
Dynamic Shadow Unveils Invisible Semantics for Video Outpainting

Ruilin Li¹ Hang Yu^{1*} Jiayan Qiu²

¹School of Computer Engineering and Science, Shanghai University, China

²School of Computing and Mathematical Sciences, University of Leicester, United Kingdom
{ruilinli, yuhang}@shu.edu.cn, jq46@leicester.ac.uk

Abstract

Conventional video outpainting methods primarily focus on maintaining coherent textures and visual consistency across frames. However, they often fail at handling dynamic scenes due to the complex motion of objects or camera movement, leading to temporal incoherence and visible flickering artifacts across frames. This is primarily because they lack instance-aware modeling to accurately separate and track individual object motions throughout the video. In this paper, we propose a novel video outpainting framework that explicitly takes shadow-object pairs into consideration to enhance the temporal and spatial consistency of instances, even when they are temporarily invisible. Specifically, we first track the shadow-object pairs across frames and predict the instances in the scene to unveil the spatial regions of invisible instances. Then, these prediction results are fed to guide the instance-aware optical flow completion to unveil the temporal motion of invisible instances. Next, these spatiotemporal guidances of instances are used to guide the video outpainting process. Finally, a video-aware discriminator is implemented to enhance alignment among dynamic shadows and the extended semantics in the scene. Comprehensive experiments underscore the superiority of our approach, outperforming existing state-of-the-art methods in widely recognized benchmarks.

1 Introduction

Shadows are an integral part of many natural scenes, providing rich but often ignored cues about both invisible semantics. While shadows in static images can help infer approximate categories and coarse shapes of invisible objects [44], dynamic shadows in videos offer even more, as they contain spatiotemporal information of invisible instances like unveiling the motion and actions of objects and even dynamic light sources in the scene.

Despite recent video outpainting methods [5, 8, 36] making significant progress in generating contextually matching content with coherent textures and scene structure. However, this consistency often applies to broader backgrounds rather than specific instances: objects may lack detail, exhibit distortions, or fail to preserve identity across frames, or become temporally misaligned with dynamic shadows, leading to inconsistencies with the motion and action of extended semantics.

Specifically, given an input video, conventional video outpainting methods often suffer from temporal inconsistency and action misalignment when an instance is temporarily outside the visible scene. As shown in Fig. 1(a), while the shadow moves, conventional methods tend to generate instances that remain static. In contrast, our method aligns the generated instance with the motion of the dynamic shadow, thereby preserving consistency across frames. Additionally, conventional approaches often lead to action misalignment, as shown in Fig. 1(b), where the actions of the generated instance do

*Corresponding author

not correspond to the dynamic shadow in the scene. In contrast, our method captures the variations of the dynamic shadow and generates appropriate actions for the instance. Beyond addressing the aforementioned temporal inconsistency and action misalignment with dynamic shadows, our method is also capable of generating a harmonious light source, even when it is not explicitly visible in the ground truth, shown in Fig. 1(c).

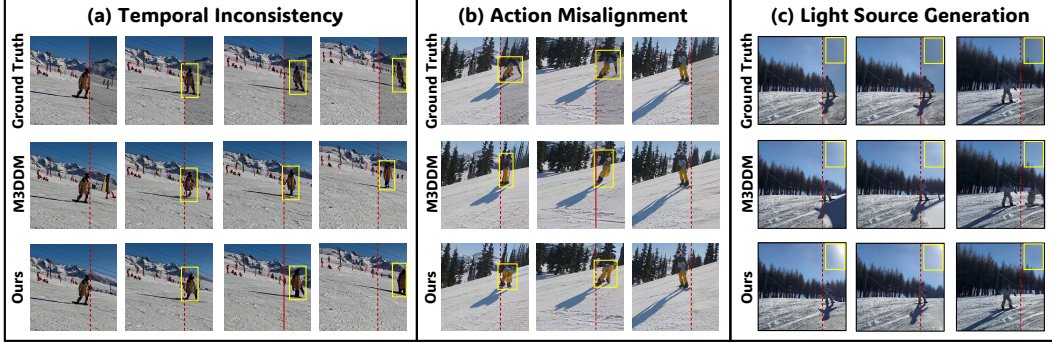


Figure 1: **Failure example of previous methods.** Many previous methods, including intensively trained models for video outpainting, often generate semantics that are temporally inconsistent with dynamic shadows and misaligned with them within the scene, as they primarily focus on maintaining coherent textures and visual consistency. In contrast, our method extends consistent semantics with dynamic shadow alignment and even generates a harmonious light across the frames.

To perform video outpainting by leveraging the hidden information in dynamic shadows, we propose a novel framework that explicitly models the associations between dynamic shadows and their corresponding instances, enabling instance-level spatiotemporal consistency even when the instances are temporarily invisible.

Specifically, we first track the shadow-instance pairs across frames and predict the instances outside the scene to unveil the spatial regions of invisible instances. Then, the prediction results are fed to guide the instance-aware optical flow completion to unveil the temporal motion of invisible instances. Next, the prediction results and the completed optical flow are used to guide the video outpainting process. Finally, a video-aware discriminator is implemented to enhance spatiotemporal alignment among dynamic shadows and the extended semantics in the scene.

Our main contributions are as follows:

- We propose a shadow-aware module that tracks and predicts the spatial region of instances, and completes optical flow in invisible areas to generate spatiotemporal guidance for outpainting.
- We introduce an instance-aware spatiotemporal guidance module that employs separate adapters for spatial predictions and completed flows to guide video outpainting, enhancing temporal consistency and semantic coherence.
- We present the first framework that explicitly leverages dynamic shadows for instance-aware video outpainting, by modeling the associations between shadows and their corresponding instances in the visible scene to unveil invisible semantics based on the cues of dynamic shadows.

