FASTERASER: ACHIEVING EFFICIENT AND EFFECT-AWARE VIDEO OBJECT REMOVAL VIA SPARSITY-DRIVEN DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

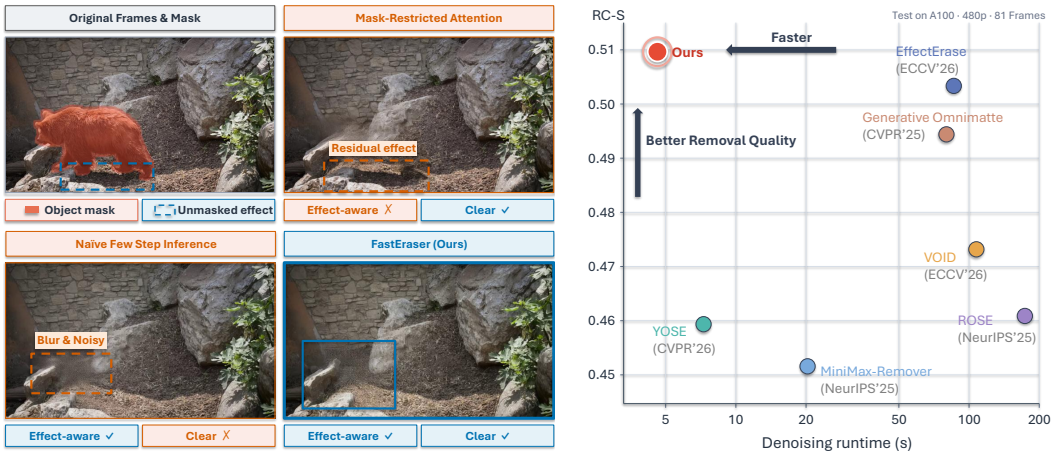
Effect-aware video object removal (VOR) aims to eliminate not only a target object but also its associated environmental effects, such as shadows and reflections. Although recent generation-based VOR methods achieve compelling reconstruction quality, their billion-scale architectures and iterative denoising processes incur substantial computational cost. Existing acceleration strategies either apply mask-guided spatial sparsity, which overlooks effects outside the input mask, or directly reduce the number of denoising steps, which introduces noticeable blur and noise. Few-step distillation offers a promising alternative, yet conventional Distribution Matching Distillation (DMD) relies heavily on the asymmetry induced by classifier-free guidance (CFG), making it poorly suited to VOR models that are inherently CFG-free. We propose FastEraser, an efficient framework for effect-aware VOR built upon Sparsity-Driven Distillation (SDD) and Pixel Reconstruction Loss (PRL). SDD replaces CFG-induced asymmetry with a dense-to-sparse attention discrepancy between the real and fake score models, restoring the driving signal required for stable few-step adaptation without restricting removal to the input mask. PRL further anchors the student prediction to the ground-truth video by decoding its denoised latent through a frozen VAE and applying direct supervision in RGB space, thereby preserving removal capability and high-frequency details. Extensive experiments on four benchmarks demonstrate that FastEraser achieves performance comparable to existing 50-step models using only 4 denoising steps, substantially reducing inference cost while maintaining effect removal quality and reconstruction fidelity.

1 INTRODUCTION

Video object removal (VOR) is a fundamental task in video editing, with broad applications in film production and content creation. With the rapid progress of video generation Kong et al. (2024); Yang et al. (2025); Wan et al. (2025), an increasing number of VOR methods have been developed on top of pretrained video generation models, leveraging their strong generative priors to achieve high-quality reconstruction of occluded regions Zi et al. (2026); Motamed et al. (2026); Miao et al. (2026); Wu et al. (2026a); Lee et al. (2025); Fu et al. (2026). Although these models substantially improve object removal quality, they incur considerable computational overhead. These models often contain billions of parameters and rely on tens of iterative denoising steps, resulting in inference times of tens of seconds per video.

