGVI, which propagates information across frames. In the outpainting scenario, motion must be extrapolated far beyond the observed boundaries, resulting in flow bleeding, distorted warping, and inaccurate motion prediction around scene edges and dynamic objects. These artifacts expose the limitations of inpainting-trained flow completion networks in handling unobserved regions. We further evaluate ProPainter by applying its full inpainting pipeline to the outpainting task. Although it performs well for inpainting, it fails to extend the scene over large un-



Current Reference Frame



(a)



(b)



(c)



(d)

Next Reference Candidates

Figure 4. Comparison of fixed-stride and similarity-based reference selection. While the fixed-stride method selects frame (d) with redundant content, our method adaptively selects frame (a), which offers more informative cues for outpainting.

observed regions. The model cannot generate new structures or motion beyond the original frame boundaries, resulting in spatially limited and incoherent generation. This indicates that inpainting-based reconstruction pipelines are inherently constrained by their reliance on visible context and struggle to generalize to large-scale scene expansion. Quantitative results under a 33% horizontal outpainting setting further confirm these observations. Table 1 reports significant degradation in both perceptual and temporal quality when ProPainter is applied to outpainting. The notably poor FVD score indicates severe temporal inconsistency, underscoring the fundamental gap between reconstruction-oriented inpainting and generation-oriented outpainting.

Implications and Design Motivation. These findings demonstrate that, although propagation mechanisms developed for video inpainting provide useful priors for maintaining temporal consistency, they cannot be directly applied to video outpainting due to fundamental differences in both domain and objective. In contrast, diffusion-based video outpainting methods primarily rely on generative synthesis. While fine-tuning large diffusion models can enhance perceptual realism, such approaches often overlook structural fidelity and fail to preserve alignment with the original scene layout. We argue that effective video outpainting requires integrating the complementary strengths of both paradigms—the structural reliability of propagation and the generative flexibility of diffusion—while mitigating their respective inefficiencies. To this end, we propose *Seen-to-Scene*, an efficient propagation-guided diffusion framework that unifies structure propagation and content generation within a single diffusion process.

B. Reference Frame Selection

Why reference frames are needed. Video outpainting differs fundamentally from inpainting in that it must generate large outpainting regions while maintaining temporal and spatial coherence. To preserve structural consistency across frames, information from previously observed content should be propagated to the expanded regions. However, direct propagation from adjacent frames alone is often insufficient, since nearby frames tend to carry redundant information and contribute little to unseen areas. Moreover, propagating sequentially through all intermediate frames leads to excessive computational cost and accu-

mulated warping errors. Therefore, selecting a small number of representative reference frames that can provide diverse structural and motion cues is crucial for efficient and reliable outpainting.

Reference Frame Selection vs Fixed Stride. To evaluate the effectiveness of our reference selection strategy, we compare it against a fixed-stride selection scheme under identical video sequences. In the fixed-stride setting, reference frames are uniformly sampled at constant temporal intervals, regardless of scene dynamics. Such a strategy often leads to redundant reference choices, particularly when camera motion exhibits cyclic behavior (e.g., panning and returning to the same view). As a result, multiple references may contribute little new information for outpainting and may even fail to capture critical regions that should be propagated. In contrast, our similarity-based selection method adaptively identifies frames that provide complementary and non-redundant information relative to the target frame. This allows the propagation stage to effectively leverage structurally diverse cues from different temporal contexts, especially for expanding unseen regions. As illustrated in Fig. 2, while the fixed-stride method selects frame (d) as the next reference, our method dynamically chooses frame (a), which contains more relevant and distinctive content for the outpainting area. This example demonstrates that our approach not only reduces redundant references but also enhances the informativeness and diversity of propagated content, leading to more consistent and complete video outpainting.

Effect of window size on reference selection. We further investigate the impact of the temporal window size w used during reference selection. This experiment aims to evaluate whether our adaptive reference selection can maintain comparable performance to full-frame propagation while offering significantly improved computational efficiency. We conduct the evaluation on the DAVIS dataset using 48-frame sequences and measure performance across various window sizes. For comparison, we also include a baseline where propagation is performed for all frames without reference selection. Table 2 shows the quantitative results under different window sizes. Across all settings, the PSNR, SSIM, and LPIPS values remain identical, indicating that spatial quality and perceptual similarity are not affected by the choice of window size. This consistency