2 Related Work

Video Shadow Processing. Conventional video shadow processing methods seek to track and segment dynamic shadows in complex scenes [50, 6, 2, 15]. Unlike static image shadow processing, which only focuses on individual frames, dynamic shadow processing enables the capture of spatiotemporal information in objects and light sources over time. Then, a sequence of methods [11, 18, 35] is developed to enhance the tracking and segmentation of moving objects. ViShadow [38] not only detects shadows and their corresponding objects in video frames, but also continuously tracks each shadow, object, and their relationships throughout the entire video sequence. Additionally, several methods in video shadow processing, such as [38, 52], not only track shadows and objects but

also enable video editing by removing instances, all while preserving temporal coherence. However, none of the existing methods deeply look into the spatiotemporal hidden information of dynamic shadows to generate invisible semantics.

Video Outpainting. Conventional video outpainting methods [5, 42, 41, 30] seek to extend the content beyond its original borders based on the dynamic scene, maintaining inter-frame and intra-frame consistency in videos. Although image outpainting [4, 14, 44, 24, 19] has been extensively studied, video outpainting still needs to be fully researched. Dehan [5] proposes a modified Generative Adversarial Network (GAN) [49] for video outpainting, performing flow estimation and background estimation before integrating them into a complete result. Recently, with the powerful generative capabilities of diffusion models [29, 22, 23, 25], some diffusion-based approaches have been introduced. M3DDM [8] presents global frame-guided training with a coarse-to-fine inference pipeline to handle the artifact accumulation issue. Meanwhile, MOTIA [36] proposes a test sample-specific fine-tuning strategy to learn the patterns of each sample. However, none of these methods can take advantage of the dynamic shadows in the scene to reasonably extend the invisible areas in the video.

3 Method

In this section, we detail the working scheme of the proposed video outpainting framework, which comprises four sequential stages, as illustrated in Fig. 2.

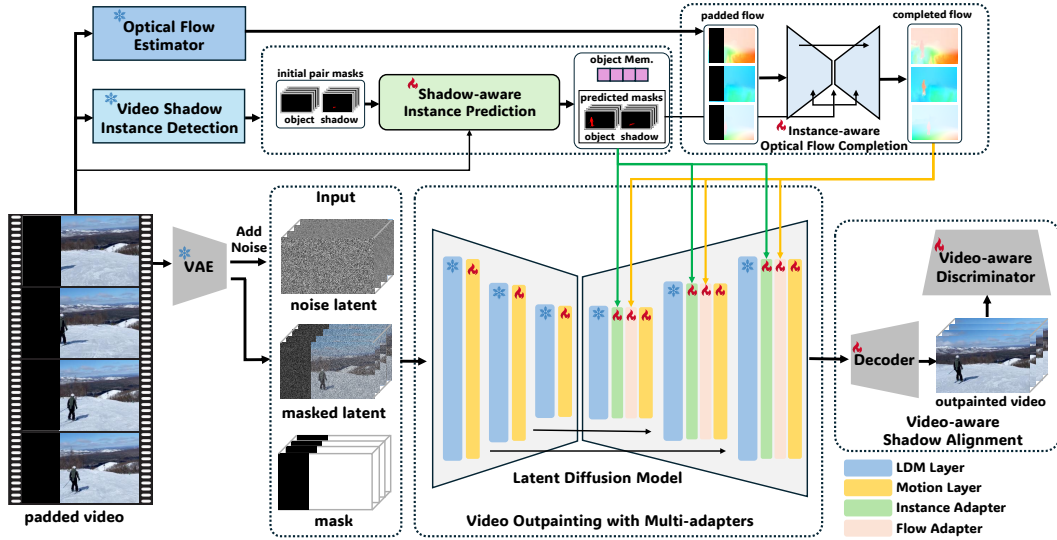


Figure 2: Illustration of the proposed framework. We first extract the shadow-object pair mask from the video and perform shadow-aware instance prediction to obtain spatial information of instances beyond the scene. This is then fed into an instance-aware optical flow completion module to generate temporal guidance for video outpainting. With the predicted spatiotemporal guidance, we perform video outpainting with two types of adapters. Finally, a video-aware discriminator is employed to enhance temporal consistency between the extended semantics and dynamic shadows.

3.1 Shadow-aware Instance Prediction

Tracking the mask of foreground instances in a video sequence is easy while using existing video object/instance segmentation methods [38, 27, 3]. However, predicting the masks of invisible instances that are temporarily outside the scene, either before they appear or after they leave, remains challenging due to their complex motions and actions.

To address this problem, we propose a novel shadow-aware instance prediction module that utilizes dynamic shadow information in the scene to recover the masks of instances while they are temporarily outside the visible scene.

Specifically, given the input video $V_{\text{in}} = \{I^0, I^1, \dots, I^T\}$, where each frame $I^t \in \mathbb{R}^{H \times W \times 3}$ and T is the total number of frames, we extend the video borders according to user-defined target aspect ratio to obtain the padded video $V = \{\hat{I}^0, \hat{I}^1, \dots, \hat{I}^T\}$, where each padded frame $\hat{I}^t \in \mathbb{R}^{(H+p_h) \times (W+p_w) \times 3}$ results from adding p_h pixels horizontally and p_w pixels vertically to the original frame. And we define $M_{\text{vis}} \in \{0, 1\}^{H \times W}$ as the visible region mask of the video, where 1 indicates the visible area and 0 indicates the non-visible area, which is defined as:

$$V_{\text{in}} = V \odot M_{\text{vis}}, \quad V_{\text{invis}} = V \odot (1 - M_{\text{vis}}) \quad (1)$$

where $\odot$ denotes the Hadamard product, V_{invis} denotes the region of V that needs to be outpainted, representing the content beyond the original field of view.

We first adopt a pretrained video shadow instance detection model [38] to obtain the initial visible shadow-object pairs from the input video, which is defined as:

$$\{(M_{\text{o_vis}}^{t,n}, M_{\text{s_vis}}^{t,n})\}_{t=1}^T = \text{Video_Instance_Detection}(V_{\text{in}}), \quad n = 1, 2, \dots, N \quad (2)$$

where we assume that each frame contains N shadow-object pairs, and each pair consists of an object mask $M_{\text{o_vis}}^{t,n}$ and its corresponding shadow mask $M_{\text{s_vis}}^{t,n}$, indicating the visible region of the object and its shadow at time t , respectively.

