WHAT IS THIS OBJECT AND WHERE IS IT GOING?

Anonymous authors
Paper under double-blind review

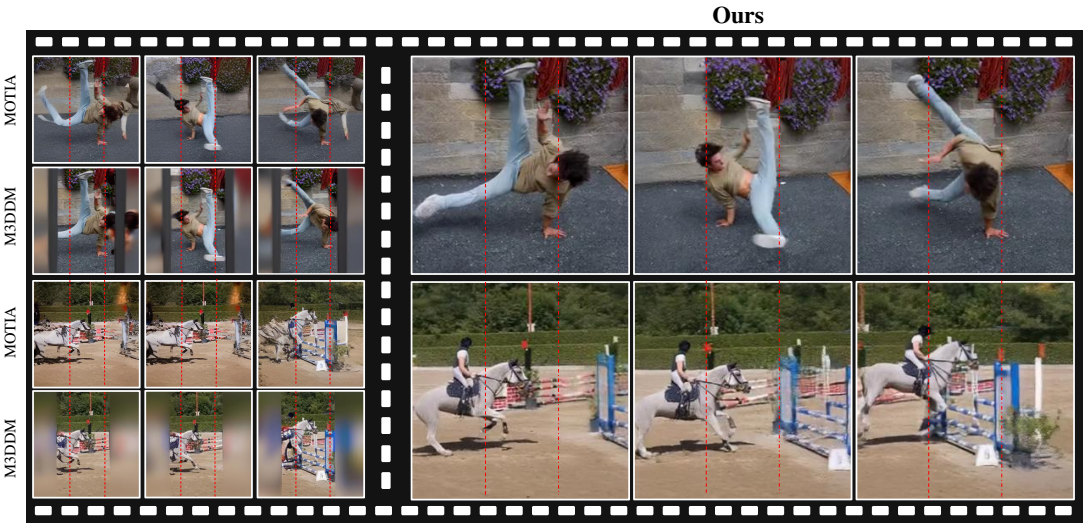


Figure 1: **Qualitative comparison of video outpainting results.** Our method better preserves object appearance and motion consistency than MOTIA and M3DDM. Red dashed lines indicate the input boundaries.

ABSTRACT

Video outpainting aims to extend content beyond the original frame boundaries while preserving inter-frame and intra-frame consistency. Existing methods have achieved promising results, but still fall short in complex scenes and motions. To address these issues, we propose RAVO, a retrieval-augmented test-time adaptation framework that complements incomplete object observations with relevant videos to support object-consistent video outpainting. However, directly exploiting retrieved videos is non-trivial due to discrepancies in visual characteristics and dynamics between the retrieved and source objects. To bridge this discrepancy, we introduce Object-Specific Adaptation (OSA), which learns transferable part-level correspondences between source and retrieved objects while retaining source-specific identity cues. Furthermore, conventional 3D attention tightly couples spatial and temporal interactions, limiting the effective incorporation of heterogeneous object-level cues during adaptation. We therefore propose Spatial-Temporal Decoupled Adaptation (STDA), which separately adapts spatial appearance and temporal dynamics to mitigate their mutual interference. Beyond these methodological contributions, we introduce ObjectBench to address the lack of benchmarks for object-level evaluation, providing an object-centric assessment of appearance and motion consistency. Extensive experiments show that RAVO achieves state-of-the-art performance on standard benchmarks and consistently outperforms existing methods on ObjectBench, demonstrating broad efficacy.

1 INTRODUCTION

As multi-platform environments continue to expand, content repurposing has become increasingly important for adapting videos to diverse viewing conditions. For instance, distributing videos across

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

different platforms often requires reformatting them to match different display aspect ratios such as adapting existing videos to horizontal or vertical formats. In the process, differences in aspect ratios can result in parts of the original video being cropped or empty regions appearing along the frame boundaries, posing a practical challenge in video editing and content adaptation.

Video outpainting, which extends videos beyond their original field of view, provides an effective solution to these issues. Early approaches include propagation-based video outpainting (Dehan et al., 2022) and GAN-based image expansion (Kimura & Rekimoto, 2018; Ying & Bovik, 2020). While these methods can extend observed content, they often struggle to generate content in unobserved regions. Motivated by these limitations and rapid advances in generative models have shifted the focus toward generation-based approaches. Utilizing diffusion models (Ho et al., 2020; 2022) with strong generative capability, they have achieved promising results.