To reduce this cost, YOSE Wu et al. (2026a) uses only tokens corresponding to the user-provided object mask as queries in the attention computation of the diffusion transformer (DiT) Peebles & Xie (2023), thereby reducing the cost of attention over long video sequences. As illustrated by the efficiency–quality comparison in Fig. 1 (b), this strategy substantially lowers the inference cost. However, it has a fundamental limitation when applied to the increasingly important task of *effect-aware VOR*. In this task, the model receives only an object mask and must independently identify and remove secondary visual effects associated with the target object, such as shadows and reflections. Because these effects often extend beyond the object mask, restricting attention queries to masked tokens can leave the regions containing those effects insufficiently updated. Consequently, residual effects remain visible after removal, as shown in the upper-right example in Fig. 1 (a). This limitation

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



(a) Effect removal and reconstruction quality.

(b) Removal quality versus denoising runtime.

Figure 1: Existing acceleration methods either introduce blur or fail on effect-aware VOR. Our FastEraser enables fast inference while removing both objects and their effects, achieving the best efficiency–quality trade-off.

motivates us to pursue a more broadly applicable acceleration strategy, such as reducing the number of denoising steps.

However, we observe that direct step reduction, e.g., from 50 steps to 2 or 4 steps, introduces conspicuous blur and noise into the reconstructed regions, as illustrated in the lower-left example of Fig. 1 (a). Such degradation severely limits the usability of the resulting videos for subsequent editing. This observation motivates us to explore post-training distillation techniques that explicitly adapt the model to few-step inference, e.g., Distribution Matching Distillation (DMD) Yin et al. (2024b;a). DMD has demonstrated strong performance in distilling multi-step text-to-image/video diffusion models into few-step generators. However, its underlying mechanism reveals a structural mismatch with many VOR methods. As discussed in Decoupled DMD Liu et al. (2026), DMD can be viewed as a combination of CFG augmentation and distribution matching. CFG Ho & Salimans (2022) Augmentation provides the key driving signal for step distillation: it exploits the discrepancy between a CFG-enabled real score model and a CFG-free fake score model to steer the student toward few-step generation. DMD is feasible in text-to-image/video models, where CFG is widely used. In contrast, VOR is a strongly conditioned generation task that does not rely on explicit text prompts, and many existing VOR models are therefore natively CFG-free Lee et al. (2025); Fu et al. (2026); some methods, such as MiniMax-Remover Zi et al. (2026), even eliminate cross-attention modules used to inject text embeddings in the DiT. Consequently, both the real and fake score models operate without CFG, eliminating the CFG-induced asymmetry that drives conventional DMD. Without this critical signal, our experiments show that directly applying DMD to VOR models leads to unstable distillation and degrades the student model's removal capability.

To address this challenge, we introduce Sparsity-Driven Distillation (SDD), which establishes an asymmetry between the real score model and fake score model under identical CFG-free settings. SDD constructs this asymmetry through attention sparsity: the real score model retains dense attention, whereas the fake score model employs weaker sparse attention. The resulting dense-to-sparse score discrepancy supplies an explicit driving signal for few-step adaptation, enabling the DMD framework to operate effectively even when the underlying VOR model is natively CFG-free. Because attention sparsity applies to all queries independently of the object mask, SDD can still remove associated effects beyond the masked region.

We further observe that, for the challenging task of effect-aware VOR, using SDD as the sole learning signal during distillation can degrade the student's removal capability. We therefore introduce Pixel Reconstruction Loss (PRL), which explicitly constrains the student's removal ability using ground-truth clean videos. PRL directly estimates the clean latent from the student's predicted velocity and decodes it into RGB frames using a frozen VAE decoder. Reconstruction supervision is then applied against the ground-truth video in RGB space, with gradients propagated through the

decoder to the student model. Compared with directly adding a flow-matching velocity loss to the student model in latent space, this pixel-space supervision improves the visual naturalness of the generated videos while reducing blur and noise in the reconstructed regions.

Building on above strategies, we introduce FastEraser, a powerful and efficient model for effect-aware video object removal. We evaluate FastEraser extensively on four benchmarks Fu et al. (2026); Li et al. (2026) using complementary metrics, including the removal-oriented RC-S and RC-T metrics Li et al. (2026), the pixel-level PSNR and SSIM metrics, and scores produced by VLM. The experimental results demonstrate that FastEraser achieves performance comparable to existing 50-step models using only 4 denoising steps, substantially reducing inference cost without sacrificing effect removal or reconstruction fidelity. These results establish FastEraser as an efficient and practical solution for effect-aware video object removal.