After obtaining the initial masks of visible shadow-object pairs in the scene, we aim to predict the shadow-object masks for instances that are temporarily outside the visible scene while simultaneously refining the initial tracking results, which may be inaccurate due to shadow overlap and discontinuity in complex dynamic scenes, as discussed in [44]. For simplicity, we denote the initial masks of a single shadow-object pair as $(M_{\text{o_vis}}^t, M_{\text{s_vis}}^t)$ at each time step.

To predict the instance masks outside the visible scene, we treat each shadow-object pair as a token in a spatiotemporal transformer [33], which is widely used in video understanding [12, 45, 51] for capturing long-range spatiotemporal coherence.

Specifically, we first apply a shared image encoder to extract high-level semantic features F_i^t of each padded frame $\hat{I}^t$. And the visible object mask and the shadow mask are fed it into a mask encoder to obtain their feature embeddings F_o^t and F_s^t , respectively, which are defined as:

$$F_i^t = \text{Image_Encoder}(\hat{I}^t) \quad (3)$$

$$F_o^t = \text{Mask_Encoder}(M_{\text{o_vis}}^t), \quad F_s^t = \text{Mask_Encoder}(M_{\text{s_vis}}^t) \quad (4)$$

where $\text{Image_Encoder}(\cdot)$ and $\text{Mask_Encoder}(\cdot)$ are convolutional networks (CNNs) designed to extract features from the input image and the respective masks. Note that a standalone image encoder helps reduce redundant computation over V when predicting multiple shadow-object pairs.

Next, we feed the extracted features of a single pair into the transformer for spatiotemporal fusion across frames. Specifically, for each shadow-object pair, we construct a sequence of length $T + 1$, where the 0-th input token is a learnable vector z^0 used to aggregate global information of this shadow-object pair, and the remaining T tokens correspond to the local instance features at each time step. For each time t , we concatenate F_i^t , F_o^t , and F_s^t along channels and feed the result into a feature encoder, obtaining the instance feature z_t of time t used for inter-frame fusion, which is defined as:

$$z^t = \text{Feature_Encoder}(\text{concat}(F_i^t, F_o^t, F_s^t)), \quad t = 1, 2, \dots, T \quad (5)$$

$$h^0, h^1, \dots, h^T = \text{Inter-Frame_Fusion}(z^0, z^1, \dots, z^T) \quad (6)$$

where $\text{Feature_Encoder}(\cdot)$ is a convolutional network designed to encode instance features of a single frame at time t , and $\text{Inter-Frames_Fusion}(\cdot)$ is a transformer-based module to fuse inter-frame information of instances. The 0-th output feature h^0 is the aggregated global embedding of the shadow-object pair, serving as shadow-object memory in the subsequent outpainting module, while the 1st to T -th output features h^t are inter-frame fused features capturing spatiotemporal correlations.

When multiple shadow-object pairs are detected, their predictions often influence each other, making it necessary to fuse spatial information across pairs. For object mask prediction, each pixel can belong to only one object. In contrast, for shadow mask prediction, a pixel may belong to several shadow regions, since shadows from different pairs may overlap.

To achieve inter-pair feature fusion within a single frame across pairs, we perform spatial fusion on the features h_i^t for all pairs at each time step t . The temporal fused features h_i^t are then passed

through another transformer for spatial fusion across shadow-object pairs, which is defined as:

$$\hat{h}^{t,1}, \hat{h}^{t,2}, \dots, \hat{h}^{t,N} = \text{Inter-Pair_Fusion}(h^{t,1}, h^{t,2}, \dots, h^{t,N}), \quad t = 1, 2, \dots, T \quad (7)$$

where each $\hat{h}^{t,n}$ is a spatiotemporal fused feature across frames and shadow-object pairs.

The spatiotemporal fused feature is passed through a lightweight mask prediction head to predict the object mask $\hat{M}_o^{t,n}$ and the shadow mask $\hat{M}_s^{t,n}$ of each shadow-object pair at each time step:

$$\hat{M}_s^{t,n}, \hat{M}_o^{t,n} = \text{Mask_Decoder}(\hat{h}^{t,n}), \quad t = 1, 2, \dots, T, \quad n = 1, 2, \dots, N \quad (8)$$

where $\text{Mask_Decoder}(\cdot)$ are convolutional networks designed to decode from each mask logit for the object and its corresponding shadow masks at time t .

Finally, the predicted outputs $\hat{M}_s^{t,n}$ and $\hat{M}_o^{t,n}$ are used to compute the training loss for supervising the object and shadow mask prediction in the completed scene. Additionally, the aggregated global feature h^0 serves as the memory embedding of a shadow-object pair and is passed to the video outpainting module to guide the recovery of the appearance of the object. Besides, the mask logits of each pair $\hat{h}^t$ are passed to the next module for guiding optical flow completion, facilitating instance-level temporal consistency outside the visible scene.

3.2 Instance-aware Optical Flow Completion

Pretrained flow completion modules are commonly used in video generation networks [47, 46, 13], as they provide a simpler solution for handling invisible flow compared to directly extending complex RGB content [40]. This approach leverages the inherent simplicity of flow completion to enhance video generation efficiency by using the completed flow to propagate pixels, thereby reducing the burden of video outpainting and better preserving temporal coherence.

To address this problem, we propose a novel instance-aware optical flow completion module, which recovers the instance-level optical flow outside the visible scene and provides temporal consistency guidance for the video outpainting process.