Table 2. Quantitative evaluation and efficiency analysis of different window sizes for reference selection on the DAVIS dataset

| Window Size (w) | PSNR \uparrow | LPIPS \downarrow | SSIM \uparrow | FVD \downarrow | Num. of Refs \downarrow |
|---------------------|-----------------|--------------------|-----------------|------------------|---------------------------|
| All Frames | 21.65 | 0.151 | 0.767 | 183.19 | 70.65 |
| $w = 2$ | 21.65 | 0.151 | 0.767 | 183.49 | 36.03 |
| $w = 3$ | 21.65 | 0.151 | 0.767 | 183.05 | 26.46 |
| $w = 4$ | 21.65 | 0.151 | 0.767 | 182.93 | 20.15 |
| $w = 5$ | 21.65 | 0.151 | 0.767 | 183.43 | 15.92 |
| $w = 6$ | 21.65 | 0.151 | 0.767 | 183.98 | 13.49 |
| $w = 7$ | 21.65 | 0.151 | 0.767 | 184.08 | 11.65 |

can be attributed to the design of our latent-space propagation. Since propagation occurs within the downsampled latent representation rather than at the pixel level, minor variations in reference selection have limited influence on pixel-domain metrics. Latents capture semantically aligned structure and motion cues that are robust to small temporal shifts, resulting in nearly identical spatial quality across different window configurations. However, the FVD scores exhibit slight fluctuations depending on w , reflecting the sensitivity of temporal dynamics to reference selection. The best FVD is obtained with $w = 4$, which strikes an effective balance between temporal coverage and motion alignment accuracy. When w is too small, reference diversity is reduced, increasing local flow accumulation errors. In contrast, excessively large windows include frames with greater motion variation, which introduces minor misalignment across long temporal gaps. This trade-off explains the gentle U-shaped trend in FVD as w increases, confirming that our method achieves optimal temporal coherence at $w = 4$.

Efficiency of Reference selection. Across all DAVIS sequences, the average video length is approximately 70.65 frames. As shown in Table 2, the number of selected references gradually decreases as the temporal window size w increases. This indicates that our method performs efficient reference sampling without the need to propagate through every frame sequentially. Instead, it achieves the same temporal coverage using a significantly smaller set of key reference frames, reducing redundant propagation while maintaining the necessary contextual information for outpainting.

C. Flow Completion Network (FCNet)

Domain gap between inpainting and outpainting. Although flow completion networks trained for video inpainting can effectively estimate motion within observed regions, they exhibit notable instability when applied to video outpainting. This discrepancy stems from a fundamental domain gap between the two tasks. Inpainting assumes dense visual context within bounded regions, where the network can infer missing motion based on adjacent observations. In contrast, outpainting involves large unobserved areas beyond the original frame boundaries, requiring motion extrapolation into entirely unseen regions. Consequently, inpainting-trained flow completion models tend to produce unreliable predictions near boundary extensions, resulting

in warped distortions, spatial discontinuities, and incomplete motion fields in the extrapolated zones. These artifacts propagate across frames and severely degrade temporal consistency in the generated videos.

Flow Visualization and Analysis. To further analyze the effect of domain adaptation, we visualize the flow completion results obtained from the inpainting pre-trained model and our outpainting fine-tuned version, as shown in Fig. 5. For a fair comparison, both models are trained without reference frame selection and follow the same training configuration. It is important to note that the goal of our flow completion is not to hallucinate flow in unseen source regions, but to accurately complete motion in the areas necessary for warping during propagation. The objective is therefore to generate stable and geometrically consistent motion fields that enable reliable information transfer from source to outpainting regions. The inpainting pre-trained flow completion network exhibits severe *flow bleeding* near boundary extensions or dynamic objects, where dynamic object motion incorrectly propagates into static background regions. In addition, its flow fields tend to rapidly lose magnitude toward the unobserved regions, indicating that the model struggles to infer motion continuity beyond visible boundaries. This limitation results in warped artifacts and incomplete propagation in the final outpainted frames. In contrast, our fine-tuned outpainting-specific flow completion network produces flow fields that are both spatially consistent and semantically aware of scene structure. Dynamic and static regions are clearly separated. Dynamic object motion is confined within its boundaries, and background flow remains stable without undesired bleeding. Moreover, the flow strength is maintained smoothly across extrapolated areas, allowing motion trajectories to extend naturally into the outpainting space. These results confirm that fine-tuning under the outpainting setting effectively adapts the flow completion process to handle large-scale scene expansion while preventing inter-region interference.

Propagated Results Visualization and Analysis. To directly verify the necessity of fine-tuning the flow completion network for the video outpainting domain, we conduct a qualitative comparison using pixel-level warping results, as illustrated in Fig. 6. Specifically, we propagate source content using two flow completion networks: (i) pre-trained for video inpainting and (ii) fine-tuned under the outpainting setting. For clarity of visualization, the warp-

