From Draft to Draft-Free: One-Step Video Object Removal via Privileged Distillation and Fast Planting

Zizhao Chen[1], Ping Wei[1†], Guang Dai[2], Jingdong Wang[4], and Mengmeng Wang[3,2†]

[1] State Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
[2] SGIT AI Lab, State Grid Corporation of China
[3] Zhejiang University of Technology
[4] Baidu

https://github.com/bigD233/D2DF

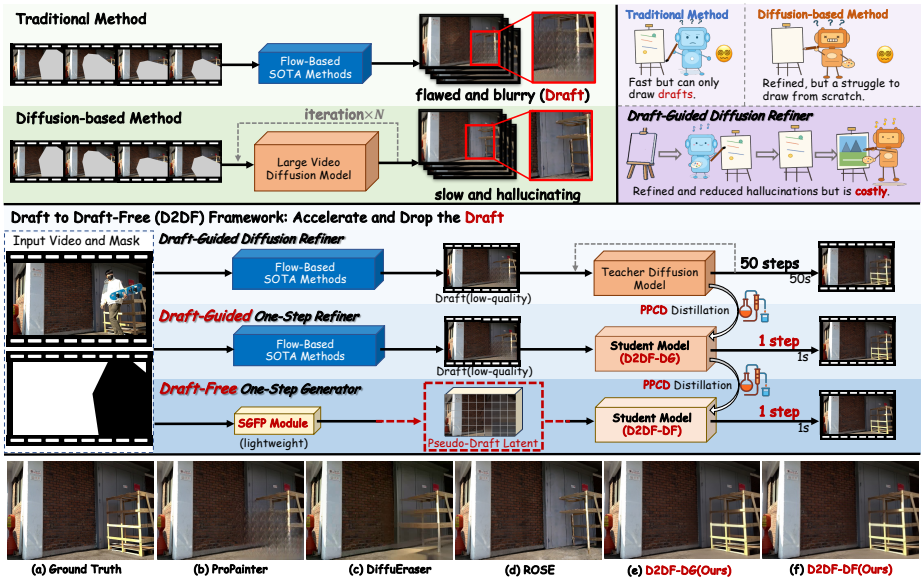


Fig. 1: Overview of our three-staged **D2DF** framework. To address the limitations of traditional and diffusion-based methods, we combine them as a refiner. However, this incurs significant time costs and depends on external drafts. Through PPCD distillation and the SGFP module, we obtain the one-step video object removal models D2DF-DG and D2DF-DF. Visual comparisons from (a) to (f) demonstrate that D2DF exhibits powerful background reconstruction regardless of whether drafts are used as prior information.

Abstract. Video object removal is a fundamental yet challenging task in video editing. Despite recent progress, existing methods typically fall into two categories. Traditional approaches based on optical flow or attention mechanisms often introduce noticeable artifacts and yield unnatural results. In contrast, diffusion-based methods improve visual realism but demand multiple denoising steps, lim-

† Corresponding authors. Contact: pingwei@xjtu.edu.cn, mengmewang@gmail.com.

iting their practicality. To address these issues, we propose From-Draft-to-Draft-Free (**D2DF**), a framework that distills the ability of transforming coarse drafts into refined videos into a one-step video generation model. Within D2DF, a teacher model is trained to refine low-quality removal results ("drafts") into high-fidelity videos by multiple steps. Then, through Prior-Privileged Consistency Distillation (**PPCD**), we distill this capability into a student model that performs one-step removal conditioned on the draft. To eliminate draft dependency, we introduce a Self-Guided Fast Planting (**SGFP**) module based on our Temporal Masked Transformer that autonomously generates scene-consistent pseudo-drafts in latent space, enabling a fully **draft-free** one-step model. Extensive experiments show that both draft-conditioned and draft-free versions achieve state-of-the-art performance on multiple metrics, surpassing traditional and multi-step generative methods in both quality and efficiency. The denoising process for a single video takes only about 1 second.

Keywords: Object Removal · Video Inpainting · One-Step Video Generation

1 Introduction

Video object removal (VOR) aims to remove undesired objects from videos and reconstruct the background in the occluded regions. As an essential task in video inpainting [8,46,63], it has a wide range of applications. The challenge is intrinsically difficult, often requiring globally consistent spatial–temporal content generation [1,7,41] for regions that cannot be propagated from adjacent frames.

Traditional VOR methods rely on optical-flow propagation or shallow transformer architectures [9, 21, 23, 59, 60, 65]. While they propagate background textures, their limited generative capacity yields structural distortions and blurry reconstructions, especially for large occlusions (Fig. 1(b)). This "draft" output contains noticeable flaws.

The emergence of large video diffusion models [12,20,29,30,33,56] offers a path to solving this "generative capacity" bottleneck, significantly enhancing realism. However, two critical problems remain. First, **Hallucination**: weak priors cause diffusion models to "hallucinate"—generating plausible but semantically incorrect content (Fig. 1(c)(d)), failing deterministic reconstruction. Second, **Slowness**: diffusion inference is intrinsically slow (tens of DDIM steps [37]), hindering practical deployment.

To solve the hallucination issue, we observe that flow-based methods efficiently propagate reliable pixel information across frames. This process is fast and fills parts of the masked region with valid content. While their "draft" output can be blurry (Fig. 1(b)), it provides a strong, partially reconstructed starting point. We use this "draft" to guide a diffusion "refiner" to avoid hallucination. To solve the slowness issue, we argue that the full multi-step refinement is unnecessary given this strong "draft" condition. The task is simplified enough that the "refiner" can be distilled into a single, efficient step. Our ultimate goal, however, is an autonomous model that learns this paradigm but discards the external "draft" prior at inference.

Based on this insight, we propose the "From Draft to Draft-Free" framework **(D2DF)**, which progressively derives a draft-guided one-step refiner (D2DF-DG) and a fully draft-free one-step generator (D2DF-DF). Our framework (Fig. 1) has three stages:

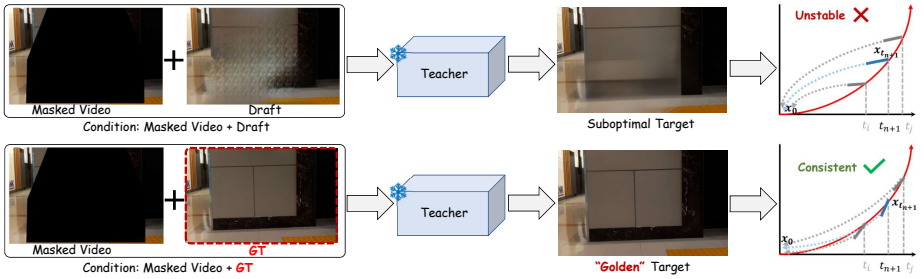


Fig. 2: Comparison of D-LDM outputs under different conditions. The middle column shows the results obtained by 4-step DDIM denoising of the teacher network. The solid red line represents the ground-truth denoising trajectory. The dashed line represents the solution trajectory of the teacher. The short solid line indicates the target provided by the teacher at the current timestep.

Stage I: The Draft-Guided Diffusion Teacher. We first train a multi-step diffusion teacher (D-LDM) conditioned on external "drafts" (e.g., ProPainter [65]). The teacher refines the flawed draft into a high-quality object removal video through 50 steps.