Specifically, we first use the pretrained optical flow estimation model RAFT [31] and the visible region mask M_{vis} to obtain the optical flow of padded video V and initialize these flows with Laplacian filling for the initialized optical flow O_{in} , which is defined as :

$$O_{\text{in}} = \text{Laplacian_Filling}(\text{RAFT}(V) \odot M_{\text{vis}}) \quad (9)$$

Then, the initialized optical flow is fed into a pseudo-3D U-Net [26] encoder-decoder structure with skip connections from the encoder to the corresponding layers in the decoder, which takes the initialized optical flow O_{in} and invisible region mask $(1 - M_{\text{vis}})$ and output the completed optical flow O_{out} , which is defined as:

$$F_{\text{flow}} = \text{Flow_Encoder}(O_{\text{in}}), \quad F_{\text{mask}} = \text{Mask_Encoder}((1 - M_{\text{vis}})) \quad (10)$$

$$O_{\text{out}} = \text{U-Net}(\text{concat}(F_{\text{flow}}, F_{\text{mask}}), \hat{h}^t) \quad (11)$$

where $\text{Flow_Encoder}(\cdot)$ is a convolutional network that extracts features from the initialized optical flow and the invisible region mask. The U-Net architecture is conditioned on the spatiotemporal features $\hat{h}^t$, which enables the network to incorporate instance-level mask features, enhancing the prediction of the completed optical flow outside the scene.

Specifically, to enable optical flow completion of the padded areas with instance-awareness, we introduce a cross-attention mechanism [33] after the temporal 2D convolutions in U-Net. This mechanism allows the network to effectively incorporate conditions from the dynamic foreground, enhancing its ability to complete the optical flow, which is defined as :

$$Q = W_Q(x), \quad K = W_K(\text{concat}(\hat{h}^1, \hat{h}^2, \dots, \hat{h}^T)) \quad (12)$$

$$V = W_V(\text{concat}(\hat{h}^1, \hat{h}^2, \dots, \hat{h}^T)) \quad (13)$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (14)$$

where x denotes the latent feature from the U-Net, and $\hat{h}^t$ represents the conditional mask features of time t . W_Q , W_K , and W_V are learned linear projection layers that map inputs into the query, key, and value spaces of dimension d_k .

Finally, we utilize the instance logits generated by the previous module to achieve instance-aware optical flow completion, particularly in regions with sharp motion boundaries in previously invisible areas. In regions around the padded boundaries, the module promotes smoothness where there is no moving instance, while preserving sharp motion boundaries when moving instances are present.

3.3 Video Outpainting with Multi-adapters

To enhance the temporal coherence and spatial consistency at the instance level in video outpainting, we incorporate spatiotemporal instance conditioning and flow conditioning into a video diffusion model via a two-branch architecture, complemented by a multi-scale adapter proposed by [43], which provides different types of guidance across the U-Net blocks of latent diffusion model (LDM).

For the spatial consistency of video outpainting, we first combine the mask logit $\hat{h}_t^n$ of the t -th frame and the memory embedding h_0^n for each pair n at time t . Then, the combined features of all objects are concatenated and projected to obtain the condition for the k -th block and t -th frame in up-blocks:

$$f^{t,n} = \hat{h}_t^{t,n} + h_0^{0,n}, \quad n = 1, 2, \dots, N \quad (15)$$

$$C_{k,t}^{mask} = \text{Mask_Projection}(\text{concat}(f^{t,1}, f^{t,2}, \dots, f^{t,N})) \quad (16)$$

$$x'_{k,t} = x_{k,t} + \gamma_1 \times \text{Cross_Attn}(x_{k,t}, C_{k,t}^{mask}), \quad t = 1, 2, \dots, T \quad (17)$$

where the parameter γ_1 controls the influence of the instance condition on the generated video. And $\text{Cross_Attn}(\cdot)$ denotes the cross-attention mechanism applied at the t -th frame and the k -th block, where $x_{k,t}$ is used to project the query, and $C_{k,t}^{mask}$ is used to project the key and value.

For the temporal coherence of video outpainting, We inject them into the U-Net with the multi-scale optical flow features extracted from O_{out} . We combine the latent features x_k of the k -th block of the U-Net with the condition from flow features C_k through element-wise addition. The combined feature is then processed by a linear layer, and its output is fed directly into the cross attention layer before each motion module in the up-blocks:

$$C_k^{flow} = \text{Flow_Projection}(O_{out}) \quad (18)$$

$$x'_k = x_k + \gamma_2 \times \text{Cross_Attn}(x_k, C_k^{flow}) \quad (19)$$

where the parameter γ_2 controls the influence of the flow condition on the generated video, and x_k is used to project the query, and C_k^{flow} is used to project the key and value.

Finally, after the diffusion process, the denoised latents are obtained by integrating the spatiotemporal conditioning signals through the U-Net. The instance and flow conditioning help guide the denoising process, ensuring both spatial consistency and temporal coherence in the generated video.

3.4 Video-aware Shadow Alignment Discriminator

After obtaining the denoised video latent, the latents are fed to the decoder to transform them from the latent space back to pixel space, producing the outpainted video.

Our outpainting module builds on a pre-trained image LDM [29], whose autoencoder, trained on individual images, causes flickering when applied to temporally coherent sequences, leading to shadow-object inconsistency in videos.

To address this problem, we incorporate dedicated temporal layers into the decoder, proposed in [28], to capture spatio-temporal dependencies across video frames. These layers are subsequently fine-tuned on video data, allowing the decoder to reconstruct temporally coherent textures and structures of instances within the outpainted regions.

we propose a video-aware shadow alignment discriminator that focuses on the dynamic shadow-object pairs in the scene. After obtaining the output video V_{out} of decoder, extract masks of shadow-object pairs M_o^n and M_s^n . For each shadow-object pair, we take the object mask and the shadow mask to extract mask features. For each frame, we also extract image features and then concatenate them

along the channel dimension. The resulting feature dimension is $[T, w, h, c_s + c_o + c_i]$, where T is the number of frames, w and h are the width and height of the image feature, c_s is the number of channels in the shadow mask feature, c_o is the number of channels in the object mask feature, and c_i is the number of channels in the image feature. Then, the result is fed into a local shadow-instance alignment video discriminator to ensure that each pair of extended semantics is visually coherent, which is defined as :

$$\begin{aligned}\min_D V(D) &= \frac{1}{2} \mathbb{E}_{x \sim p_{data}(x)} [(D(x) - b)^2] + \frac{1}{2} \mathbb{E}_{z \sim p_z(z)} [(D(V_{out}, M_o^n, M_s^n))^2], \\ \min_G V(G) &= \frac{1}{2} \mathbb{E}_{z \sim p_z(z)} [(D(V_{out}, M_o^n, M_s^n) - c)^2]\end{aligned}\tag{20}$$

where D denotes the alignment discriminator. a and b denote the ground truth real and fake labels, respectively. c denotes the value that G wants D to believe for the fake data.