Despite these strengths, existing methods still struggle to maintain object appearance when extending partially observed objects into the outpainting regions. Insufficient appearance cues lead to heavy reliance on pretrained generative priors, resulting in structural distortions or hallucinated object parts in Fig. 2. This limitation also affects motion consistency, as video generation relies on implicit temporal correlations across frames. Insufficient appearance information can make these correlations unreliable, leading to motion discontinuities and inconsistent trajectories. Together, these challenges compromise object-level consistency in video outpainting

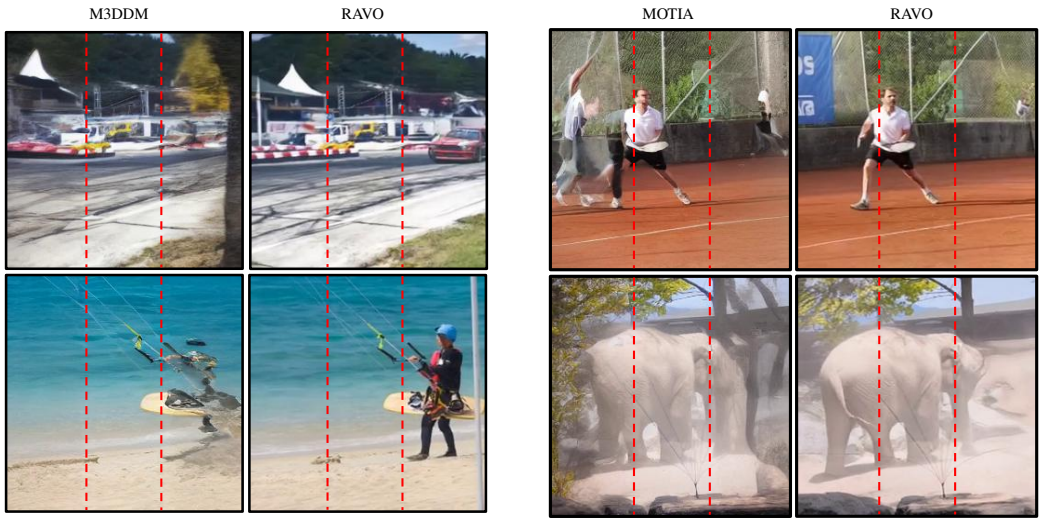
To address these issues, we propose RAVO, a retrieval-augmented test-time adaptation framework for video outpainting. Unlike prior approaches that rely on limited object information from the source video, RAVO retrieves semantically related object videos to provide additional object-specific guidance for partially observed objects during adaptation. Specifically, We first retrieve videos related to the source object from a large-scale video dataset (Wang et al., 2023), leveraging them as guidance for unobserved object regions. Since directly applying the retrieved videos is non-trivial due to discrepancies between the source and retrieved objects, we introduce Object-Specific Adaptation (OSA) to bridge these discrepancies. OSA partitions the retrieved objects into part-level regions and exploits diverse part-level cues to learn correspondences that identify transferable object information. Through this process, we alleviate appearance inconsistencies caused by limited object information while preserving source-specific characteristics.

To further leverage the resulting object information to improve motion consistency, RAVO adopts a two-stage adaptation process that separately handles spatial appearance and temporal dynamics. Conventional 3D attention jointly models spatial and temporal interactions (Ma et al., 2025), which can entangle appearance and motion learning, causing interference between them. We therefore introduce Spatial-Temporal Decoupled Adaptation (STDA), which separates the adaptation of these two aspects into distinct stages. This allows RAVO to leverage the enhanced appearance information through spatial-temporal decoupling, thereby improving object motion consistency across frames.

Together, these two designs substantially enhance object consistency in both appearance and motion. Additionally, to address the lack of benchmarks for object-level evaluation in video outpainting, we introduce ObjectBench to assess these improvements in terms of object appearance and motion fidelity across frames. Extensive experiments demonstrate that RAVO achieves state-of-the-art performance on standard benchmarks and outperforms existing methods on ObjectBench.

Our main contributions are summarized as follows:

- We introduce RAVO, a retrieval-augmented test-time adaptation framework for object-consistent video outpainting that complements incomplete object observations with reference videos. To this end, we develop Object-Specific Adaptation (OSA) to identify transferable part-level object information while preserving source-specific characteristics, and Spatial-Temporal Decoupled Adaptation (STDA) to separately adapt spatial appearance and temporal dynamics while reducing their mutual interference.
- RAVO achieves strong performance on standard video outpainting benchmarks and ObjectBench. In particular, we introduce ObjectBench for object-centric evaluation of appearance fidelity and motion consistency, and demonstrate consistent improvements over existing methods in object-level consistency.



(a) Object Structural Distortion

(b) Object Hallucination

Figure 2: **Failure examples of existing video outpainting methods.** Existing methods can still suffer from generation failures in outpainting regions, such as object hallucination and object distortion. In contrast, RAVO effectively preserves object appearance. Red dashed lines indicate the boundaries of the input video.