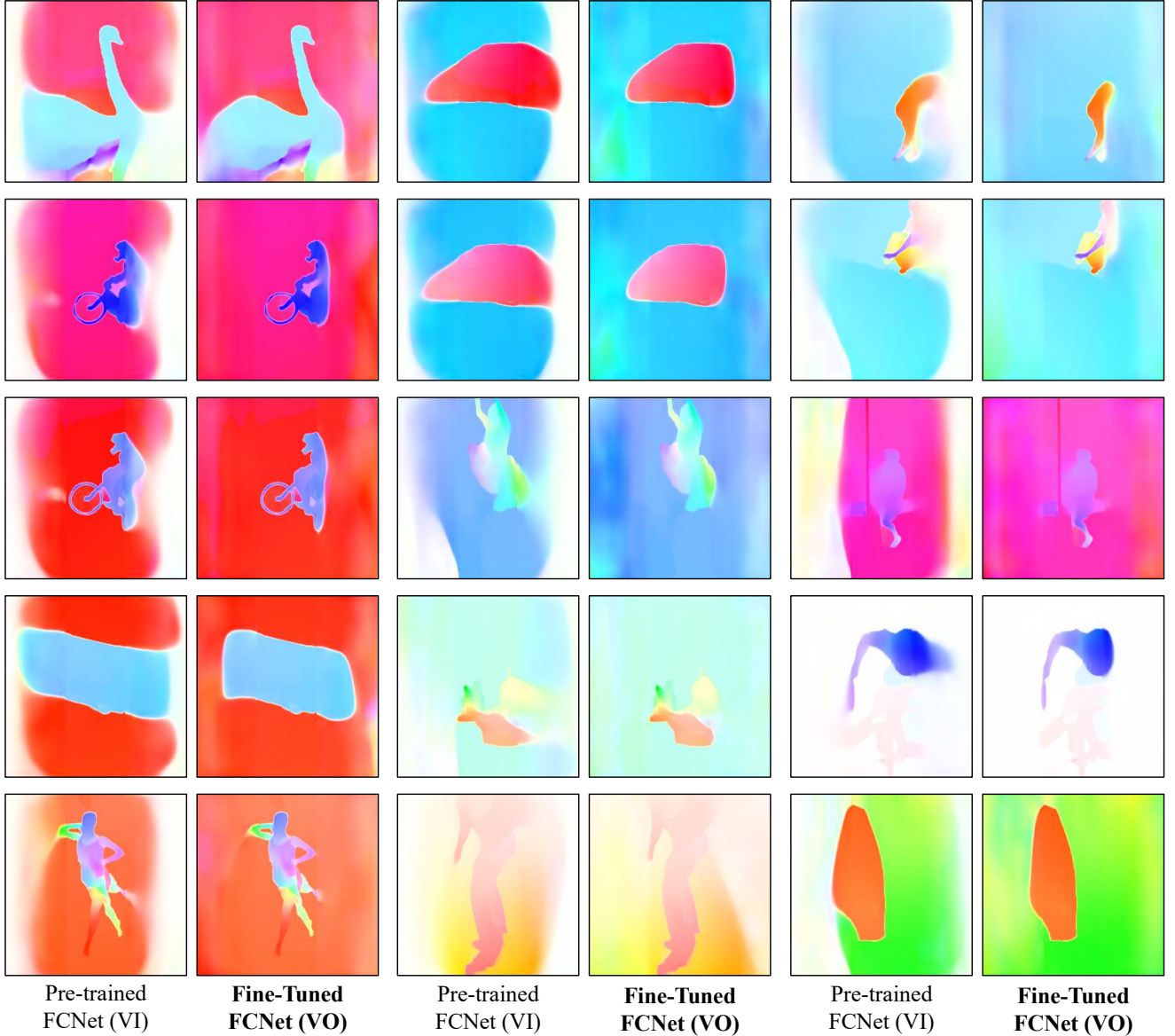


Figure 5. **Visualization of flow completion results.** Our outpainting fine-tuned model produces geometrically stable and semantically separated motion fields compared to inpainting-pretrained flow completion.

ing is performed at the pixel level rather than in the latent space. When the inpainting-pretrained network is used, the propagation fails to extend accurately over large unobserved regions, resulting in incomplete motion transfer and significant spatial distortion near the extrapolated boundaries. This limitation arises because inpainting-trained models are optimized for local completion within visible regions and thus cannot extrapolate motion trajectories beyond the original field of view. As a consequence, even source-visible structures are often misaligned or missing in the propagated results, and the distortion becomes more pronounced as the distance from the original frame increases. Moreover, unstable flow estimation for dynamic objects leads to trajectory ghosting and flow bleeding artifacts, produc-

ing residual trails and inconsistent motion patterns. In contrast, when the flow completion network is fine-tuned on the video outpainting setting, the propagated content is spatially coherent and extends seamlessly into the expanded regions. The fine-tuned model learns to infer long-range and cross-boundary motion more robustly, resulting in stable propagation even for highly dynamic scenes. As shown in Fig. 6, this adaptation effectively mitigates flow bleeding and geometric distortion, demonstrating that domain-specific fine-tuning is crucial for reliable motion estimation in video outpainting.