Moreover, to ensure photorealistic reconstructions, a patch-wise temporal discriminator [10] built with 3D convolutions is implemented. Additionally, the image-level shadow-instance alignment discriminators proposed by [44] are also adapted to enhance alignment between the extended semantics and their shadows. By integrating these discriminators during finetuning, the decoder learns to generate outpainted content that not only matches the spatial quality of individual frames but also preserves temporal coherence, reducing flickering, shadow misalignment, and even light source inconsistencies in the final video output.

4 Implementation Details

Video Shadow Instance Detector We directly use ViShadow [38] as our video shadow instance detector, which provides robust paired tracking of objects and their corresponding shadows, even in cases where one of them is temporarily invisible or occluded. The initial tracking results provide object-shadow relationships that serve as a heuristic to guide the tracking and prediction mask of instances outside the scene.

Diffusion Module Our method is built upon Stable Diffusion v1-5. The temporal modules are initialized with the weights from pretrained motion modules [9] to obtain additional motion priors. Each spatial 2D convolutional layer is followed by a temporal 1D convolutional layer, which is pretrained in text-to-video tasks on the WebVid dataset [1].

Multi-instance Pair Prediction Since the video shadow instance detector is limited to identifying shadow-object pairs with visible shadows, we employ YOLO [34] and SAM2 [27] to track and segment objects without visible shadows, particularly those near boundary regions. To maintain computational efficiency, we limit the number of instance pairs to a maximum of 10.

Outpainting Strategy For long video outpainting, M3DDM [8] adopts a hybrid coarse-to-fine pipeline, performing coarse outpainting over larger time intervals followed by fine outpainting over shorter intervals. In contrast, MOTIA [36] sequentially splits the long video into short clips with temporal overlap for outpainting. Our approach, however, takes the lifespan of instances into account. Specifically, we dynamically split the long video based on the predicted instance masks to ensure spatial consistency of instances across frames. In scenarios with multiple instances, we prioritize splitting and outpainting at segments with the highest instance overlap.

Training Details Our method is implemented using PyTorch [20] and trained on eight NVIDIA RTX A6000 GPUs. During training, we employ the AdamW optimizer [17] with a fixed learning rate of 1×10^{-4} and the warm-up learning rate step is 1k. Following the training scheme of [8], the model is trained on the WebVid dataset [1] for 6 epochs with a batch size of 16, utilizing gradient accumulation to stabilize optimization under memory constraints. During the inference process, we use the PNDMScheduler [16] to guide the denoising steps in the reverse diffusion process. Also, we use 50 inference steps and a scaled linear β schedule that starts at 0.00085 and ends at 0.012.

5 Experiments

In this section, we describe our experimental setup and present quantitative and qualitative comparisons, ablation studies, and a user study to validate the effectiveness of our method.

5.1 Datasets

We evaluate our proposed approach on two common datasets, DAVIS [21] and YouTube-VOS [39], which are widely used benchmarks for video outpainting.

DAVIS [21] dataset collects 150 videos in total (60 for training, 30 for validation, and 60 for testing), with each video annotated with multiple foreground instances per frame. It was originally introduced as a benchmark for Video Object Segmentation (VOS) to segment and track object instances with precise and temporally consistent annotations across video frames.

YouTube-VOS [39] dataset collects 4,000 high-resolution YouTube videos, totaling over 340 minutes, providing rich diversity in object appearance, motion patterns, and scene complexity. Each video is densely annotated with high-quality pixel-level masks for multiple objects in selected frames.

Additionally, since these two datasets only provide foreground object segmentation annotations and lack instance-level shadow annotations, we introduce a dataset with shadow annotations to further validate the effectiveness of our approach.

SOBA-VID [38] dataset collects 292 videos with a total of 7045 frames, capturing shadow-object interactions across diverse scenarios. It emphasizes varied shadow patterns, dynamic backgrounds, occlusions, and lighting conditions. The dataset is split into 232 training videos with 5863 sparsely annotated frames and 60 testing videos with 1182 densely annotated frames, resulting in 637 shadow-object pairs in total. On average, each video contains 24.1 frames and 2.2 shadow-object pairs.

5.2 Comparisons to Baseline Methods

Qualitative Comparisons. As shown in Fig. 3, we show qualitative results comparing our method to baseline approaches. M3DDM and MOTIA fail to align extended semantics with dynamic shadows. In contrast, our method generates visual content consistent with shadows, improving temporal coherence and shadow alignment. This is due to our shadow-object prediction and flow completion modules, which provide effective spatiotemporal guidance for video outpainting. Additionally, the video-aware shadow alignment discriminator enhances spatiotemporal consistency across frames.