2 RELATED WORK

Video Diffusion Models. Video generation has progressed rapidly in recent years, with diffusion-based models becoming a dominant paradigm (Ho et al., 2022). Early approaches extended image diffusion models to video by adding temporal modules to U-Net architectures (Ho et al., 2022). More recent DiT-based models adopt 3D full attention to jointly model spatial and temporal information within a unified spatiotemporal representation (Peebles & Xie, 2023; Kong et al., 2024; Wan Team et al., 2025), leading to substantial improvements in visual quality and temporal consistency. Large-scale models such as HunyuanVideo, Sora, MovieGen, and Wan further demonstrate the effectiveness of this paradigm (Kong et al., 2024; Brooks et al., 2024; Movie Gen Team, 2024; Wan Team et al., 2025). Building on this progress, RAVO adopts Wan2.1-VACE (Jiang et al., 2025), which provides strong generative performance and flexible conditioning, as its backbone. This enables RAVO to utilize pretrained video priors while incorporating source-specific information for video outpainting.

Video Outpainting. Early video outpainting methods primarily relied on explicit propagation to extend observed content beyond the frame boundary. Dehan (Dehan et al., 2022) uses optical flow to warp background information from neighboring frames and completes the remaining regions with an image completion network. With the emergence of diffusion models, later approaches increasingly shifted toward generative formulations. M3DDM (Fan et al., 2023) trains a masked 3D diffusion model for video outpainting, while Follow-Your-Canvas (Chen et al., 2025) distributes generation across spatial windows to support large-scale, high-resolution expansion. More recent methods have focused on improving source-specific and spatiotemporal consistency.

MOTIA (Wang et al., 2024) performs input-specific adaptation to capture appearance and motion patterns from each source video. Unboxed (Yu et al., 2025) introduces 3D scene constraints for static regions and diffusion-based synthesis for dynamic content. Dynamic Shadow (Li et al., 2025) further models shadow-object relations and instance-aware motion cues to handle temporarily invisible objects in dynamic scenes. Seen-to-Scene (Jeon et al., 2026) combines propagation and diffusion-based generation within a unified framework to preserve observed content while synthesizing unseen regions. Despite these advances, existing methods still derive most of their outpainting cues from the source video itself or from pretrained generative priors. When the source contains only partial object observations, such information may be insufficient to recover object-specific appearance and

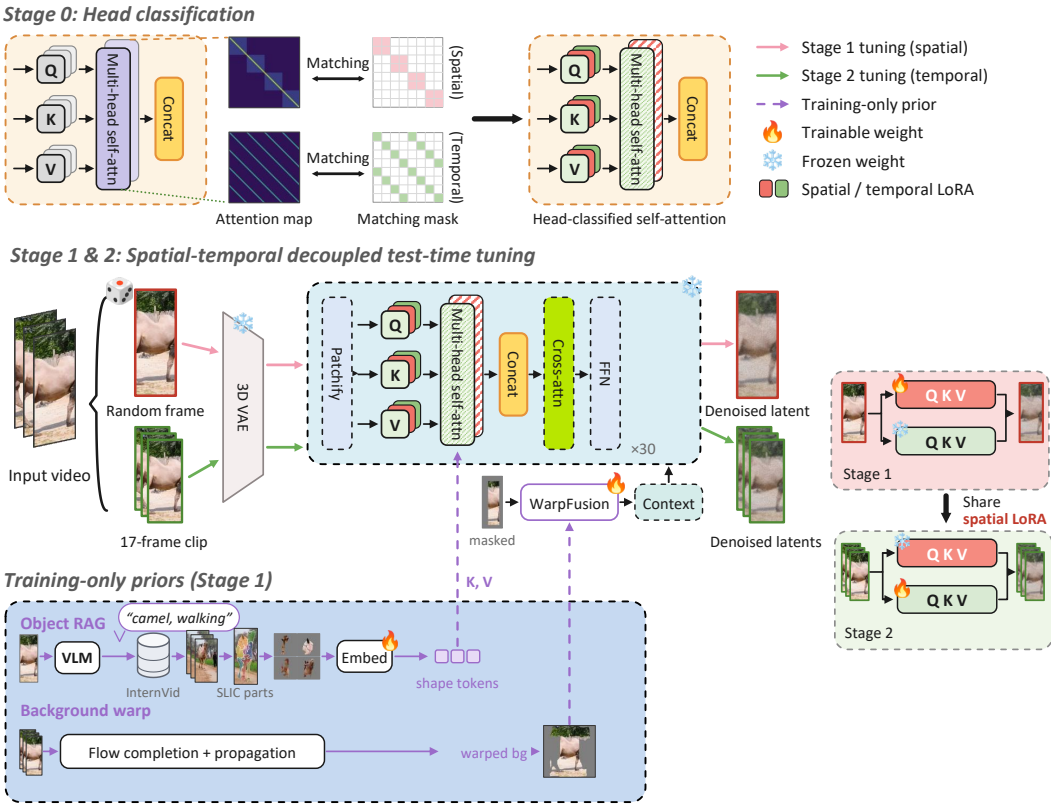


Figure 3: **Overview of RAVO.**

motion beyond the visible region. RAVO addresses this limitation by introducing retrieved videos as external object-level references during test-time adaptation.