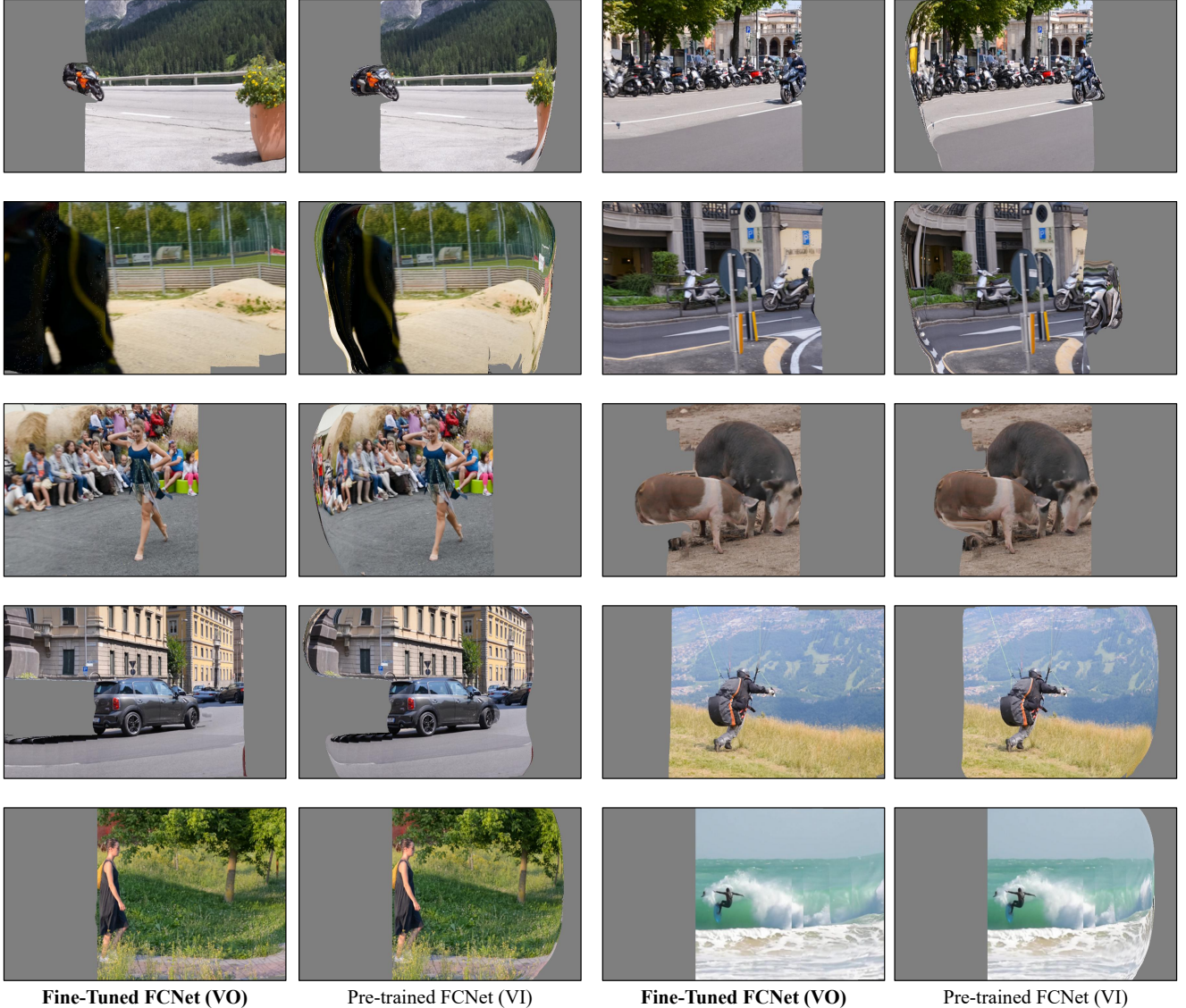


Figure 6. **Visualization of pixel-level warping using different flow completion networks.** Our outpainting fine-tuned model achieves stable propagation, while the inpainting pre-trained model fails to propagate content effectively to outpainting regions.

C.1. Latent Propagation

Reference-guided latent propagation. Instead of accumulating flow information across all intermediate frames, we utilize only a compact set of reference frames. Each reference frame is carefully selected to maximize informational diversity and to provide complementary content for the outpainting regions of the target frame. This design replaces dense sequential propagation with sparse, targeted correspondence, significantly reducing both temporal dependency length and the number of warping operations required. Formally, the computational complexity is reduced from $\mathcal{O}(N^2)$ pairwise propagation to $\mathcal{O}(N \cdot R)$, where R is the small number of reference frames ($R \ll N$). By restricting propagation to this compact set, redundant frame-to-frame accumulation is eliminated, which not only mit-

igates flow drift and error compounding but also leads to substantial improvements in runtime efficiency.