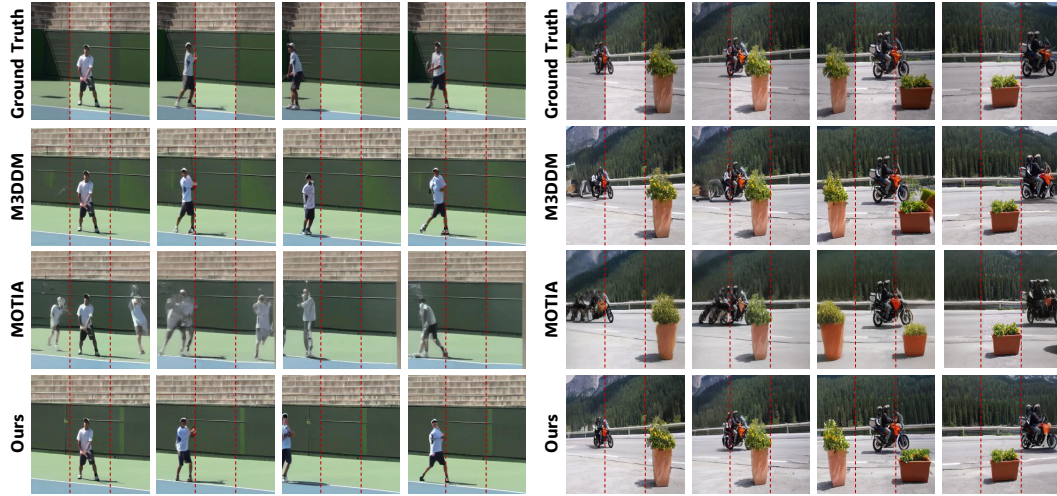


Figure 3: Qualitative comparison of video outpainting results with a mask ratio of 0.666 on the DAVIS [21] and YouTube-VOS [39] datasets.

Quantitative Comparisons. We compare our method with several state-of-the-art approaches on the widely used video outpainting benchmarks DAVIS [21] and YouTube-VOS [39], using four common evaluation metrics: PSNR, SSIM [37], LPIPS [48], and FVD [32]. As shown in Tab. 1, our approach achieves significant improvements over the state-of-the-art in terms of PSNR, SSIM, and FVD. However, the LPIPS score is slightly higher compared to MOTIA. Compared to our efficient zero-shot approach, MOTIA achieves better perceptual alignment and lower LPIPS scores

by fine-tuning the model on each test sample. Note that we follow the test setting of [8, 36] using mask ratios of 0.25 and 0.666 in the horizontal direction.

Table 1: Quantitative Comparisons on DAVIS [21] and YouTube-VOS [39]. $\uparrow$ means “better when higher”, and $\downarrow$ indicates “better when lower”. $\dagger$ denotes it is based on test sample-specific fine-tuning.

Method	DAVIS [21]				YouTube-VOS [39]			
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FVD $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FVD $\downarrow$
Dehan [5]	17.96	0.6272	0.2331	363.1	18.25	0.7195	0.2278	149.7
SDM [28]	20.02	0.7078	0.2165	334.6	19.91	0.7277	0.2001	94.81
M3DDM [8]	20.26	0.7082	0.2026	300.0	20.20	0.7312	0.1854	66.62
MOTIA $\dagger$ [36]	20.36	0.7578	0.1595	286.3	20.25	0.7636	0.1727	58.99
Ours	20.81	<u>0.7254</u>	<u>0.1842</u>	234.7	20.32	0.7719	<u>0.1793</u>	40.78

5.3 Ablation Studies

Ablation Study on Multi-adaption. We further conduct a visual analysis to evaluate the effects of different adaptations on the quality of generated results, as shown in Fig. 4. It can be seen that, although all methods can generate plausible results shortly after the object becomes invisible, as the invisibility duration increases, notable differences emerge. Specifically, without the instance adapter, objects tend to exhibit artifacts or even disappear entirely. The absence of object memory embedding causes significant changes in the appearance of the object, while omitting the flow adapter leads to blurred object motion. In contrast, our method leverages dynamic shadow to generate temporally consistent objects, maintaining both appearance and motion fidelity over longer periods of invisibility.

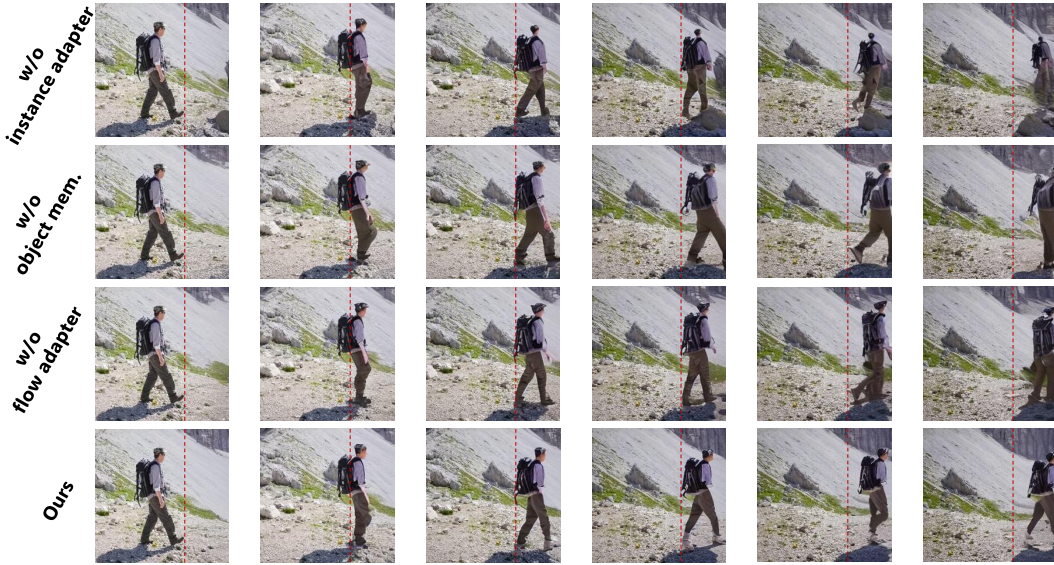


Figure 4: Visual results of multi-adaption ablation study on the DAVIS [21] dataset. The region to the right of the red line is masked out and needs to be generated.

Ablation Study on Outpainting. We conduct an ablation study on the YouTube-VOS [39] dataset to analyse the contribution of each module, shown in Tab. 2. The results demonstrate that both the instance and flow adapters significantly improve the video generation quality, with noticeable enhancements in PSNR, SSIM, and LPIPS. Additionally, the video-aware discriminator significantly helps reduce the FVD, further improving the overall performance of the model.