3 METHOD

Given a video sequence of N frames, with each frame $I_i \in \mathbb{R}^{3\times h\times w}$, our goal is to generate an expanded sequence with frames $\hat{I}_i \in \mathbb{R}^{3\times H\times W}$, where $i = 1, \ldots, N$, $H > h$, and $W > w$, while preserving the original content and ensuring both spatial and temporal consistency.

As illustrated in Fig. 3, our framework builds upon the pretrained Wan2.1-VACE (Jiang et al., 2025) as its backbone and leverages external object information to complement insufficient observations in the source video. We retrieve related object videos from a large-scale database and employ Object-Specific Adaptation (OSA) to extract transferable local appearance cues while preserving source-specific characteristics (Sec. 3.1). The resulting guidance is incorporated through Spatial-Temporal Decoupled Adaptation (STDA), which adapts spatial attention heads using retrieved appearance information and source-derived background propagation, while separately optimizing temporal attention heads using consecutive source frames (Sec. 3.2).

3.1 Retrieval-Augmented Appearance Guidance

Partially observed objects provide incomplete object appearance information for outpainting, making it difficult to preserve object appearance in the outpainted regions. We therefore retrieve related object videos from an external database to supplement the missing appearance information. To bridge the appearance gap between the source and retrieved objects, we first identify suitable references, then adapt their local object regions as spatial guidance. We describe these two steps as Object-Aware Reference Retrieval and Object-Specific Adaptation.

Object-Aware Reference Retrieval. We employ a retrieval-augmented approach to learn object appearance priors at test time from retrieved references. Based on an object-centric description generated by a vision-language model (Bai et al., 2025), we retrieve relevant videos from a large-scale video database (Wang et al., 2023). Our retrieval pipeline first performs initial text-based retrieval using BM25 (Robertson & Zaragoza, 2009), followed by attribute-based filtering and object tracking. The candidates are then refined using CLIP (Radford et al., 2021) and ViCLIP (Wang et al., 2023) similarities to select a suitable reference. The selected video provides complementary appearance information for Object-Specific Adaptation.

Object-Specific Adaptation. Given the retrieved video, we first use SAM 3 (Carion et al., 2025) to isolate the target object from the background, retaining only the object region for subsequent processing, as shown in Fig. 3(b). We then apply SLIC (Achanta et al., 2012) within the object mask to partition the retrieved object into local superpixel regions. At each adaptation step, one region is randomly sampled with its RGB content and binary mask. The sampled region is encoded by the pretrained VAE encoder and concatenated with its corresponding mask. The combined feature is then processed by a learnable embedding module to obtain reference tokens. The source tokens provide the query Q_{src}, while the reference tokens are projected into key K_{ref} and value V_{ref} , which are concatenated with K_{src} and V_{src} as

$$Q = Q_{\mathrm{src}}, \qquad K = [K_{\mathrm{src}}; K_{\mathrm{ref}}], \qquad V = [V_{\mathrm{src}}; V_{\mathrm{ref}}]. \tag{1}$$

This formulation keeps the query based solely on the source tokens while allowing them to attend to local features from the retrieved object. The part-level reference tokens are used only during spatial adaptation.

3.2 Spatial-Temporal Decoupled Adaptation

The part-level reference tokens from OSA are used during spatial adaptation. Since VACE employs unified 3D self-attention to jointly model spatial and temporal interactions, spatial adaptation can also affect temporal dynamics. We therefore propose Spatial-Temporal Decoupled Adaptation (STDA), as illustrated in Fig. 3(c). STDA separately optimizes spatial and temporal attention heads to preserve appearance adaptation while maintaining temporal consistency.

Spatial-Temporal Head Classification. Following prior work on spatial-temporal attention decoupling (Ma et al., 2025), we classify the unified 3D attention heads into spatial and temporal groups based on their attention patterns. Specifically, each head is compared with pseudo spatial and temporal attention maps that characterize within-frame and cross-frame interactions, respectively. The resulting head partition is fixed throughout the subsequent adaptation stages.

Spatial Appearance Adaptation. Following the attention head classification, we perform two-stage adaptation by applying separate LoRAs to the spatial and temporal heads. In the first stage, we optimize the spatial LoRA θ_{spat} while keeping the temporal LoRA frozen. At each step, we uniformly sample a source frame I_i, where $i \in \{1, \ldots, N\}$, and randomly sample boundary ratios from $[0, 1]$ along its four sides to construct a pseudo-outpainting task. We define the enclosed area as the known region and mask the remaining area. The masked frame and its corresponding mask are then fed into the VACE conditioning pathway, allowing the model to learn source-specific appearance patterns by recovering the masked regions.