C.2. Latent Refinement

To refine warped multi-frame latents for video outpainting, we introduce a dual-branch refinement module that operates directly on the warped latent, the reference latent, and their associated binary masks. In the first branch, we concatenate the reference latent, the warped context latent, an outpainting mask, and a valid-warp mask, and use a lightweight CNN to predict spatially varying offsets and modulation masks for a deformable convolution, enabling the network to adaptively resample informative neighbors while implicitly correcting local misalignments in latent space. In parallel, a second branch processes the reference latent, warped latent, and outpainting mask using stan-

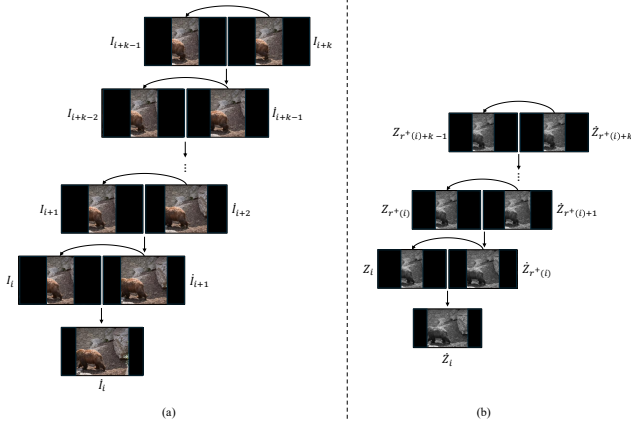


Figure 7. **Visualization of propagation strategies.** (a) Conventional sequential propagation accumulates flow across all intermediate frames, while (b) our reference-guided latent propagation performs efficient warping using only selected references in the latent space.

standard convolutions to produce a complementary refinement signal that emphasizes appearance and structural consistency around the outpainted regions. The outputs of the two branches are summed and passed through an additional convolutional layer to yield a refined latent, allowing the network to learn how to locally fuse deformation-aware and content-aware evidence. While we employ predicted optical flow to obtain the initial warped latents, our refinement module deliberately avoids using flow at this stage; instead, it learns offsets directly in latent space conditioned on visibility and outpainting masks. This design choice contrasts with prior refinement strategies that explicitly rely on optical flow fields, and is particularly important in the outpainting setting, where the predicted flow must be extrapolated into unobserved regions, making flow-based refinement susceptible to residual misalignments and artifacts. For bi-directional propagation, we apply the same refinement module independently to the forward-propagated and backward-propagated latents obtained from bi-directional warping, and then concatenate the two refined latents and pass them through a final convolutional layer to produce a unified, refined bi-directional propagation latent. By decoupling refinement from explicit flow and aggregating forward and backward evidence through content-adaptive latent offsets, our approach mitigates the propagation of errors from imperfect flow and yields propagated latents that are structurally well aligned between the source content and the bi-directionally warped regions.

C.3. Pseudo Code for Latent Propagation

For clarity, we also provide the pseudo code in Algorithm 1 summarizes the overall process of our Reference-Guided Latent Propagation. Here, I denotes the input video frame sequence and M represents the corresponding outpainting mask sequence that specifies outpainting regions to be generated. \mathcal{W} denotes a local temporal window used to select

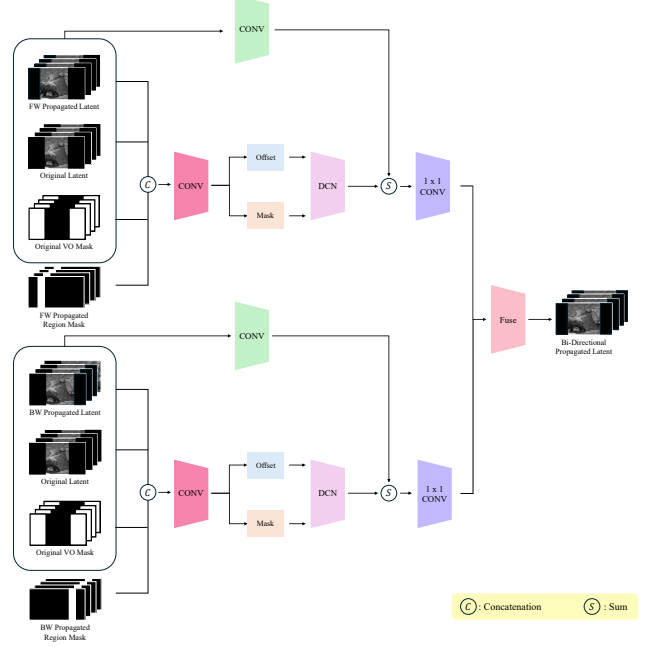


Figure 8. **Architecture of the latent refinement module.** It refines propagated latent features through spatial alignment and residual correction, enhancing temporal consistency and visual coherence in the outpainted video.

reference candidates, and \mathcal{R} indicates the selected reference frame set. F represents the optical flow field between frames, where FW and BW correspond to forward and backward directions in the temporal domain, respectively. Each z_i refers to a latent extracted from the corresponding frame, and Z_i denotes the final bidirectionally fused latent feature of frame i . The alignment module \mathcal{A} aligns propagated and original latent features using the propagated masks m , and \mathcal{F} denotes the fusion operator that integrates forward and backward aligned features to produce temporally consistent latent representations.