Ablation Study on Shadow-aware Prediction. We conduct an ablation study to assess the impact of dynamic shadow information on object mask predictions across multiple datasets, as shown in Tab. 3. Our results show that incorporating dynamic shadow improves both object and shadow mask accuracy.

Note that we introduce the SOBA-VID [38] dataset to further validate the impact of shadows, which provides annotations for both objects and shadows. In this dataset, without shadow information incorporated, objects and shadows are treated as independent instances for prediction.

Table 2: Ablation study on the YouTube-VOS [39] dataset.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FVD $\downarrow$
w/ instance adapter	19.89	0.7583	0.1881	64.51
w/ flow adapter	19.94	0.7602	0.1865	61.34
w/ instance & flow adapter	20.15	0.7657	0.1813	51.92
w/ video-aware discriminator	20.11	0.7628	0.1827	42.66
Ours	20.32	0.7719	0.1793	40.78

5.4 User study

We conducted a user study comparing our method against M3DDM [8] and MOTIA [36] using the YouTube-VOS [39] dataset with a horizontal mask ratio of 0.66 as source videos. We collected preferences from 40 volunteers, each evaluating 20 randomly selected result sets. The evaluation focused on two aspects: visual quality (e.g., clarity, color fidelity, and texture detail) and temporal consistency (e.g., motion smoothness, object continuity, and temporal coherence). As shown in Tab. 4, our method is preferred over both baselines in terms of both visual quality and temporal consistency.

Table 3: Ablation study of mask IoU with and without shadow incorporated across datasets.

Method	DAVIS	YouTube-VOS	SOBA-VID
	Object	Object	Object / Shadow
w/o shadow	0.75	0.78	0.65 / 0.56
w shadow	0.82	0.84	0.76 / 0.65

Table 4: User preference distribution (%) across three methods.

Method	Visual Quality	Temporal Consistency
M3DDM [8]	18.0%	22.3%
MOTIA [36]	25.6%	15.7%
Ours	56.4%	62.0%

6 Conclusion

In this paper, we propose a novel instance-aware video outpainting framework that leverages shadow-object associations to enhance the spatial and temporal consistency of generated content. By explicitly tracking and modeling shadow-object pairs, our method effectively unveils both the spatial region and temporal motion of instances that are invisible in the input video. The integration of spatiotemporal instance guidance with optical flow completion, along with a video-aware discriminator, ensures temporal consistency and coherent scene extension even in complex dynamic scenarios. Extensive experiments on several benchmarks demonstrate that our approach significantly outperforms existing state-of-the-art methods.

Limitations. Our method exploits shadow cues to improve instance generation beyond the visible scene. The informativeness of shadows depends on the angle of the light source. For example, lower angles usually cast longer shadows, which carry more semantic hints of invisible instances, such as their motion, actions, or even the location of the light source. In scenes without visible shadows, these cues are absent, and our approach falls back to standard instance-aware video outpainting.

Societal Impacts. Our approach can extrapolate scenes beyond the original camera view, supporting applications in film production, surveillance, and autonomous driving. However, the generated regions are synthetic and may not reflect reality, raising concerns about authenticity and trust. Following [7], we apply watermarking to label generated content and reduce the risk of misuse.

Acknowledgements

This work is supported by National Natural Science Foundation of China (Grant No.62302287) and projects of the Shanghai Committee of Science and Technology, China (Grant No.23ZR1423500).

References

- [1] Max Bain et al. “Frozen in time: A joint video and image encoder for end-to-end retrieval”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 1728–1738.
- [2] Zhihao Chen et al. “Triple-cooperative video shadow detection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 2715–2724.
- [3] Ho Kei Cheng et al. “Putting the object back into video object segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 3151–3161.
- [4] Yen-Chi Cheng et al. “Inout: Diverse image outpainting via gan inversion”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 11431–11440.
- [5] Loi  Dehan et al. “Complete and temporally consistent video outpainting”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 687–695.
- [6] Xinpeng Ding et al. “Learning shadow correspondence for video shadow detection”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 705–722.
- [7] Patrick Esser et al. “Structure and Content-Guided Video Synthesis with Diffusion Models”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 7346–7356.
- [8] Fanda Fan et al. “Hierarchical masked 3d diffusion model for video outpainting”. In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, pp. 7890–7900.
- [9] Yuwei Guo et al. “Animatediff: Animate your personalized text-to-image diffusion models without specific tuning”. In: *arXiv preprint arXiv:2307.04725* (2023).
- [10] Phillip Isola et al. “Image-to-image translation with conditional adversarial networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134.
- [11] Hao Jiang and Mark S Drew. “Shadow-resistant tracking in video”. In: *2003 International Conference on Multimedia and Expo. ICME’03. Proceedings (Cat. No. 03TH8698)*. Vol. 3. IEEE. 2003, pp. III–77.
- [12] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. “Occlusion-aware video object inpainting”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 14468–14478.
- [13] Zhen Li et al. “Towards an end-to-end framework for flow-guided video inpainting”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 17562–17571.
- [14] Han Lin, Maurice Pagnucco, and Yang Song. “Edge guided progressively generative image outpainting”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 806–815.
- [15] Lihao Liu et al. “Scotch and soda: A transformer video shadow detection framework”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 10449–10458.
- [16] Luping Liu et al. “Pseudo Numerical Methods for Diffusion Models on Manifolds”. In: *International Conference on Learning Representations*. 2022. URL: <https://openreview.net/forum?id=PlKWVd2yBkY>.
- [17] Ilya Loshchilov and Frank Hutter. *Decoupled Weight Decay Regularization*. 2019. arXiv: 1711.05101 [cs.LG]. URL: <https://arxiv.org/abs/1711.05101>.
- [18] Sohail Nadimi and Bir Bhanu. “Physical models for moving shadow and object detection in video”. In: *IEEE transactions on pattern analysis and machine intelligence* 26.8 (2004), pp. 1079–1087.
- [19] Mahdi Naseri, Jiayan Qiu, and Zhou Wang. “Generated Contents Enrichment”. In: *arXiv preprint arXiv:2405.03650* (2024).
- [20] Adam Paszke et al. “Pytorch: An imperative style, high-performance deep learning library”. In: *Advances in neural information processing systems* 32 (2019).
- [21] Federico Perazzi et al. “A benchmark dataset and evaluation methodology for video object segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 724–732.