Additionally, we propagate observed background from neighboring frames to identify recoverable regions and those requiring generation. Specifically, we employ SAM (Carion et al., 2025) to separate the source object from the background and propagate only the latter in latent space. The propagated latents are paired with a propagation mask M_i^{prop}, indicating valid propagation locations outside the known region. We concatenate the propagated latents and their corresponding propagation mask to construct additional VACE conditioning features. These features are then concatenated with the existing VACE conditioning features. The combined features are fused through a learnable convolution before being passed to the existing condition embedder. This provides clearer spatial context for completing partially observed objects using reference tokens.

In addition to VACE conditioning, we obtain reference tokens R from a randomly sampled object part at each step. Following Eq. equation 1, their projected keys K_{ref} and values V_{ref} are concatenated with K_{src} and V_{src}, respectively, allowing Q_{src} to attend to diverse local appearance cues.

Together with VACE conditioning, this enables the model to leverage retrieved appearance information when reconstructing partially observed objects. We optimize the spatial LoRA using the following velocity prediction loss to reconstruct the sampled source frame:

$$\mathcal{L}_{\text{spat}} = \mathbb{E}_{i,m,r,\sigma,\epsilon}\left[\left\|v_{i,\sigma} - v_{\theta_{\text{spat}}}\left(z_{i,\sigma}, \sigma \mid c_i(m), R(r)\right)\right\|_2^2\right], \tag{2}$$

where z_i denotes the clean latent of the sampled source frame, $z_{i,\sigma} = (1-\sigma)z_i + \sigma\epsilon$ is its noisy latent, and $v_{i,\sigma} = \epsilon - z_i$ is the target velocity, where $\epsilon \sim \mathcal{N}(0, I)$. The model $v_{\theta_{\text{spat}}}$ predicts the velocity conditioned on the fused VACE conditioning $c_i(m)$ and reference tokens $R(r)$, where m and r denote the sampled outpainting mask and object part, respectively.

Temporal Adaptation. After spatial adaptation, we retain the learned spatial LoRA and freeze its parameters while optimizing the temporal LoRA to learn source-specific motion patterns. At each step, we randomly select a starting index $s \in \{1, \ldots, N - F_{\text{temp}} + 1\}$ and sample $F_{\text{temp}} = 17$ consecutive frames $\{I_s, \ldots, I_{s+F_{\text{temp}}-1}\}$ from the source video. We apply the same pseudo-outpainting strategy and background propagation as in spatial adaptation to reconstruct the masked source sequence. To further decouple temporal motion from spatial appearance, we employ a motion loss following prior work (Ling et al., 2024; Zhao et al., 2023). Specifically, we define the target and predicted motion representations as the differences between consecutive velocities, given by $\Delta v_{j,\sigma} = v_{j,\sigma} - v_{j-1,\sigma}$ and $\Delta \hat{v}_{j,\sigma} = \hat{v}_{j,\sigma} - \hat{v}_{j-1,\sigma}$, respectively.

We then define the motion loss using cosine similarity between the target and predicted motion representations:

$$\mathcal{L}_{\text{motion}} = \mathbb{E}_{s,m,\sigma,\epsilon}\left[1 - \text{CosSim}\left(\Delta v_\sigma, \Delta \hat{v}_\sigma\right)\right]. \tag{3}$$

Finally, the overall temporal adaptation objective is $\mathcal{L}_{\text{temp}} = \mathcal{L}_{\text{vel}} + \mathcal{L}_{\text{motion}}$, where $\mathcal{L}_{\text{vel}}$ follows the velocity prediction objective in spatial adaptation, applied to the sampled source sequence.

3.3 Benchmark Construction

Existing video outpainting metrics primarily evaluate frame-level fidelity or distribution-level realism. However, these metrics are largely influenced by background regions and do not explicitly evaluate the appearance and motion consistency of objects extending into generated regions. To address this limitation, we introduce ObjectBench, an object-centric benchmark for appearance and motion fidelity. We construct ObjectBench from 90 DAVIS sequences at two expansion ratios for comparison with existing video outpainting methods. The overall score is defined as

$$S_{\text{object}} = \sqrt{S_{\text{app}} \cdot S_{\text{motion}}}, \tag{4}$$

where S_{app} and S_{motion} denote object appearance and motion fidelity, respectively.