D. Details of Video Diffusion Models

D.1. Inference Stage

During inference, we replace the ground-truth latent used in training with a randomly initialized Gaussian noise, drawn from a standard normal distribution, as the initial conditional latent. This noise serves as the starting point for generation, enabling the diffusion model to generate unseen content beyond the original frame boundary. To maintain structural and temporal coherence, the propagated latent features obtained from the reference-guided latent propagation are concatenated with the noise latent along the channel dimension. This concatenated representation provides both generative flexibility and spatial guidance, allowing the model to produce realistic and consistent outpainting content. No text prompts or additional conditioning signals are used during inference. Unlike prior diffusion-based approaches [1, 3, 4], we do not perform any fine-tuning,

Algorithm 1 Reference-Guided Latent Propagation

```
1: Input: Video Sequence  $\{I_0, I_1, \dots, I_{T-1}\}$ ,  
2:   Mask Sequence  $\{M_0, M_1, \dots, M_{T-1}\}$ ,  
3:   Window size  $w$   
4: Output: Latent Sequence  $\{Z_0, Z_1, \dots, Z_{T-1}\}$   
  
5: Initialize:  $\mathcal{R} = \{I_0\}$ ,  $c = 0$   
  
6: while  $c < T - 1$  do  
7:   Define  $\mathcal{W} = \{I_{c+1}, \dots, I_{c+w}\}$   
8:   Select  $r = \arg \min_{I_j \in \mathcal{W}} \text{SSIM}(g(I_c), g(I_j))$   
9:   Append  $r$  to  $\mathcal{R}$ ;  $c \leftarrow r$   
10: end while  
  
11: Flow Extraction and Completion  
  
12: for  $i = 0, \dots, T - 1$  do  
13:   for each reference frame  $r > i$  do  
14:     Accumulate flow  $F_{r \rightarrow i} = F_{r \rightarrow r-1} \circ \dots \circ F_{i+1 \rightarrow i}$   
15:     Propagate latent feature  $z_r$  via  $F_{r \rightarrow i}$  to  $z_i$   
16:     Propagate mask  $m_r$  via  $F_{r \rightarrow i}$  to  $m_i$   
17:      $z_i^{\text{aligned}} = \mathcal{A}(z_i^{\text{orig}}, z_i^{\text{prop}}, m_i^{\text{orig}}, m_i^{\text{prop}})$   
18:   end for  
19:   for each reference frame  $r < i$  do  
20:     Accumulate flow  $F_{r \rightarrow i} = F_{r \rightarrow r-1} \circ \dots \circ F_{i+1 \rightarrow i}$   
21:     Propagate latent feature  $z_r$  via  $F_{r \rightarrow i}$  to  $z_i$   
22:     Propagate mask  $m_r$  via  $F_{r \rightarrow i}$  to  $m_i$   
23:      $z_i^{\text{aligned}} = \mathcal{A}(z_i^{\text{orig}}, z_i^{\text{prop}}, m_i^{\text{orig}}, m_i^{\text{prop}})$   
24:   end for  
25: end for  
26: for  $i = 0, \dots, T - 1$  do  
27:    $Z_i = \mathcal{F}(z_i^{\text{aligned}, FW}, z_i^{\text{aligned}, BW})$   
28: end for
```

test-sample-specific optimization, complex inference strategy, or post-processing refinement. Our framework operates in a purely feed-forward manner, relying solely on the propagation-driven structural priors to achieve coherent and temporally stable video outpainting.

D.2. Extension to Long Video Outpainting

Sampling a long video sequence in a single diffusion pass requires excessive memory and restricts the effective temporal context available to the model. To alleviate this limitation, we utilize a sliding-window-based sampling strategy that enables scalable inference while preserving temporal consistency. Given a video of length T , we define a set of overlapping temporal windows $\mathcal{V} = \{[s_i, e_i]\}_{i=1}^N$, where each window satisfies $e_i - s_i \leq W$, and adjacent windows overlap with a stride S . At each diffusion timestep t_k , we apply the U-Net to every window independently and then average the per-frame noise predictions to obtain a unified

global noise estimate:

$$\epsilon_t = \frac{1}{|\mathcal{V}(t)|} \sum_{i: t \in [s_i, e_i]} \epsilon_t^{(i)}, \quad (1)$$

where $\mathcal{V}(t)$ denotes the set of windows that include frame t . The averaged $\epsilon_{1:T}$ is then passed once to the scheduler's update step to refine the entire latent sequence.

E. YouTube-VOS Test Set

To ensure transparency and reproducibility, we disclose the list of sequences used from the YouTube-VOS test set. Unlike prior works [1, 3–5] that did not explicitly specify their evaluation subset, we randomly selected 60 sequences without any subjective bias to provide a fair and diverse benchmark for evaluating generalization performance. We release these sequence identifiers to facilitate future research, fair comparison, and reproducibility of our results.