- [22] Jiayan Qiu et al. “Hallucinating visual instances in total absentia”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 264–282.
- [23] Jiayan Qiu et al. “Matching seqlets: An unsupervised approach for locality preserving sequence matching”. In: *IEEE transactions on pattern analysis and machine intelligence* 43.2 (2019), pp. 745–752.
- [24] Jiayan Qiu et al. “Scene essence”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 8322–8333.
- [25] Jiayan Qiu et al. “World from blur”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 8493–8504.
- [26] Zhaofan Qiu, Ting Yao, and Tao Mei. “Learning spatio-temporal representation with pseudo-3d residual networks”. In: *proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 5533–5541.
- [27] Nikhila Ravi et al. “SAM 2: Segment Anything in Images and Videos”. In: *arXiv preprint arXiv:2408.00714* (2024). URL: <https://arxiv.org/abs/2408.00714>.
- [28] Robin Rombach et al. “High-Resolution Image Synthesis With Latent Diffusion Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 10684–10695.
- [29] Robin Rombach et al. “High-resolution image synthesis with latent diffusion models”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.
- [30] Sheyang Tang et al. “Controllable Data Generation with Hierarchical Neural Representations”. In: *Forty-second International Conference on Machine Learning*.
- [31] Zachary Teed and Jia Deng. “Raft: Recurrent all-pairs field transforms for optical flow”. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer. 2020, pp. 402–419.
- [32] Thomas Unterthiner et al. “Towards accurate generative models of video: A new metric & challenges”. In: *arXiv preprint arXiv:1812.01717* (2018).
- [33] A Vaswani. “Attention is all you need”. In: *Advances in Neural Information Processing Systems* (2017).
- [34] Ao Wang et al. “Yolov10: Real-time end-to-end object detection”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 107984–108011.
- [35] Yang Wang. “Real-time moving vehicle detection with cast shadow removal in video based on conditional random field”. In: *IEEE transactions on circuits and systems for video technology* 19.3 (2009), pp. 437–441.
- [36] Fu-Yun Wang et al. “Be-your-outpainter: Mastering video outpainting through input-specific adaptation”. In: *European Conference on Computer Vision*. Springer. 2025, pp. 153–168.
- [37] Zhou Wang et al. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.
- [38] Zhenghao Xing et al. “Video Instance Shadow Detection Under the Sun and Sky”. In: *IEEE Transactions on Image Processing* (2024).
- [39] Ning Xu et al. “Youtube-vos: A large-scale video object segmentation benchmark”. In: *arXiv preprint arXiv:1809.03327* (2018).
- [40] Rui Xu et al. “Deep flow-guided video inpainting”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 3723–3732.
- [41] Xiaoyu Xu et al. “Relationship spatialization for depth estimation”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 615–637.
- [42] Xiaoyu Xu et al. “Visual Relationship Transformation”. In: *European Conference on Computer Vision*. Springer. 2024, pp. 251–272.
- [43] Hu Ye et al. “Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models”. In: *arXiv preprint arXiv:2308.06721* (2023).
- [44] Hang Yu et al. “Shadow-Enlightened Image Outpainting”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 7850–7860.
- [45] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. “Learning joint spatial-temporal transformations for video inpainting”. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*. Springer. 2020, pp. 528–543.

- [46] Kaidong Zhang, Jingjing Fu, and Dong Liu. “Flow-guided transformer for video inpainting”. In: *European conference on computer vision*. Springer. 2022, pp. 74–90.
- [47] 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*. 2022, pp. 5982–5991.
- [48] Richard Zhang et al. “The unreasonable effectiveness of deep features as a perceptual metric”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 586–595.
- [49] Shengyu Zhao et al. “Large Scale Image Completion via Co-Modulated Generative Adversarial Networks”. In: *International Conference on Learning Representations (ICLR)*. 2021.
- [50] Haipeng Zhou et al. “Timeline and boundary guided diffusion network for video shadow detection”. In: *Proceedings of the 32nd ACM International Conference on Multimedia*. 2024, pp. 166–175.
- [51] Qianyu Zhou et al. “TransVOD: End-to-end video object detection with spatial-temporal transformers”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.6 (2022), pp. 7853–7869.
- [52] Shangchen Zhou et al. “Propainter: Improving propagation and transformer for video inpainting”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 10477–10486.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The contributions and scope of the paper have been truthfully described throughout the abstract and the introduction, and concisely summarized at the end of the introduction. Each of the listed contributions is backed up by experimental results or literature review throughout the paper.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: The limitations of the work are discussed in Section 6.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include any theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We included detailed descriptions of our model architecture in Section 3, references to publicly available pre-trained weights, data processing procedures, and other training details in Section 4 and further in the supplementary material. We will also open-source our code and model checkpoints upon paper acceptance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: As stated in the response above, we provided detailed instructions on how to replicate our experiment results in the main paper and further in the supplementary material. We will release our code and models upon paper acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provided detailed instructions on replicating the training and evaluation procedures in Section 4, Section 5, and further in the supplementary material. We did not perform delicate tuning for the hyperparameters.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We did not compute error bars due to the lack of adequate computational resources. However, we tried our best to provide faithful and reliable quantitative evaluation results, including a user study. All experiments are conducted with the same and commonly used random seed.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We described our compute setup as well as the training time and inference runtime in Section 4 and further in the supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Authors carefully read the NeurIPS Code of Ethics and preserved anonymity.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The societal impacts of the work are discussed in Section 6.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [\[Yes\]](#)

Justification: We mentioned in Section 6 that the generated content from our model can be watermarked to prevent potential misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [\[Yes\]](#)

Justification: We cite the original paper that produced the code package or dataset. All the pre-trained models and tools we used are specified in Section 4 and further in the supplementary material.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: This research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.