Object Appearance Fidelity. We employ DINOv2 (Oquab et al., 2023) features to evaluate object appearance fidelity against ground-truth (GT). Specifically, we extract object regions using segmentation masks and suppress background pixels. We then compute the cosine similarity between DINO features of the generated and GT objects at corresponding frames. The appearance fidelity score is defined as

$$S_{\text{app}} = \frac{1}{T}\sum_{t=1}^{T} \cos\left(\phi(O_{\text{gen}}^t), \phi(O_{\text{gt}}^t)\right), \tag{5}$$

where ϕ denotes the DINO feature extractor, and O_{gen}^t and O_{gt}^t represent the masked object regions from the generated and GT frames at time t, respectively.

Object Motion Fidelity. Following MTBench (Shi et al., 2025), we evaluate object motion using both local and global trajectory similarity. We sample query points within the GT object mask and independently track them in the ground-truth and generated videos using CoTracker (Karaev et al., 2024), with identical initial coordinates to establish point correspondence. Local motion fidelity measures the cosine similarity between consecutive displacement vectors, while global motion fidelity evaluates overall trajectory similarity using Fréchet distance (Eiter & Mannila, 1994). The

two scores are defined as

$$S_{\text{local}} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}\left(1+\frac{1}{|\mathcal{J}_n|}\sum_{t\in\mathcal{J}_n}\cos(\mathbf{u}_t^n,\mathbf{v}_t^n)\right),$$
$$S_{\text{global}} = \frac{1}{N}\sum_{n=1}^{N}\exp\left(-\frac{d_F(\mathcal{T}_{\text{gt}}^n,\mathcal{T}_{\text{gen}}^n)}{\sigma}\right). \tag{6}$$

where $\mathbf{u}_t^n$ and $\mathbf{v}_t^n$ denote the consecutive displacement vectors of the n-th tracked point in the ground-truth and generated videos, respectively. $\mathcal{J}_n$ represents its valid tracking intervals, while d_F denotes the Fréchet distance between the corresponding trajectories $\mathcal{T}_{\text{gt}}^n$ and $\mathcal{T}_{\text{gen}}^n$. The parameter σ controls the distance scaling. The final motion fidelity score is computed as $S_{\text{motion}} = \frac{S_{\text{local}}+S_{\text{global}}}{2}$



Figure 4: Qualitative comparison between RAVO and State-of-the-art methods.

4 EXPERIMENTS

4.1 EXPERIMENT SETUP

Implementation Details. We adopt Wan2.1-VACE (Jiang et al., 2025; Wan Team et al., 2025) as our pretrained backbone. The LoRA rank is set to 16 for both stages. We first train the spatial LoRA

Table 1: **Quantitative Comparison on DAVIS and YouTube-VOS.** Following prior benchmarks, our method achieves the best overall performance. Gray rows indicate methods requiring input-specific adaptation.

Method	Publication	DAVIS				YouTube-VOS			
		PSNR↑	LPIPS↓	SSIM↑	FVD↓	PSNR↑	LPIPS↓	SSIM↑	FVD↓
Dehan Dehan et al. (2022)	CVPRW'22	17.96	0.233	0.627	363.1	–	–	–	–
M3DDM Fan et al. (2023)	ACMMM'23	20.26	0.204	0.708	300.0	21.27	0.168	<u>0.749</u>	267.5
MOTIA Wang et al. (2024)	ECCV'24	20.36	<u>0.160</u>	<u>0.758</u>	286.3	20.87	0.176	0.745	319.5
Follow-Your-Canvas Chen et al. (2025)	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. Li et al. (2025)	NeurIPS'25	<u>20.81</u>	0.184	0.725	<u>234.7</u>	–	–	–	–
OutDreamer Zhong et al. (2025)	arXiv'25	20.30	0.174	0.757	268.9	–	–	–	–
GlobalPaint Pan et al. (2026)	arXiv'26	20.91	0.154	0.762	227.8	–	–	–	–
Ours	–	**21.98**	**0.131**	**0.775**	**221.3**	**22.11**	**0.130**	**0.802**	**251.4**

for 500 steps using AdamW (Loshchilov & Hutter, 2019), with a learning rate of 1×10^{-4} and a weight decay of 0.1. In the second stage, we freeze the spatial LoRA and train only the temporal LoRA for 500 steps, with the same learning rate and a weight decay of 0.5. All experiments are conducted on two NVIDIA RTX 3090 GPUs.

Evaluation Datasets. Following existing video outpainting benchmarks, evaluation is conducted on DAVIS (Pont-Tuset et al., 2017) and YouTube-VOS (Xu et al., 2018). For DAVIS 2017, 90 video sequences are used. As prior works (Chen et al., 2025; Yu et al., 2025) lack a standardized evaluation protocol for YouTube-VOS and often rely on selected high-quality subsets, 60 videos are randomly sampled from the official test set to minimize selection bias. Following the standard protocol, horizontal outpainting is evaluated at mask ratios of 0.25 and 0.66 for all test videos.