Stage II: The Draft-Guided One-Step Refiner (D2DF-DG). Our goal is to distill the 50-step teacher into a 1-step student (D2DF-DG). A naive consistency distillation [26, 38, 39] would fail for the reason that when conditioned on the flawed draft, the teacher must reconstruct from varying levels of damage. This forces it onto an unstable solution trajectory that significantly deviates from the ideal path to the ground truth (GT). As Fig.2 illustrates, this noisy path provides an inconsistent and Suboptimal Target for the student, making convergence difficult. To solve this, we introduce **Prior-Privileged Consistency Distillation (PPCD)**. We replace the flawed draft with the GT as the condition of the teacher. This "truth-injected" strategy compels the teacher to generate a perfectly consistent and stable trajectory. This ideal path, which stays aligned with the GT, provides a "Golden" Target for the student, dramatically improving distillation effectiveness. Crucially, the student (D2DF-DG) still accepts the draft as input, but now learns a clean, one-step mapping to this "golden" output.

Stage III: The Draft-Free One-Step Generator (D2DF-DF). Despite the impressive performance of D2DF-DG, it still relies on external drafts. To eliminate this dependency, we design the **Self-Guided Fast Planting (SGFP)** module, a lightweight network based on Temporal Masked Transformer (TMT) that constructs a "pseudo-draft" in the latent space. This differs fundamentally from drafts in explicit pixel-space from typical methods like ProPainter [65]. We then use PPCD again to distill D2DF-DG (as teacher) into our final student, D2DF-DF. This model inherits the one-step speed but is now fully autonomous, achieving end-to-end generation without external priors.

Our D2DF-DF achieves superior performance on three benchmarks [34, 51, 57], exhibiting strong robustness, especially under large-mask scenarios. Crucially, our models set a new standard for efficiency: D2DF-DF achieves state-of-the-art results while being over 40 times faster than competing SOTA diffusion methods like ROSE [29]. This combination of quality, speed, and autonomy represents a significant step forward for practical video editing.

Our contributions can be summarized as follows:

• We propose a novel From-Draft-to-Draft-Free (D2DF) framework, a progressive methodology to decouple latency and dependency, producing two models: a high-efficiency one-step refiner (D2DF-DG) and a fully autonomous one-step generator (D2DF-DF).
• In D2DF, we introduce Prior-Privileged Consistency Distillation (PPCD), a novel "truth-injected" distillation where a teacher uses GT to create a golden target.
• We design the Self-Guided Fast Planting (SGFP) module, a lightweight latent-space prior generator, combined with PPCD, to create a fully autonomous, end-to-end, one-step model.

2 Related Work

2.1 Diffusion-based Video Generation

Diffusion models have recently achieved remarkable success in visual generation by iteratively denoising noisy inputs [4, 6, 15, 32, 45, 49, 53]. Following advances in text-to-image diffusion, recent works extend this paradigm to video generation. Text-to-video (T2V) models [3, 13, 44, 48, 50, 56] integrate temporal modules into 2D UNet architectures, while image-to-video (I2V) methods such as DynamiCrafter [52] and PixelDance [58] demonstrate strong motion and scene modeling ability. More recently, Transformer-based diffusion models (e.g., DiT, SORA-like [5,30]) further improve temporal reasoning and visual fidelity.

2.2 Video Inpainting

Most existing video inpainting approaches adopt a two-stage framework. In the first stage, prior information is extracted from the input video, such as motion cues [16, 41, 55], homography transformations [19] or low-resolution reconstructions [42]. In the second stage, based on these priors, the missing regions are filled by a carefully designed generative network. Typical architectures include image inpainting models [9, 55, 61], 3D convolutional networks [16], and Transformer-based frameworks [60]. Recent studies have combined the advantages with the spatio-temporal modeling power of video Transformers [9, 21, 60, 65], leading to a series of two-stage hybrid methods. Recently, researchers have introduced video diffusion models for video inpainting [2, 7, 11, 18, 20, 29]. These models substantially improve the realism and coherence of the generated videos, yet they still suffer from hallucination in the reconstructed regions. To address these issues, our work introduces the SGFP module to provide strong priors, together with a Prior-Privileged Consistency Distillation (PPCD) framework that enables one-step video generation, effectively balancing quality and efficiency.

2.3 Consistency Model

Diffusion models achieve impressive generative performance but suffer from slow inference due to their multi-step sampling process. To accelerate generation, several strategies have been proposed, such as improved ODE solvers [25], neural operators [64], and model distillation [28, 35, 36]. Among them, consistency distillation [39] has emerged as an effective solution. Instead of relying on iterative denoising, it trains a student consistency model to directly predict clean samples from intermediate noisy states, enabling one-step generation while preserving sample quality. Subsequent works, such

as Latent Consistency Models (LCM) [26, 27], extend this paradigm into latent space for higher efficiency and lower memory cost. Recent improvements [10, 24, 38] further enhance stability and fidelity, making consistency distillation a promising framework for efficient generative modeling. However, consistency models still face challenges in adapting to task-specific applications, especially for video generation. To address this, we propose Prior-Privileged Consistency Distillation (PPCD), which provides more stable and consistent solution trajectories for video generation tasks.

3 Method

3.1 Preliminaries

Latent Diffusion Models (LDM) [33] establish the generative modeling process in the latent space. The training objective matches the predicted denoised sample $\mathbf{x}_0$ with the ground truth through a simplified variational bound:

$$\mathcal{L}_{LDM} = \mathbb{E}_{x,t,\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|_2^2\right], \tag{1}$$

where $\mathbf{x}_t$ is the noisy input obtained through the forward diffusion process with timestep t, and $\boldsymbol{\epsilon}_\theta$ is a neural network that predicts the noise component. This objective effectively establishes a progressive denoising mapping that transforms pure noise $\mathbf{x}_T$ into clean data $\mathbf{x}_0$ through a Markov chain of T steps. This mapping can be formulated as a parameterized reverse process $\mathbf{p}_\theta : (\mathbf{x}_t, t) \longmapsto \mathbf{x}_{t-1}$. However, this requires an iterative denoising process.

Consistency Model [39] attempts to establish mappings through a consistency function, defined as $\mathbf{f} : (\mathbf{x}_t, t) \longmapsto \mathbf{x}_\delta$, where δ is a fixed small positive number. Its consistency is reflected in one of the key properties of this function:

$$\mathbf{f}(\mathbf{x}_t, t) = \mathbf{f}(\mathbf{x}_{t'}, t'), \quad \forall t, t' \in [\epsilon, T]. \tag{2}$$

Ideally, the consistency function established through learnable neural networks f_θ enables the consistency model to recover the original data $\mathbf{x}_0$ from the initial distribution $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, T^2\boldsymbol{I})$ in one step of denoising. Learnable parameters θ are trained via Consistency Distillation [10, 38, 39]. The distillation goal is to minimize the output discrepancy between random adjacent data points, thereby ensuring self-consistency in the sense of probability. Formally, the consistency loss is defined as follows:

$$\mathcal{L}(\theta, \theta^-; \Phi) = \mathbb{E}_{\boldsymbol{x},t}\left[d\left(\mathbf{f}_\theta(\mathbf{x}_{t_{n+1}}, t_{n+1}), \mathbf{f}_{\theta^-}(\hat{\mathbf{x}}^\phi_{t_n}, t_n)\right)\right], \tag{3}$$

where $0 = t_1 < t_2 < \cdots < t_N = T$, and n is uniformly distributed over the set $\{1, 2, \cdots, N-1\}$. $\Phi(\cdot)$ denotes the one-step ODE [25] solver applied to PF-ODE [40], the model parameter θ^- is obtained from the exponential moving average (EMA) of θ, and $d(\cdot, \cdot)$ is a chosen metric function for measuring the distance between two samples. $\hat{x}^\phi_{t_n}$ is determined using the equation $\hat{\mathbf{x}}^\phi_{t_n} := \Phi(\mathbf{x}_{t_{n+1}}, t_{n+1}, t_n; \phi)$. $\hat{\mathbf{x}}^\phi_{t_n}$ is estimated by a teacher model with an ODE solver in consistency distillation.