2 Related Work

2.1 Video Object Removal

Video object removal (VOR) has traditionally been formulated as mask-guided video inpainting, where the region occupied by a target object is filled using spatial context and temporal correspondence. Early learning-based approaches propagate visible content through optical flow or aligned features, as exemplified by E2FGVI Li et al. (2022) and ProPainter Zhou et al. (2023). More recently, diffusion priors have enabled stronger content synthesis for large missing regions,e.g., Minimax-remover Zi et al. (2026) and Diffueraser Li et al. (2025). However, these methods primarily remove only the object specified by the user-provided mask and cannot explicitly eliminate object-induced effects outside the masked region, such as shadows and reflections. To address this limitation, a growing body of recent work Lu et al. (2021); Lin et al. (2023); Suhail et al. (2023); Lee et al. (2025) requires models to recover a clean background in which both the target object and its associated effects are removed. Such effect-aware removal produces more visually complete results and better reflects the requirements of practical applications. ROSE Miao et al. (2026) and EffectErase Fu et al. (2026) construct dedicated training datasets using 3D rendering engines or real-world data collection, while VOID Motamed et al. (2026) further extends this direction by exploring the removal of physical consequences caused by the target object, such as collisions and other object–environment interactions. Overall, recent research in video object removal exhibits a clear trend toward increasingly comprehensive removal of both target objects and their broader visual and physical effects.

2.2 Efficient Video Generation

The high computational cost of video diffusion models has motivated several acceleration strategies. One major direction is sampling-step distillation, which reduces the number of DiT forward passes at inference time. Progressive Distillation Salimans & Ho (2022) repeatedly halves the sampling steps by training each student step to reproduce two steps of a deterministic teacher sampler. Consistency Models Song et al. (2023) learn to produce the same prediction from points along a single probability-flow ODE trajectory; LCM Luo et al. (2023) adapts this principle to latent diffusion models. In contrast, DMD Yin et al. (2024b;a) matches the student and teacher output distributions, while CausVid Yin et al. (2025) extends distribution matching to efficient autoregressive video generation.

Another complementary line of work reduces the computation through sparse attention. Originally developed for long-context large language models, MoBA Lu et al. (2026) partitions key–value tokens into blocks and dynamically routes each query only to its most relevant blocks, establishing a content-adaptive block-sparsity paradigm. This idea has subsequently been adapted to video DiTs. Sparse VideoGen Xi et al. (2025) profiles spatial and temporal attention patterns and dispatches specialized sparse kernels, while VMoBA Wu et al. (2026b) introduces video-specific spatiotemporal block partitioning and dynamic block selection. VSA Zhang et al. (2026b) learns a differentiable coarse-to-fine sparse router for both training and inference, and SLA Zhang et al. (2026a) combines exact sparse attention for critical interactions with linear attention for lower-importance ones.

VSA Zhang et al. (2026b) and BLADE Gu et al. (2026) also combine step distillation with sparse attention for video generation. However, effect-aware VOR differs fundamentally from the tasks addressed by these methods, as discussed in Section A.1.