Evaluation Metrics. We evaluate our method using standard metrics, including PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), and FVD (Unterthiner et al., 2018). These metrics measure reconstruction quality, perceptual similarity, and the distributional differences between generated and original videos. All metrics are computed following the evaluation protocols of prior works (Chen et al., 2025; Wang et al., 2024). Additionally, we employ the proposed ObjectBench (Sec. 3.3) to evaluate object-level fidelity. Object Appearance Fidelity measures DINO similarity. Object Motion Fidelity averages local and global trajectory similarities. Their geometric mean yields the overall ObjectBench score.

4.2 Comparision with baseline

Quantitative Comparison. Table 1 presents the quantitative comparison with existing video outpainting methods on DAVIS and YouTube-VOS. Our method consistently outperforms existing approaches across all metrics on both datasets. Notably, it reduces LPIPS by 14.9% on DAVIS and 22.6% on YouTube-VOS compared to the best-performing baselines, while achieving a 7.1% improvement in SSIM on YouTube-VOS. These results demonstrate improved visual fidelity and temporal coherence in video outpainting.

Quantitative Comparison. Table 2 presents the ObjectBench evaluation on DAVIS under two mask ratios. Our method consistently outperforms existing video outpainting methods across all metrics, with particularly notable improvements under the challenging mask ratio of 0.66. These results demonstrate the effectiveness of our method in preserving object appearance and motion when objects extend beyond the observed regions.

Qualitative Comparison. Figure X presents a qualitative comparison between our method and existing video outpainting approaches. M3DDM struggles to generate coherent content beyond observed regions, while MOTIA and Follow-Your-Canvas often exhibit object distortions and temporal inconsistencies, particularly when objects are severely truncated or undergo large motions. In contrast, our method reconstructs plausible object appearances while maintaining coherent motion across frames. These results demonstrate the effectiveness of our method in preserving object appearance and motion consistency, even under challenging outpainting conditions.

Table 2: **ObjectBench evaluation on DAVIS.** Object appearance and motion consistency are evaluated under two mask ratios.

Method	Mask Ratio 0.25			Mask Ratio 0.66		
	$S_{\text{object}} \uparrow$	$S_{\text{app}} \uparrow$	$S_{\text{motion}} \uparrow$	$S_{\text{object}} \uparrow$	$S_{\text{app}} \uparrow$	$S_{\text{motion}} \uparrow$
M3DDM	0.8081	0.9285	0.7125	0.6391	0.7135	0.5853
MOTIA	0.8028	0.9245	0.7056	0.6419	0.7074	0.5938
Follow-Your-Canvas	0.8131	0.9300	0.7189	0.6565	0.7120	0.6166
Ours	**0.8332**	**0.9442**	**0.7422**	**0.7188**	**0.7832**	**0.6676**

4.3 Ablation Study

To validate our proposed design, we conduct an ablation study on the DAVIS dataset, with quantitative and qualitative results presented in Table 3 and Fig. 5, respectively. The baseline, which fine-tunes only the pre-trained video diffusion model, struggles to reconstruct objects extending beyond the observed regions. Introducing retrieved reference tokens provides additional shape information, improving the reconstruction of missing object structures, as illustrated in Fig. 5. In contrast, applying spatial-temporal decoupling alone helps preserve source-specific motion but remains limited in reconstructing severely truncated objects due to insufficient visual information. Finally, combining both components enables the model to recover missing object details while preserving the original motion patterns, achieving the best overall performance in Table 3.

Table 3: Quantitative ablation study on DAVIS, evaluating the contributions of retrieved references and spatial-temporal decoupling.

Method	PSNR↑	LPIPS↓	SSIM↑	FVD↓
Baseline	20.95	0.155	0.752	252.3
w/o Decoupling	21.25	0.158	0.758	275.7
w/o Retrieval	21.47	0.139	0.762	244.7
Ours	**21.98**	**0.131**	**0.775**	**221.3**



Figure 5: Visualizing the effects of retrieval and spatial-temporal decoupling.

5 Conclusion

In this work, we presented RAVO, retrieval-augmented video outpainting framework designed to improve object-level consistency in both appearance and motion. RAVO retrieves relevant reference videos to supplement missing visual information and employs object-specific adaptation to reconstruct partially observed objects. Spatial-temporal decoupled adaptation further separates appearance and motion learning to preserve source-specific motion. We also introduced ObjectBench to evaluate object appearance and motion fidelity. Extensive experiments demonstrate that RAVO improves video outpainting quality and object-level consistency over existing methods.