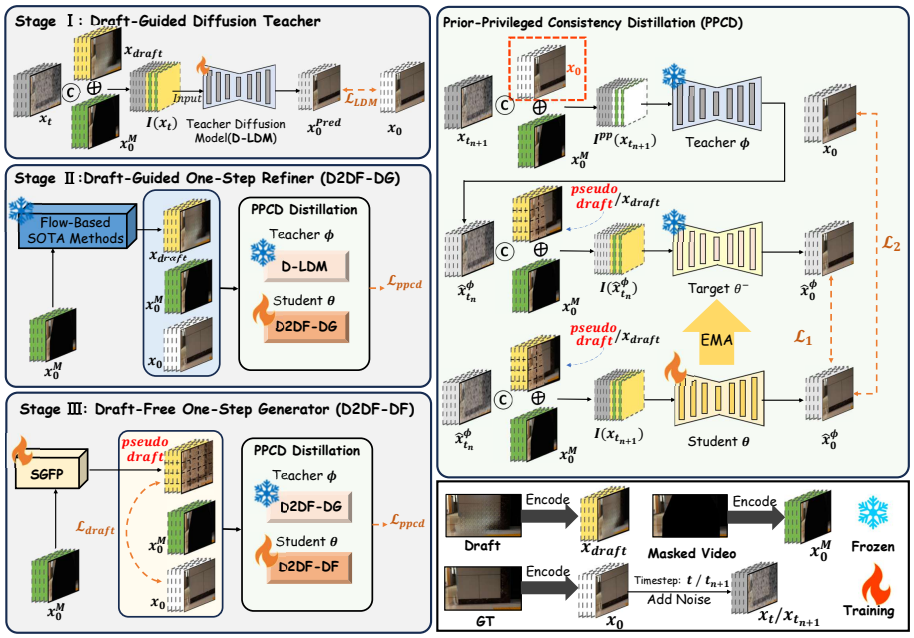


Fig. 3: Pipeline of D2DF. On the left are the three stages of our framework, while on the right is the PPCD distillation we propose.

3.2 Stage I: Draft-Guided Diffusion Teacher

Our first stage simplifies the task into a refinement process. We introduce a strong conditional prior (**draft**) from existing methods (e.g., ProPainter [65]). This process establishes a sufficiently powerful teacher network to provide a mapping from draft to refined video in multiple steps. Thus, we train a Draft-Guided Diffusion Teacher model (D-LDM), denoted as ϕ. This model is conditioned on both the draft latent $\mathbf{x}_{\text{draft}}$ and the masked video latent $\mathbf{x}_0^M$. At timestep t, the model input $I(\mathbf{x}_t)$ is defined as the concatenation of the noised input $\mathbf{x}_t$ and the conditional information:

$$I(\mathbf{x}_t) = Cat(\mathbf{x}_t, \mathbf{x}_0^M + \mathbf{x}_{\text{draft}}), \tag{4}$$

where $\mathbf{x}_t$ denotes the noised input through the forward diffusion process and $Cat(\cdot)$ denotes the latents concatenation.

The D-LDM model is then trained by substituting this conditional input $I(\mathbf{x}_t)$ for $\mathbf{x}_t$ in the LDM training objective (Eq.1). The objective for the teacher network $\boldsymbol{\epsilon}_\phi$ becomes:

$$\mathcal{L}_{D-LDM} = \mathbb{E}_{x,t,\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\phi(I(\mathbf{x}_t), t)\|_2^2\right]. \tag{5}$$

This strong draft conditioning significantly reduces the mapping complexity, making a one-step approximation feasible. The resulting D-LDM ϕ achieves strong generative performance via iterative (e.g., 50-step) DDIM sampling.

3.3 Stage II: Prior-Privileged Consistency Distillation for Draft-Guided One-Step Refiner

In the second stage, we distill the D-LDM teacher model (ϕ) into a one-step, draft-guided student model, **D2DF-DG**. While the theory of consistency distillation (CD) is promising, it suffers from high sampling variance from the teacher model, especially when conditioned on imperfect drafts, which destabilizes the distillation process.

To overcome this, we introduce Prior-Privileged Consistency Distillation (PPCD). Standard distillation from a teacher, conditioned on flawed drafts, leads to unstable and suboptimal trajectories. Our key idea is to inject the ground-truth video as a "privileged" prior into the teacher model ϕ during distillation only. This provides an ideal, "golden" target trajectory and ensures the mapping is anchored in a convergent solve space.

Formally, the teacher model serves as a trajectory provider. To ensure it provides a stable and reliable consistency trajectory, we replace the draft latent $\mathbf{x}_{\text{draft}}$ with the ground-truth latent $\mathbf{x}_0$ as shown in Fig.3(PPCD). The teacher's privileged input at timestep t becomes:

$$\mathbf{I}^{pp}(\mathbf{x}_t) = Cat(\mathbf{x}_t, \mathbf{x}_0^M + \mathbf{x}_0). \tag{6}$$

Given this privileged input, the teacher network produces a denoised latent $\hat{\mathbf{x}}_{t_n}^{\phi} = \Phi(I^{pp}(\mathbf{x}_{t_{n+1}}), t_{n+1}, t_n; \phi)$ at the next timestep along the diffusion trajectory.

Based on the principle of consistency distillation (Eq.3), the student network $\mathbf{f}_\theta$ (and its EMA target $\mathbf{f}_{\theta^-}$) is trained to match the consistency property. However, the student still operates using the **draft** as input, as it will at inference time. We obtain the target latent and prediction latent:

$$\hat{\mathbf{x}}_0^{\phi} = \mathbf{f}_{\theta^-}(Cat(\hat{\mathbf{x}}_{t_n}^{\phi}, \mathbf{x}_0^M + \mathbf{x}_{\text{draft}}), t_n), \tag{7}$$

$$\mathbf{x}_0^{pred} = \mathbf{f}_{\theta}(Cat(\hat{\mathbf{x}}_{t_{n+1}}^{\phi}, \mathbf{x}_0^M + \mathbf{x}_{\text{draft}}), t_{n+1}), \tag{8}$$

where $\mathbf{f}_{\theta^-}$ denotes the target model and θ^- is an EMA version of the student network θ. The meanings of t_n and t_{n+1} are the same as in Eq.3. Ultimately, our PPCD utilizes the following objectives to optimize the student network θ:

$$\mathcal{L}_{\text{ppcd}} = \mathbb{E}_{\boldsymbol{x},t}\left[d\left(\mathbf{x}_0^{pred}, \hat{\mathbf{x}}_0^{\phi}\right)\right] + \alpha\mathbb{E}_{\boldsymbol{x},t}\left[d\left(\mathbf{x}_0^{pred}, \mathbf{x}_0\right)\right], \tag{9}$$

where $d(\cdot,\cdot)$ is the mean square error of two samples and α is a hyperparameter. The first term enforces consistency with the privileged trajectory, while the second term adds a direct ground-truth constraint.

Using the trained D-LDM as the teacher, we apply PPCD to obtain **D2DF-DG**, a one-step, draft-conditioned object removal refiner model.