0c7a4680db, 0d349f8286, 2e21c7e59b
2e129b0b09, 3b72dc1941, 3f2012d518
4b31a18d91, 4f5b3310e3, 5c3d2d3155
06a5dfb511, 6a75316e99, 6cced81d30
7daa6343e6, 8dea7458de, 9c4419eb12
13c3cea202, 24e2b52a4d, 37b4ec2e1a
37dc952545, 45fd60997a, 54ad024bb3
83a5056a16, 95ef69d827, 97b38cabcc
97fa40286c, 397dcc3a0, 459e70cd8e
03664dc880, 4035d3275c, 9787f452bf
40718bb478, 90949b2059, 547416bda1
607001c98f, 1320830fd2, 4348676053
6031809500, a9839ec6f2, ac4653b61d
b8bd20a472, b175cd8138, b492f67a89
b715879d5a, ba5dde67e9, c9ef04fe59
c16d9a4ade, cbea8f6bea, d1dd586cfd
d7ff44ea97, d59c093632, da9713ef3e
dab44991de, dc197289ef, e10236eb37
e1925724ab, eb49ce8027, f9eedb8691
f58212429d, fab725059c, fb104c286f

F. Additional Results

To demonstrate the robustness and temporal coherence of our approach, we present multiple frame samples generated from various sequences under different scenarios in Fig. 9. The results include both static and moving camera settings, single and multiple dynamic objects, and scenes containing humans, animals, and complex natural environments. Across these diverse cases, our method consistently produces temporally coherent and visually realistic outpainting results, preserving scene structure and motion continuity without noticeable flickering or spatial inconsistency.

We further visualize our approach under varying output resolutions, aspect ratios and outpainting direction. As shown in Figure 10, our model effectively adapts to spatially expanded canvases, preserving coherent scene geom-

etry and temporal consistency even at wide outpainting ratios. These results demonstrate that the proposed framework generalizes well beyond the training resolution, producing stable and high-quality content across diverse spatial configurations.

G. Supplementary Video Demonstrations.

We additionally provide demo videos in the supplementary materials, showcasing qualitative comparisons between our method and other existing approaches under identical experimental conditions. Our results exhibit realistic visual synthesis with superior temporal consistency and structural coherence, while effectively suppressing hallucination artifacts across diverse scenarios.