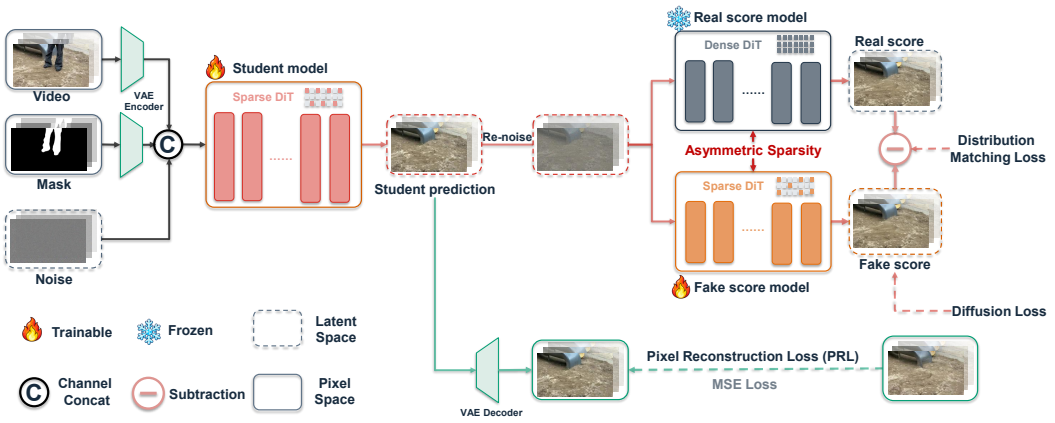


Figure 2: Training framework of our FastEraser. SDD exploits a dense real and a sparse fake score model to drive few-step distillation, while PRL directly supervises the decoded student prediction in RGB space. Flame and snowflake symbols denote trainable and frozen modules.

3 METHOD

We first review Distribution Matching Distillation and sparse attention in Section 3.1, then introduce Sparsity-Driven Distillation and Pixel Reconstruction Loss in Sections 3.2 and 3.3, respectively. The FastEraser training framework is illustrated in Figure 2.

3.1 PRELIMINARIES

Distribution Matching Distillation. Unlike trajectory-based methods such as LCM Luo et al. (2023), which align a few-step student with the denoising trajectory of a multi-step teacher, Distribution Matching Distillation (DMD) Yin et al. (2024b;a) distills a multi-step diffusion model by matching its output distribution with that of a few-step generator. Let G_θ denote the student generator, where $\mathbf{z}$ is the initial Gaussian noise and c is the conditioning input. Given a student sample $\mathbf{x}_0^g = G_\theta(\mathbf{z}, c)$, DMD perturbs it to a randomly sampled noise level, $\mathbf{x}_t^g \sim q_t(\mathbf{x}_t \mid \mathbf{x}_0^g)$, and minimizes

$$\mathcal{L}_{\mathrm{DM}} = \mathbb{E}_t \left[D_{\mathrm{KL}} \left(p_{\theta,t}(\mathbf{x}_t \mid c) \,\|\, p_{\mathrm{real},t}(\mathbf{x}_t \mid c) \right) \right]. \tag{1}$$

Since these probability densities are intractable, DMD estimates the student gradient using a frozen real score model and an online fake score model:

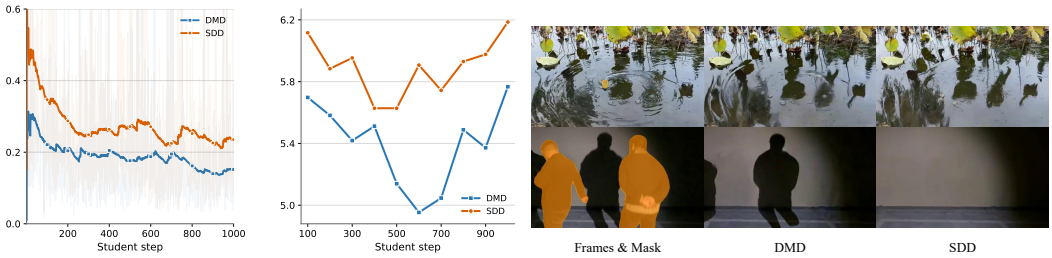
$$\nabla_\theta \mathcal{L}_{\mathrm{DM}} \propto \mathbb{E} \left[\left(s_{\mathrm{fake}}(\mathbf{x}_t^g, t, c) - s_{\mathrm{real}}(\mathbf{x}_t^g, t, c) \right) \frac{\partial G_\theta(\mathbf{z}, c)}{\partial \theta} \right]. \tag{2}$$

The frozen real score model represents the target distribution defined by the pretrained teacher, while the fake score model is continuously updated on student-generated samples to track the evolving student distribution. Their score difference provides the update direction for aligning the student distribution with the target distribution.

Sparse Attention in DiTs. Self-attention is computationally expensive in video diffusion transformers because a video produces a long sequence of spatiotemporal tokens. Recent studies such as MoBA and SLA Lu et al. (2026); Zhang et al. (2026a) show that each query can attend to only a subset of the most relevant keys while producing results comparable to those of dense attention.