3.4 Stage III: Self-Guided Fast Planting (SGFP) for Draft-Free One-Step Generator

The D2DF-DG model from Stage II is fast but still depends on drafts generated by external methods, incurring inference overhead. In our final stage, we eliminate this dependency by proposing the **Self-Guided Fast Planting (SGFP)** module, resulting in the final draft-free model, **D2DF-DF**.

The SGFP module (Fig.4) learns to reconstruct a “pseudo-draft" directly in latent space using spatio-temporal information from unoccluded regions. The module is built upon our Temporal Masked Transformer (TMT) architecture, which aims to rapidly reconstruct masked regions. Given a masked draft latent $\mathbf{x}_0^M \in \mathbb{R}^{C\times T\times H\times W}$ and a binary mask $\mathbf{M} \in \{0,1\}^{1\times T\times H\times W}$ (where 0 denotes masked regions and 1 denotes known regions), the objective of SGFP is to get a completed latent tensor $\mathbf{x}_{recon}$.

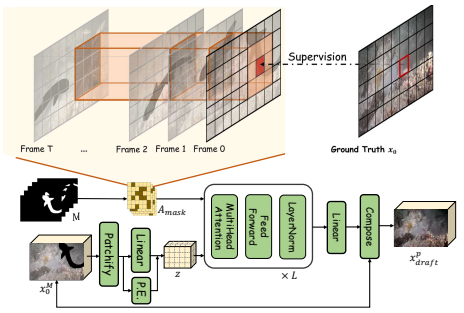


Fig. 4: Pipeline of SGFP. The entire reconstruction operation is in the latent space, constrained by the attention mask based on valid patches and spatial windows.

To handle the high-dimensional video latents while maintaining computational tractability, we first partition the spatial dimensions (H, W) into a series of non-overlapping *patches* of size $P \times P$. This reshapes the input into a sequence of $N = T \times (H/P) \times (W/P)$ tokens, where each token is a vector of dimension $C \cdot P^2$. Each patch vector $\mathbf{p}_i$ is then embedded into a lower-dimensional space $\mathbb{R}^{D_e}$ using a linear layer, combined with a positional embedding: $\mathbf{z}_i = \text{Linear}(\mathbf{p}_i) + \mathbf{e}_{pos}(i)$, where D_e is the embedding dimension and $\mathbf{z}_i$ is the input to the TMT layers. The term $\mathbf{e}_{pos}(i)$ is a learnable positional encoding generated by a small MLP from the patch's normalized 3D coordinates (t, h, w), providing spatiotemporal awareness.

The key to the “self-guided” nature of SGFP lies in the construction of its self-attention mechanism. The computation must rely solely on known, unmasked information. We first compute a mask ratio $r_i = \text{Mean}(\mathbf{M}_{patch,i})$ for each patch. If r_i exceeds a threshold τ, the patch is considered “invalid”. In the self-attention calculation, all patches $\mathbf{z}_i$ can act as Queries, but only “valid” patches (i.e., $r_j \leq \tau$) are permitted to serve as Keys and Values. This is enforced via an attention mask $\mathbf{A}_{mask}$:

$$\mathbf{A}_{mask}(i,j) = \begin{cases} 0 & \text{if } r_j \leq \tau \text{ and } \text{dist}(i,j) \leq R \\ -\infty & \text{otherwise} \end{cases},$$

where $\text{dist}(i,j) \leq R$ is an optional spatial window constraint that further restricts attention to a $(2R+1) \times (2R+1)$ spatial neighborhood of the query patch i (across all time frames T), reducing complexity and reinforcing local priors. After L Transformer layers, the output sequence $\mathbf{z}_i^{(L)}$ is projected back to the original patch dimension $\mathbb{R}^{C\cdot P^2}$ and “un-folded” to reconstruct the latent variable $\mathbf{x}_{recon} \in \mathbb{R}^{C\times T\times H\times W}$. Finally, the pseudo-draft $\mathbf{x}^p_{\text{draft}}$ is generated by composing the reconstruction and the original input using the mask $\mathbf{M}$, ensuring the fidelity of unmasked regions:

$$\mathbf{x}^p_{\text{draft}} = \mathbf{x}_{recon} \odot (1 - \mathbf{M}) + \mathbf{x}_0^M \odot \mathbf{M}. \tag{10}$$

We then treat the D2DF-DG model (from Stage II) as the new teacher and employ the PPCD framework again to distill the final draft-free student, **D2DF-DF**, which now

incorporates the SGFP module. We trained the SGFP together with the subsequent generation module, incorporating constraints for pseudo-drafts:

$$\mathcal{L}_{\text{draft}} = \mathbb{E}_{\boldsymbol{x},t}\left[d\left(\mathbf{x}^{p}_{draft}, \mathbf{x}_0\right)\right]. \tag{11}$$

By introducing SGFP, we obtain D2DF-DF, an end-to-end, draft-free video object removal model. The entire progressive pipeline is summarized in Figure 3.

4 Experiments

4.1 Experiment Setup

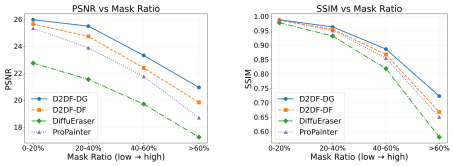
Implementation Details. We employ CogVideoX-5B-I2V [56] (480×720 resolution) as the pre-trained weights for training the first-stage D-LDM network. The D-LDM model is trained with a learning rate of 5×10^{-5} and uses 50 DDIM steps for inference. The "draft" is obtained using the ProPainter [65]. In the PPCD distillation, the hyperparameter α is set to 1, with a batch size of 4 and a learning rate of 2×10^{-5}. Sampling timesteps $t_1, t_2, ..., t_N$ are performed uniformly at 50-step intervals over 1k timesteps. The EMA decay rate of the target network is set to 0.99. For the SGFP module, the number of layers L was set to 4, the spatial radius R to 2, the patch size P to 10, the embedding dimension D_e to 512, and the mask threshold τ to 0.7. In the third-stage training, the ratio of $\mathcal{L}_{\text{draft}}$ to $\mathcal{L}_{\text{ppcd}}$ is 1:5, and the learning rate for the SGFP module parameters is 1×10^{-4}. We train only the DiT component with full-parameter SFT.

Datasets. DAVIS [31] and YouTube-VOS [54] are widely used for video inpainting with stationary random masks, but they do not provide ground-truth videos after object removal. Therefore, they are less suitable for VOR evaluation unless synthetic or subjective protocols are introduced. Based on this, we evaluate our model on three datasets: RORD [34], ROVI [51], and VPLM [57]. RORD is a large-scale real-world dataset specifically designed for object removal, containing 2,761 distinct scenes for training and 343 scenes for testing. ROVI includes 5,650 videos with a total of 9,091 object removal instances, among which 458 videos (758 objects) are reserved for testing. From the training splits of RORD and ROVI, we extract 25-frame video segments to form our training set, resulting in approximately 30k samples. For evaluation, we adopt the VPLM dataset, which provides 60 video-based object removal tasks. All three datasets provide triplets of original video, object mask, and ground-truth removed video.

Metrics. To quantitatively evaluate the performance of video object removal, we employ four established metrics: PSNR [14], SSIM [47], VFID [43], and LPIPS [62]. PSNR and SSIM are widely used for measuring low-level pixel-wise similarity between the generated video and the ground truth. To further assess perceptual quality, we use LPIPS, which leverages deep features to better align with human perception. We use VFID, an adaptation of FID for videos, to evaluate the overall temporal consistency and visual realism of the completed video sequences. Additionally, we also introduced flow warping error [17] as a metric for measuring temporal consistency in the zero-shot validation. We strictly follow the settings of ProPainter [65] for calculating metrics.