References

Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurélien Lucchi, Pascal Fua, and Sabine Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012. doi: 10.1109/TPAMI.2012.120.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng,

Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. OpenAI technical report, 2024. URL https://openai.com/index/video-generation-models-as-world-simulators/.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, et al. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. URL https://arxiv.org/abs/2511.16719.

Qihua Chen, Yue Ma, Hongfa Wang, Junkun Yuan, Wenzhe Zhao, Qi Tian, Hongmei Wang, Shaobo Min, Qifeng Chen, and Wei Liu. Infinite-canvas: Higher-resolution video outpainting with extensive content generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2150–2158, 2025. doi: 10.1609/aaai.v39i2.32213. Originally released as Follow-Your-Canvas (arXiv:2409.01055).

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 Workshops*, pp. 687–695, 2022. URL https://openaccess.thecvf.com/content/CVPR2022W/NTIRE/html/Dehan_Complete_and_Temporally_Consistent_Video_Outpainting_CVPRW_2022_paper.html.

Thomas Eiter and Heikki Mannila. Computing discrete Fréchet distance. Technical Report CD-TR 94/64, Christian Doppler Laboratory for Expert Systems, TU Vienna, 1994. URL https://www.kr.tuwien.ac.at/staff/eiter/et-archive/files/cdtr9464.pdf.

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*, pp. 7890–7900, 2023. doi: 10.1145/3581783.3612478.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. URL https://arxiv.org/abs/2204.03458.

Inseok Jeon, Minhyeok Lee, Seunghoon Lee, Minseok Kang, Suhwan Cho, and Sangyoun Lee. Seen-to-scene: Keep the seen, generate the unseen for video outpainting. *arXiv preprint arXiv:2604.14648*, 2026. URL https://arxiv.org/abs/2604.14648. Authors also identify this work as a CVPR 2026 Findings paper.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17202, 2025. URL https://arxiv.org/abs/2503.07598.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker: It is better to track together. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2307.07635.

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*, 2018. doi: 10.1145/3173574.3174001.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. URL https://arxiv.org/abs/2412.03603.

Ruilin Li, Hang Yu, and Jiayan Qiu. Dynamic shadow unveils invisible semantics for video outpainting. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://openreview.net/forum?id=irni27kAeP.

Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaian Chen, Jiaqi Wang, and Yi Jin. MotionClone: Training-free motion cloning for controllable video generation. *arXiv preprint arXiv:2406.05338*, 2024. URL https://arxiv.org/abs/2406.05338.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Yue Ma, Yulong Liu, Qiyuan Zhu, Xiangpeng Yang, Kunyu Feng, Xinhua Zhang, Zexuan Yan, Zhifeng Li, Sirui Han, Chenyang Qi, and Qifeng Chen. Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. *arXiv preprint arXiv:2506.05207*, 2025. URL https://arxiv.org/abs/2506.05207.

Movie Gen Team. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. URL https://arxiv.org/abs/2410.13720.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL https://arxiv.org/abs/2304.07193.

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. URL https://arxiv.org/abs/2601.06413.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. URL https://arxiv.org/abs/1704.00675.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

Qingyu Shi, Jianzong Wu, Jinbin Bai, Jiangning Zhang, Lu Qi, Yunhai Tong, and Xiangtai Li. Decouple and track: Benchmarking and improving video diffusion transformers for motion transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10995–11005, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Shi_Decouple_and_Track_Benchmarking_and_Improving_Video_Diffusion_Transformers_For_ICCV_2025_paper.html.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenges. *arXiv preprint arXiv:1812.01717*, 2018. URL https://arxiv.org/abs/1812.01717.

Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

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*, pp. 153–168. Springer, 2024. doi: 10.1007/978-3-031-72784-9_9. Method referred to as MOTIA.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Conghui He, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. InternVid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. URL https://arxiv.org/abs/2307.06942. Introduces ViCLIP.

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. doi: 10.1109/TIP.2003.819861.

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. *arXiv preprint arXiv:1809.03327*, 2018. URL https://arxiv.org/abs/1809.03327.

Zhenqiang Ying and Alan C. Bovik. 180-degree outpainting from a single image. *arXiv preprint arXiv:2001.04568*, 2020. URL https://arxiv.org/abs/2001.04568.

Zhongrui Yu, Martina Megaro-Boldini, Robert W. Sumner, and Abdelaziz Djelouah. Unboxed: Geometrically and temporally consistent video outpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7309–7319, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yu_Unboxed_Geometrically_and_Temporally_Consistent_Video_Outpainting_CVPR_2025_paper.html.

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*, pp. 586–595, 2018.

Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. MotionDirector: Motion customization of text-to-video diffusion models. *arXiv preprint arXiv:2310.08465*, 2023. URL https://arxiv.org/abs/2310.08465.

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. URL https://arxiv.org/abs/2506.22298.

A Appendix

You may include other additional sections here.