In practice, queries, keys, and values are partitioned into spatiotemporal blocks. Each block is represented by a pooled feature, and a lightweight routing score is computed between every pooled query block $\overline{\mathbf{q}}_i$ and key block $\overline{\mathbf{k}}_j$. The highest-scoring key–value blocks are then selected according to a retention ratio k:

$$r_{ij} = \frac{\overline{\mathbf{q}}_i \overline{\mathbf{k}}_j^\top}{\sqrt{d}}, \qquad \Omega_i = \mathrm{TopK}_j \left(r_{ij}, \lceil k n_k \rceil \right), \tag{3}$$



(a) Score discrepancy. (b) Overall Performance. (c) Qualitative comparison between DMD and SDD.

Figure 3: Oberservation and effect of Sparsity-Driven Distillation.

where n_k is the total number of key–value blocks, k (e.g., $k = 10\%$)denotes the fraction retained for each query block, and Ω_i contains the selected blocks. After routing, exact attention is evaluated only over the selected keys and values:

$$\mathbf{O}_i = \mathrm{Softmax}\left(\frac{\mathbf{Q}_i \mathbf{K}_{\Omega_i}^{\top}}{\sqrt{d}}\right)\mathbf{V}_{\Omega_i}. \tag{4}$$

Consequently, the dominant attention complexity is reduced from $\mathcal{O}(N^2 d)$ to approximately $\mathcal{O}(kN^2 d)$, while exact attention is preserved for the selected interactions.

3.2 Sparsity-Driven Distillation

Mask-guided sparsification is inherently ill-suited to effect-aware VOR, since shadows, reflections, and other object-induced effects may appear anywhere outside the input mask. We therefore turn to sampling-step distillation, which provides a more general acceleration strategy without restricting the editable region. Among existing techniques, Distribution Matching Distillation (DMD) Yin et al. (2024b;a) is one of the most widely adopted approaches. However, when we apply vanilla DMD to distill a VOR model, we observe a marked degradation in student performance, with many object-induced effects remaining unremoved, as shown in Fig. 3(c).

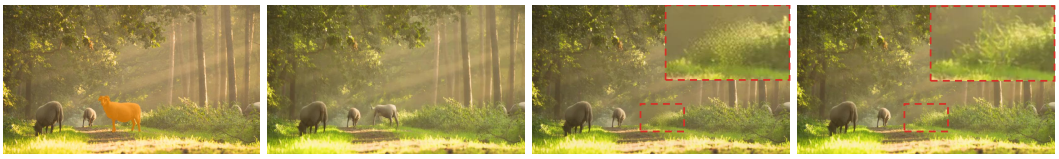
To understand this failure, we revisit DMD through the lens of Decoupled-DMD Liu et al. (2026), which identifies an important implementation asymmetry in vanilla DMD: the real score is constructed by combining conditional and unconditional predictions through CFG, whereas the fake score is obtained from a single conditional prediction. Consequently, the score difference can be algebraically decomposed as

$$s_{\text{real}}^{\text{cfg}} - s_{\text{fake}}^{\text{cond}} = \underbrace{(\omega - 1)\left(s_{\text{real}}^{\text{cond}} - s_{\text{real}}^{\text{uncond}}\right)}_{\text{CFG augmentation}} + \underbrace{s_{\text{real}}^{\text{cond}} - s_{\text{fake}}^{\text{cond}}}_{\text{distribution matching}}, \tag{5}$$

where the superscripts cond and uncond denote scores computed with and without the conditioning input, respectively, and ω is the CFG scale. This decomposition separates DMD into two functionally distinct components: CFG Augmentation (CA) and Distribution Matching (DM). Through extensive ablations, Decoupled-DMD shows that CA provides the principal driving force for converting a multi-step generator into a few-step one. Removing CA leaves only the DM objective and results in substantially degraded few-step generation. In contrast, DM only acts as a regularizer that stabilizes optimization and suppresses artifacts.

This analysis explains why vanilla DMD is less effective for VOR models. VOR is strongly conditioned on the input video and object mask and generally requires neither positive nor negative text prompts. Consequently, many VOR models are inherently CFG-free. For example, EffectErase and Generative Omnimatte Fu et al. (2026); Lee et al. (2025) operate without CFG, while MiniMax-Remover further removes text cross-attention from its DiT Zi et al. (2026). For these models, $\omega = 1$ in Eq. equation 5, causing the CA term to vanish. The distillation process is therefore left without the principal driving signal, which accounts for the degradation observed in our experiments.