Table 1: Quantitative comparisons on RORD, ROVI, and VPLM datasets. ↑ indicates higher is better. ↓ indicates lower is better. The best and second performances are highlighted in bold and underlined.

Models	RORD				ROVI				VPLM			
	PSNR ↑	SSIM ↑	VFID ↓	LPIPS ↓	PSNR ↑	SSIM ↑	VFID ↓	LPIPS ↓	PSNR ↑	SSIM ↑	VFID ↓	LPIPS ↓
FLoED [12]	26.06	0.8263	0.471	0.1557	34.42	0.9225	0.138	0.0763	34.85	0.9149	0.493	0.0816
E^{2}FGVI [21]	26.95	0.8421	0.311	0.1340	38.01	0.9693	0.096	0.0337	38.29	0.9652	0.424	0.0390
ROSE [29]	27.94	0.8959	0.270	0.1105	35.72	0.9642	0.099	0.0379	36.19	0.9520	0.378	0.0473
FuseFormer [23]	27.99	0.8659	0.306	0.1431	41.21	0.9813	0.064	0.0292	41.24	0.9814	0.236	0.0364
DiffuEraser [20]	28.64	0.8865	0.268	0.1205	38.79	0.9769	0.093	0.0323	38.76	0.9719	0.298	0.0421
FGT [60]	30.13	0.8972	0.271	0.1042	37.80	0.9725	0.089	0.0325	38.51	0.9725	0.350	0.0368
ProPainter [65]	30.97	0.9132	0.230	0.1020	40.86	0.9759	0.069	0.0263	41.15	0.9808	0.241	0.0374
MiniMax-Remover [66]	30.60	0.9166	**0.224**	0.0947	38.65	0.9775	0.069	0.0318	39.47	0.9744	0.246	0.0381
D2DF-DG(Ours)	**32.20**	**0.9318**	0.251	**0.0902**	**42.41**	0.9888	**0.062**	**0.0227**	**43.31**	**0.9877**	**0.191**	**0.0337**
D2DF-DF(Ours)	31.56	0.9202	0.268	0.1001	42.25	**0.9889**	0.063	0.0282	43.05	0.9864	0.196	0.0348



(a) PSNR (b) SSIM

Fig. 5: Comparison of different models at different mask ratios.

Table 2: Efficiency comparison of diffusion-based methods. "Prior" indicates the inference time of the prior extraction stage. "Denoising" indicates the denoising time. "Step" indicates the number of iterations for denoising.

Models	Prior(s)	Denoise(s)	Steps	Total Time(s)
ROSE [29]	–	45.76	50	45.76
FLoED [12]	2.66	31.49	25	34.15
DiffuEraser [20]	2.38	2.85	2	5.13
D2DF-DG	2.38	1.03	**1**	3.41
D2DF-DF	**0.02**	**1.03**	**1**	**1.05**

4.2 Comparisons

Quantitative Evaluation. We report quantitative results on three datasets. We compare the proposed D2DF-DG and D2DF-DF with previous video object removal methods, including FuseFormer [23], FGT [60], E^{2}FGVI [21], FLoED [12], DiffuEraser [20], ROSE [29] and ProPainter [65]. Among them, FLoED, DiffuEraser, and ROSE are methods based on diffusion models. As shown in Table.1, D2DF-DG achieves the best performance on most metrics and consistently delivers strong results across all three benchmarks. Although D2DF-DF exhibits slightly lower performance compared to D2DF-DG, it still leads state-of-the-art methods in PSNR. This fully demonstrates the powerful capability of our approach in object removal. Additionally, we observe a significant number of large objects in the RORD dataset. Therefore, we select the best models from diffusion-based (DiffuEraser) and traditional approaches (Propainter) and compare their performance at different mask ratios. As shown in Fig.5, our models demonstrate only a marginal advantage over Propainter and DiffuEraser within the 0-20% mask ratio range. However, as the mask ratio increases, our models show a more pronounced advantage in both PSNR and SSIM metrics. This highlights the robustness of our model and its superiority in removing large-area objects.

Qualitative Evaluation. For qualitative evaluation, we select diffusion-based methods ROSE [29] and DiffuEraser [20] alongside the lightweight transformer-based ProPainter

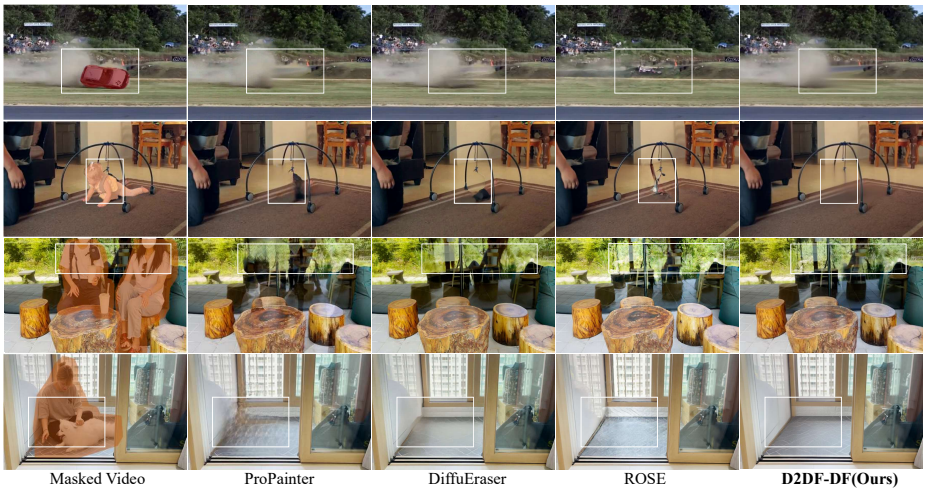
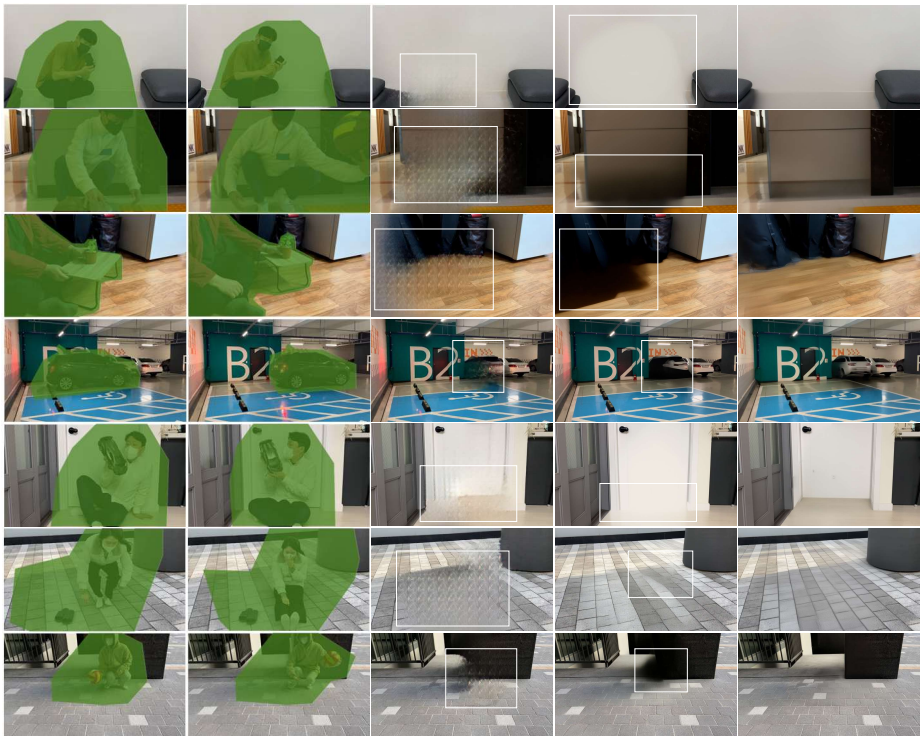


Fig. 6: Qualitative comparisons on video object removal. Our D2DF-DF model demonstrates exceptionally strong object removal and background reconstruction capabilities.