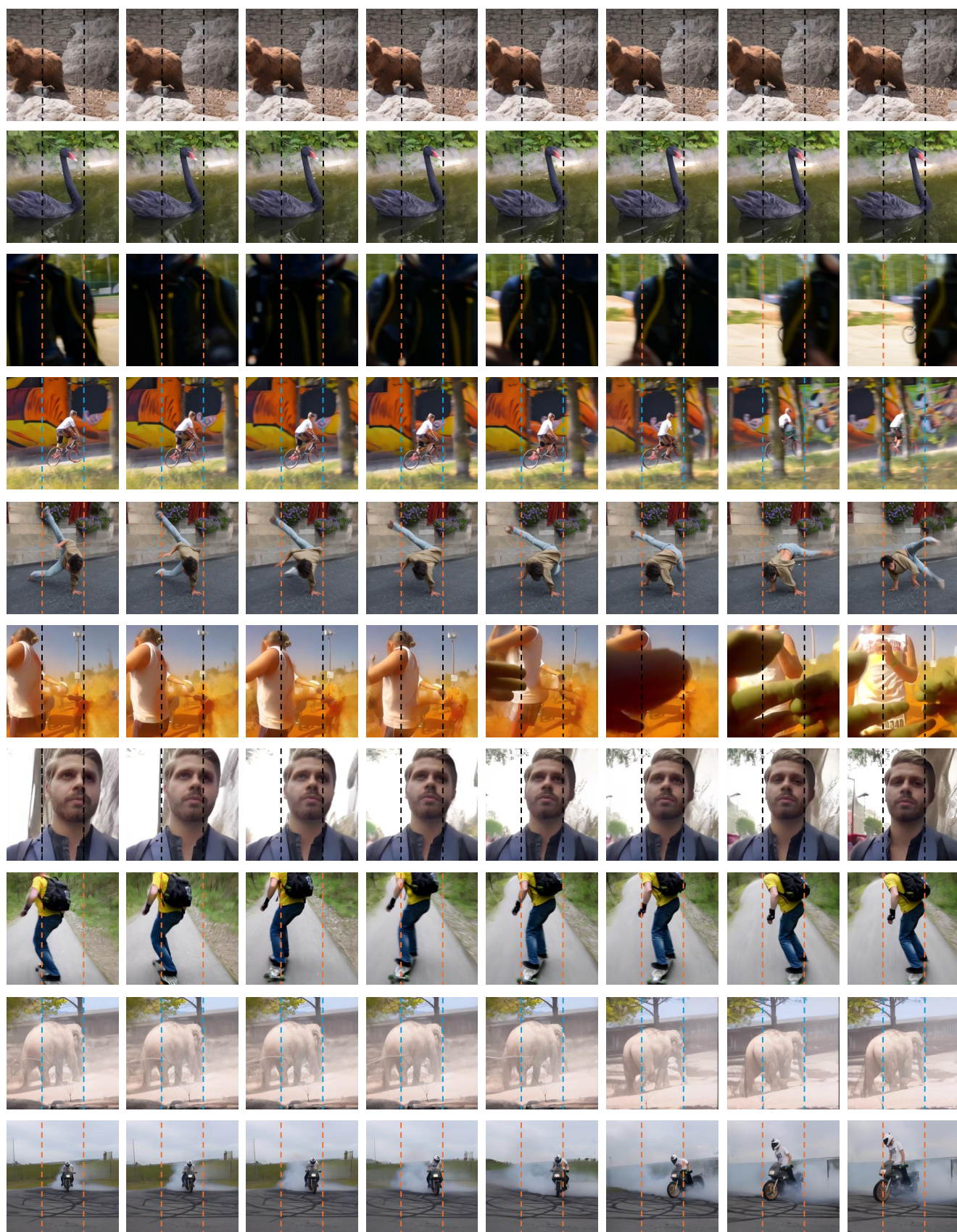


Figure 9. **Qualitative results of our method across diverse scenarios and environments.** The examples demonstrate the robustness and generalization capability of our model in handling various motion patterns, scene complexities, and outpainting configurations.

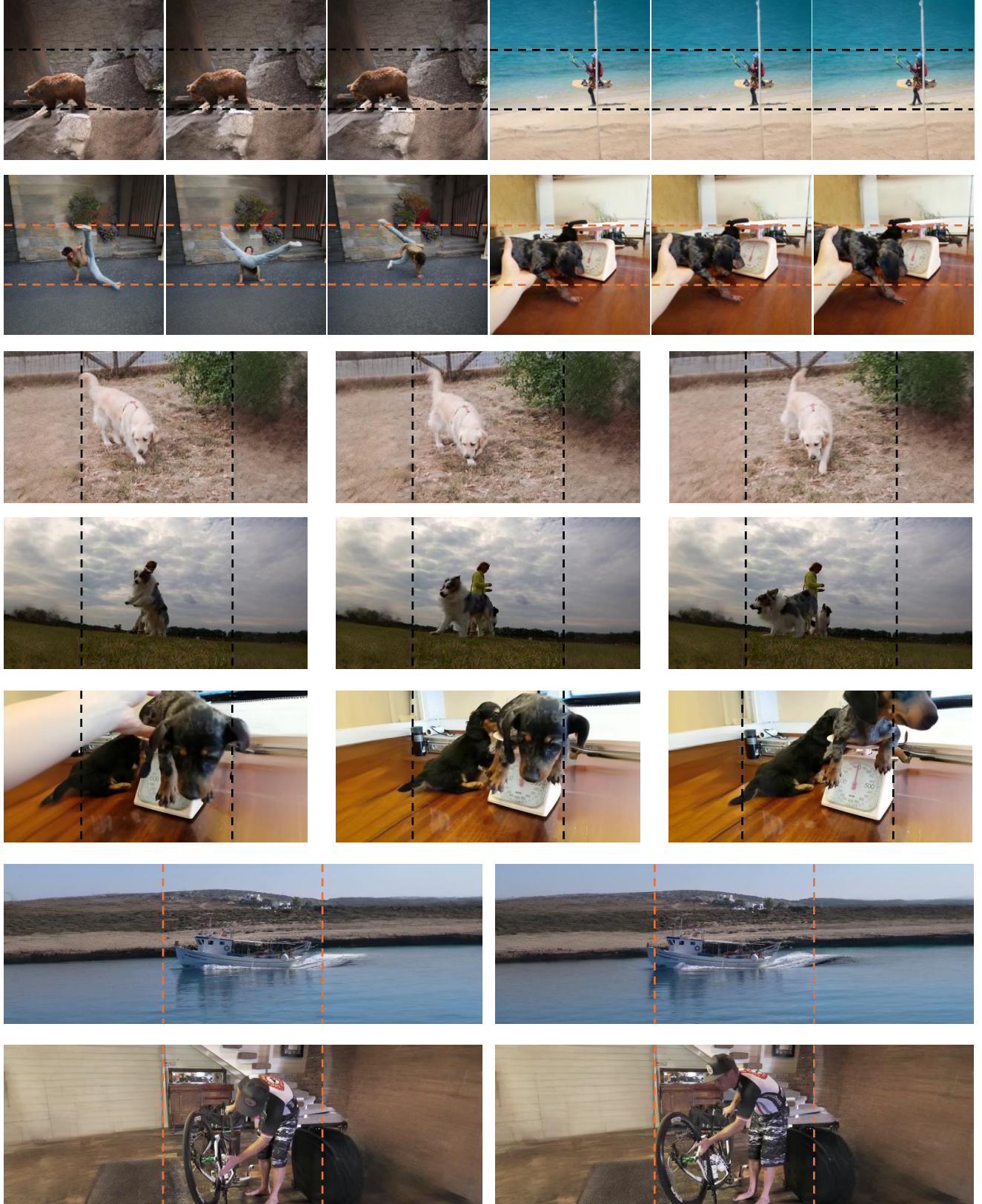


Figure 10. **Qualitative results across different spatial resolutions and aspect ratios.** Our method maintains consistent visual quality and structure from 256×256 to extended formats such as 256×512 and 256×768 .

References

- [1] 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. 6, 7
- [2] 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
- [3] 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. 6, 7
- [4] 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. 6
- [5] 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. 7
- [6] 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