[65] for comparison. As shown in Fig.6, our method D2DF-DF demonstrates strong capabilities for reconstructing background regions. When removing objects from larger areas in the last two rows, D2DF-DF still delivers highly stable performance. However, other methods exhibit issues such as inconsistent texture information, blurring, and artifacts. This highlights the powerful performance of our one-step denoising model in video object removal. The D2DF-DG delivers even more powerful results. Due to space limitations, you can refer to the appendix for details.

Efficiency Comparison with Diffusion-based Methods. Tab.2 compares the inference times of different diffusion-based methods. We test each model on a NVIDIA A100 GPU to generate 25 frames of 480×720 video, measuring the time required for both prior extraction and denoising. Results indicate that DiffuEraser [20], also utilizing ProPainter, achieves performance extremely close to our D2DF-DG. However, when switching to our SGFP module for prior extraction, we observed that this module completes the task in just 0.02 seconds. Consequently, D2DF-DF delivers the fastest inference time while maintaining superior performance. Including the time for encoding and decoding latents, our complete processing time is within 10 seconds.

Qualitative Analysis Under Extreme Occlusion Conditions. In this section, we will discuss how different models perform under extreme occlusion conditions. When objects in a video have a limited range of motion or occupy a large area, the background regions occluded by these objects cannot be obtained through frame propagation. We define this scenario as extreme occlusion. RORD [34] employs a polygonal mask annotation method, resulting in a dataset with an exceptionally high number of extreme occlusions, making it highly challenging. We selected the state-of-the-art methods ProPainter [65] and MiniMax-Remover [66], based on optical flow and diffusion models for comparison. As shown in Fig.7, for regions where information cannot prop-



(a) First Frame (b) Last Frame (c) ProPainter (d) MiniMax-Remover (e) D2DF-DF

Fig. 7: Qualitative comparisons under extreme occlusion conditions. Our DF model achieves remarkable structural consistency and generative performance in one step.

agate between frames, ProPainter directly produces blurred textures, while MiniMax-Remover (6 steps) exhibits noise artifacts. In contrast, our D2DF-DF model achieves excellent background reproduction with just one step. This demonstrates a unique advantage of our model under extreme occlusion conditions.

4.3 Robustness and Generalization Analysis

Zero-shot Evaluation on Camera-Bench. To further evaluate the generalization ability of our method, we conduct zero-shot experiments on Camera-Bench [22], as shown in Tab. 3. Without training on this benchmark, both D2DF-DG and D2DF-DF remain highly competitive against existing methods. It is worth noting that our framework is not inherently tied to optical flow. Whether optical flow is involved depends on the draft source used by D2DF-DG. Although D2DF-DG is trained with ProPainter drafts, it can still effectively refine FuseFormer drafts, which are generated without optical flow. This suggests that the cross-domain robustness of our method is not limited to dataset transfer, but also extends to different types of draft priors. Furthermore, D2DF-DF removes the external draft dependency entirely, making the final model fully free from optical-flow-related concerns during inference.

Table 3: Zero-shot performance on Camera-Bench [22]. "Draft" denotes whether the method requires an external draft during inference, and "Flow" denotes whether optical flow is involved in its inference pipeline. E_{warp} is reported in 10^{-4}. The best results are highlighted in bold.

Method	Draft	Flow	PSNR↑	SSIM↑	LPIPS↓	E_{warp} ↓ ($\times 10^{-4}$)
FuseFormer [23]	–	✗	25.05	0.9155	0.2020	0.70
ProPainter [65]	–	✓	25.42	0.9203	0.1268	1.10
ROSE [29]	–	✗	26.45	0.9286	0.0860	0.55
MiniMax-Remover [66]	–	✗	26.63	0.9324	0.0854	0.37
D2DF-DG(Ours)	FuseFormer	✗	**27.08**	0.9421	**0.0831**	0.41
D2DF-DG(Ours)	ProPainter	✓	26.96	**0.9429**	0.0884	0.93
D2DF-DF(Ours)	–	✗	26.57	0.9365	0.0896	**0.27**

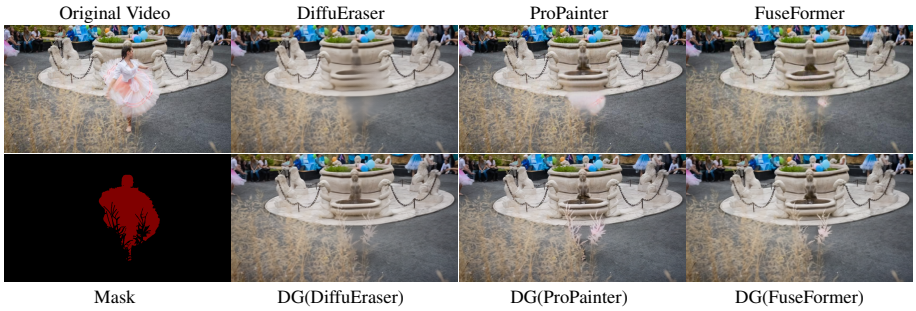


Fig. 8: Visual verification of the generalization of draft sensitivity on a sample from DAVIS. The results demonstrate that our DG can refine various types of drafts.

Draft-source Robustness. We further evaluate D2DF-DG with drafts from different sources. As shown in Fig. 8, although our model is trained with ProPainter drafts, it can consistently refine drafts generated by DiffuEraser, ProPainter, and FuseFormer, producing similarly high-quality removal results. This demonstrates that D2DF-DG does not overfit to a specific draft distribution, but learns a robust refinement ability across heterogeneous draft priors. Notably, the effectiveness on FuseFormer drafts also indicates that our framework can work with flow-free draft sources, while D2DF-DF further removes the draft dependency entirely.

4.4 Ablation Study

Study of Prior-Privileged Consistency Distillation. In Tab.4, when we train directly using the LDM approach, the model behaves as a standard LDM model, yielding suboptimal results in a single step. Consequently, employing Consistency Distillation (CD) significantly improves performance. However, introducing privileged prior knowledge (replacing the draft conditions with ground truth) further enhances results. This indicates that our ground-truth prior indeed yields a more stable consistency trajectory.
Study of Self-Guided Fast Planting Module. To investigate the effectiveness of our proposed SGFP module, we conduct detailed ablation experiments on the RORD dataset. As shown in Tab.5, the first row represents the baseline model, indicating the performance without any prior knowledge. Adding the transformer layer (TL) yields an initial

Table 4: Ablation study of proposed Prior-Privileged Consistency Distillation (PPCD). “D.T." means directly train the diffusion model as Eq.1, “CD" means the Consistency Distillation [39].

Training Methods			Metrics			
D.T.	CD	PPCD	PSNR↑	SSIM↑	VFID↓	LPIPS↓
✓			31.54	0.9192	0.271	0.1014
	✓		31.85	0.9296	0.257	0.1002
		✓	**32.20**	**0.9318**	**0.251**	**0.0902**

Table 5: Ablation study of our SGFP module. The first row means the baseline model without any extra prior. “TL" means the transformer layers, “VP” indicates whether to exclude masked patches as valid patches, “SW” indicates the spatial window constraint, and “MR” indicates reconstruction of the masked region only.

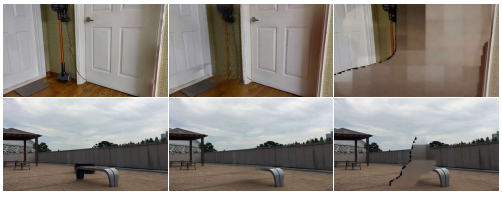
Components				Metrics		
TL	VP	SW	MR	PSNR↑	SSIM↑	LPIPS↓
				29.36	0.8920	0.1426
✓				30.73	0.9171	0.1149
✓	✓			30.78	**0.9228**	0.1093
✓	✓	✓		31.47	0.9197	0.1070
✓	✓	✓	✓	**31.56**	0.9202	**0.1001**

improvement in model performance. Excluding masked regions (VP) results in a slight further enhancement. Introducing the spatial window (SW) significantly boosts model effectiveness. Finally, reconstructing only the masked region (MR) achieves optimal model performance. These ablation experiments demonstrate the rationality and effectiveness of each design step. To validate that the latents generated by SGFP indeed contain meaningful background information, we additionally trained a diagnostic decoder. This decoder is solely used for visualization. As shown in Fig.9(c), our pseudo-drafts exhibit blocky filling because they are learned from many patch latents. Despite its visual coarseness compared to the draft in Fig.9(b), this decoding result demonstrates that SGFP successfully captures key background color tones and structures, which suffice as effective prior information.

Study of Three-stage Distillation Framework. To validate the effectiveness of the three-stage framework, we conducted experiments as shown in Table.6. As the table indicates, the performance achieved when the teacher network is LDM was not as good as that obtained with D2DF-DG as the teacher network. This shows the trained D2DF-DG model can provide more stable and consistent solution trajectories than the diffusion D-LDM. Furthermore, the comparison between CD and PPCD also demonstrated the superiority of the PPCD method.

4.5 Analysis of Failure Cases

In this section, we will discuss cases where the model fails to remove objects and analyze the reasons for these failures. We observe that when multiple objects exist in the video and these objects exhibit dynamic and complex occlusion relationships, the model



(a) Ground Truth (b) Draft (c) Pseudo-Draft

Fig. 9: Visualization of Drafts and Pseudo-Drafts. (a) Ground Truth. (b) Drafts with artifacts. (c) Pseudo-drafts from decoder.

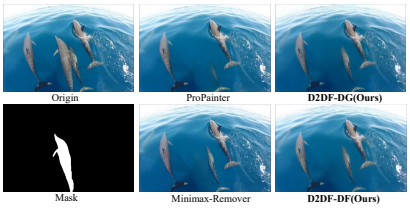


Fig. 10: Analysis of failure cases. Object removal models fail when complex occlusions exist between objects.

Table 6: Ablation study of our three-stage framework to obtain D2DF-DF. “CD" means the Consistency Distillation [39]. The best results are achieved by using PPCD for distilling D2DF-DG.

Training Settings				Metrics		
Student	Teacher	CD	PPCD	PSNR↑	SSIM↑	LPIPS↓
D2DF-DF	D-LDM	✓		31.3268	0.9179	0.1025
	D-LDM		✓	31.4484	0.9197	0.1014
	D2DF-DG	✓		31.3257	0.9173	0.1036
	D2DF-DG		✓	**31.5646**	**0.9202**	**0.1001**

is unable to fully reconstruct the occluded objects. As shown in Fig.10, we compared the state-of-the-art optical flow-based method ProPainter [65] and the state-of-the-art diffusion model-based method MiniMax-Remover [66]. The results indicate that neither method can fully restore the occluded “dolphin” in this example. Our two models, D2DF-DG and D2DF-DF, also produce imperfect results under one-step inference settings. However, they retain significantly more context compared to MiniMax-Remover. This also demonstrates the superiority of our approach.

5 Conclusion

In this paper, we present D2DF, a unified framework that progressively evolves from a draft-guided diffusion teacher to a fully draft-free one-step video object removal model. By introducing Prior-Privileged Consistency Distillation (PPCD), we effectively stabilize the distillation process and enable high-quality one-step generation. Furthermore, the proposed Self-Guided Fast Planting (SGFP) module eliminates the dependency on external draft inputs by generating scene-consistent pseudo-drafts directly in latent space. Our method not only excels in visual quality and temporal coherence but also achieves remarkable efficiency. However, the blurring that occurs in certain scenarios with the one-step denoising is an enhancement needed in future work.

Acknowledgements. Zizhao Chen and Ping Wei acknowledge the support of the National Natural Science Foundation of China (No. U23B2060, No. 62495092). Mengmeng Wang acknowledges the support of the National Natural Science Foundation of China (No. 62403429). We also acknowledge the computational resources provided by SGIT AI Lab, State Grid Corporation of China.

References

1. Bertalmio, M., Bertozzi, A.L., Sapiro, G.: Navier-stokes, fluid dynamics, and image and video inpainting. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. vol. 1, pp. I–I. IEEE (2001)
2. Bian, Y., Zhang, Z., Ju, X., Cao, M., Xie, L., Shan, Y., Xu, Q.: Videopainter: Any-length video inpainting and editing with plug-and-play context control. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–12 (2025)
3. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
4. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22563–22575 (2023)
5. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al.: Video generation models as world simulators. OpenAI Blog **1**(8), 1 (2024)
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
7. Ding, D., Pan, Y., Feng, R., Dai, Q., Qiu, K., Bao, J., Luo, C., Chen, Z.: Homogen: Enhanced video inpainting via homography propagation and diffusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22953–22962 (2025)
8. Ebdelli, M., Le Meur, O., Guillemot, C.: Video inpainting with short-term windows: application to object removal and error concealment. IEEE Transactions on Image Processing **24**(10), 3034–3047 (2015)
9. Gao, C., Saraf, A., Huang, J.B., Kopf, J.: Flow-edge guided video completion. In: European Conference on Computer Vision. pp. 713–729. Springer (2020)
10. Geng, Z., Pokle, A., Luo, W., Lin, J., Kolter, J.Z.: Consistency models made easy. arXiv preprint arXiv:2406.14548 (2024)
11. Gu, B., Luo, H., Guo, S., Dong, P.: Advanced video inpainting using optical flow-guided efficient diffusion. arXiv e-prints pp. arXiv–2412 (2024)
12. Gu, B., Luo, H., Guo, S., Dong, P., Zhou, Q.: Coherent video inpainting using optical flow-guided efficient diffusion. arXiv preprint arXiv:2412.00857 (2024)
13. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
14. Hore, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. In: 2010 20th international conference on pattern recognition. pp. 2366–2369. IEEE (2010)
15. Jin, W., Dai, Q., Luo, C., Baek, S.H., Cho, S.: Flovd: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2040–2049 (2025)
16. Kim, D., Woo, S., Lee, J.Y., Kweon, I.S.: Deep video inpainting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5792–5801 (2019)
17. Lai, W.S., Huang, J.B., Wang, O., Shechtman, E., Yumer, E., Yang, M.H.: Learning blind video temporal consistency. In: Proceedings of the European conference on computer vision (ECCV). pp. 170–185 (2018)
18. Lee, M., Cho, S., Shin, C., Lee, J., Yang, S., Lee, S.: Video diffusion models are strong video inpainter. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4526–4533 (2025)

19. Lee, S., Oh, S.W., Won, D., Kim, S.J.: Copy-and-paste networks for deep video inpainting. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4413–4421 (2019)
20. Li, X., Xue, H., Ren, P., Bo, L.: Diffueraser: A diffusion model for video inpainting. arXiv preprint arXiv:2501.10018 (2025)
21. Li, Z., Lu, C.Z., Qin, J., Guo, C.L., Cheng, M.M.: Towards an end-to-end framework for flow-guided video inpainting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17562–17571 (2022)
22. Liu, D., Wang, W., Li, C., Lyu, J.: From understanding to erasing: Towards complete and stable video object removal. arXiv preprint arXiv:2604.01693 (2026)
23. Liu, R., Deng, H., Huang, Y., Shi, X., Lu, L., Sun, W., Wang, X., Dai, J., Li, H.: Fuseformer: Fusing fine-grained information in transformers for video inpainting. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14040–14049 (2021)
24. Lu, C., Song, Y.: Simplifying, stabilizing and scaling continuous-time consistency models. arXiv preprint arXiv:2410.11081 (2024)
25. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. Advances in neural information processing systems **35**, 5775–5787 (2022)
26. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint arXiv:2310.04378 (2023)
27. Luo, S., Tan, Y., Patil, S., Gu, D., von Platen, P., Passos, A., Huang, L., Li, J., Zhao, H.: Lcm-lora: A universal stable-diffusion acceleration module. arXiv preprint arXiv:2311.05556 (2023)
28. Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J., Salimans, T.: On distillation of guided diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14297–14306 (2023)
29. Miao, C., Feng, Y., Zeng, J., Gao, Z., Liu, H., Yan, Y., Qi, D., Chen, X., Wang, B., Zhao, H.: Rose: Remove objects with side effects in videos. arXiv preprint arXiv:2508.18633 (2025)
30. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
31. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 724–732 (2016)
32. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
34. Sagong, M.C., Yeo, Y.J., Jung, S.W., Ko, S.J.: Rord: A real-world object removal dataset. In: BMVC. p. 542 (2022)
35. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022)
36. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. In: European Conference on Computer Vision. pp. 87–103. Springer (2024)
37. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
38. Song, Y., Dhariwal, P.: Improved techniques for training consistency models. arXiv preprint arXiv:2310.14189 (2023)

39. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. In: Proceedings of the 40th International Conference on Machine Learning. pp. 32211–32252 (2023)
40. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
41. Tang, N.C., Hsu, C.T., Su, C.W., Shih, T.K., Liao, H.Y.M.: Video inpainting on digitized vintage films via maintaining spatiotemporal continuity. IEEE Transactions on Multimedia **13**(4), 602–614 (2011)
42. Wang, C., Huang, H., Han, X., Wang, J.: Video inpainting by jointly learning temporal structure and spatial details. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 5232–5239 (2019)
43. Wang, T.C., Liu, M.Y., Zhu, J.Y., Liu, G., Tao, A., Kautz, J., Catanzaro, B.: Video-to-video synthesis. arXiv preprint arXiv:1808.06601 (2018)
44. Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: Videocomposer: Compositional video synthesis with motion controllability. Advances in Neural Information Processing Systems **36**, 7594–7611 (2023)
45. Wang, Y., Bao, J., Weng, W., Feng, R., Yin, D., Yang, T., Zhang, J., Dai, Q., Zhao, Z., Wang, C., et al.: Microcinema: A divide-and-conquer approach for text-to-video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8414–8424 (2024)
46. Wang, Z., Wang, J., Li, X., Li, Y.L., Lu, Y., Wang, S.: Unsupervised temporal correspondence learning for unified video object removal. IEEE Transactions on Image Processing (2023)
47. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
48. Wei, Y., Zhang, S., Qing, Z., Yuan, H., Liu, Z., Liu, Y., Zhang, Y., Zhou, J., Shan, H.: Dreamvideo: Composing your dream videos with customized subject and motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6537–6549 (2024)
49. Weng, W., Feng, R., Wang, Y., Dai, Q., Wang, C., Yin, D., Zhao, Z., Qiu, K., Bao, J., Yuan, Y., et al.: Art-v: Auto-regressive text-to-video generation with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7395–7405 (2024)
50. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7623–7633 (2023)
51. Wu, J., Li, X., Si, C., Zhou, S., Yang, J., Zhang, J., Li, Y., Chen, K., Tong, Y., Liu, Z., et al.: Towards language-driven video inpainting via multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12501–12511 (2024)
52. Xing, J., Xia, M., Zhang, Y., Chen, H., Yu, W., Liu, H., Liu, G., Wang, X., Shan, Y., Wong, T.T.: Dynamicrafter: Animating open-domain images with video diffusion priors. In: European Conference on Computer Vision. pp. 399–417. Springer (2024)
53. Xing, Z., Dai, Q., Hu, H., Wu, Z., Jiang, Y.G.: Simda: Simple diffusion adapter for efficient video generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7827–7839 (2024)
54. Xu, N., Yang, L., Fan, Y., Yang, J., Yue, D., Liang, Y., Price, B., Cohen, S., Huang, T.: Youtube-vos: Sequence-to-sequence video object segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 585–601 (2018)

55. Xu, R., Li, X., Zhou, B., Loy, C.C.: Deep flow-guided video inpainting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3723–3732 (2019)
56. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
57. Yoon, J., Yu, S., Bansal, M.: RACCooN: Versatile instructional video editing with auto-generated narratives. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 27960–27996. Association for Computational Linguistics (2025)
58. Zeng, Y., Wei, G., Zheng, J., Zou, J., Wei, Y., Zhang, Y., Li, H.: Make pixels dance: High-dynamic video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8850–8860 (2024)
59. Zeng, Y., Fu, J., Chao, H.: Learning joint spatial-temporal transformations for video inpainting. In: European conference on computer vision. pp. 528–543. Springer (2020)
60. Zhang, K., Fu, J., Liu, D.: Flow-guided transformer for video inpainting. In: European conference on computer vision. pp. 74–90. Springer (2022)
61. Zhang, K., Fu, J., Liu, D.: Inertia-guided flow completion and style fusion for video inpainting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5982–5991 (2022)
62. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: 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)
63. Zhang, R., Li, W., Wang, P., Guan, C., Fang, J., Song, Y., Yu, J., Chen, B., Xu, W., Yang, R.: Autoremover: Automatic object removal for autonomous driving videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12853–12861 (2020)
64. Zheng, H., Nie, W., Vahdat, A., Azizzadenesheli, K., Anandkumar, A.: Fast sampling of diffusion models via operator learning. In: International conference on machine learning. pp. 42390–42402. PMLR (2023)
65. Zhou, S., Li, C., Chan, K.C., Loy, C.C.: Propainter: Improving propagation and transformer for video inpainting. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10477–10486 (2023)
66. Zi, B., Peng, W., Qi, X., Wang, J., Zhao, S., Xiao, R., Wong, K.F.: Minimax-remover: Taming bad noise helps video object removal. arXiv preprint arXiv:2505.24873 (2025)