A natural question is why CA can drive few-step conversion. Decoupled-DMD does not provide a rigorous explanation, but offers a high-level hypothesis: the CFG-enabled real score represents a



Frames & Mask | w / o Auxiliary Loss | w / Velocity Regression Loss | w / Pixel Reconstruction Loss

Figure 4: Effect of auxiliary supervision. PRL improves removal completeness while preserving finer details than VRL.

more specific and deterministic decision pattern, and CA progressively internalizes this pattern into the student generator. Building on this interpretation, we further hypothesize that CA provides a quality-oriented optimization direction. Specifically, the CFG-enabled real score supplies a stronger prediction, whereas the conditional fake score tracks the relatively weaker student distribution. Their prediction discrepancy therefore provides a direction along which the few-step student can improve.

Based on this hypothesis, we propose *Sparsity-Driven Distillation* (SDD), a strategy that creates a score discrepancy without relying on CFG. Specifically, the frozen real score model retains the original dense-attention DiT, whereas the online fake score model uses a sparse-attention DiT. Sparse attention closely approximates dense attention but retains a small, structured prediction discrepancy, which provides the optimization direction missing from vanilla DMD used in CFG-free models. The fake score model is continuously updated on student-generated samples to track the evolving sparse student distribution, while the frozen dense real score model provides a fixed, higher-fidelity target.

Analogous to Eq. equation 5, the corresponding descent direction can be decomposed as

$$s_{\text{real}}^{\text{dense}} - s_{\text{fake}}^{\text{sparse}} = \underbrace{s_{\text{real}}^{\text{dense}} - s_{\text{real}}^{\text{sparse}}}_{\text{sparsity augmentation}} + \underbrace{s_{\text{real}}^{\text{sparse}} - s_{\text{fake}}^{\text{sparse}}}_{\text{distribution matching}}. \tag{6}$$

Here, $s_{\text{real}}^{\text{dense}}$ and $s_{\text{real}}^{\text{sparse}}$ denote the scores predicted by the real score model using dense and sparse attention, respectively, while $s_{\text{fake}}^{\text{sparse}}$ denotes the score predicted by the fake score model using sparse attention. The first term isolates the prediction discrepancy induced by dense and sparse attention, whereas the second performs distribution matching under a shared sparse-attention operator. This discrepancy supplies the missing driving signal for few-step adaptation. As shown in Figure 3(a), SDD provides a larger score discrepancy during distillation, consistent with the proposed architectural asymmetry. Thus, SDD demonstrates a stable stronger performance during distllation, as shown in Figure 3(b). The visual comparison in Fig. 3(c) further illustrates cleaner removal by SDD.

3.3 Pixel Reconstruction Loss

Relying only on distribution-level supervision during distillation risks weakening the student's editing ability. DMD-based methods therefore add LPIPS or GAN losses Yin et al. (2024b;a) to incorporate supervision from real data. However, these objectives mainly provide high-level perceptual guidance. Because effect-aware VOR requires removing both an object and its associated effects, we argue that auxiliary supervision from ground-truth videos should directly constrain the student's removal behavior.

A straightforward solution is to directly supervise the student's velocity predictions in latent space. Under the flow-matching Lipman et al. (2022) formulation, the noisy latent is constructed as

$$\mathbf{x}_t = (1 - \sigma_t)\mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{7}$$

where $\mathbf{x}_0$ is the ground-truth video latent and σ_t denotes the noise level. The corresponding ground-truth velocity is $\mathbf{v}^* = \boldsymbol{\epsilon} - \mathbf{x}_0$. We therefore define Velocity Regression Loss (VRL) as

$$\mathcal{L}_{\text{VRL}} = \mathbb{E}_{t,\boldsymbol{\epsilon}} \left[\| \mathbf{v}_\theta(\mathbf{x}_t, t, c) - (\boldsymbol{\epsilon} - \mathbf{x}_0) \|_2^2 \right], \tag{8}$$

where $\mathbf{v}_\theta$ is the velocity predicted by the student and c denotes the conditioning input. However, we found that VRL produces undesirable blur in the edited regions, as shown in the third column

of Fig. 4. We attribute this behavior to pointwise velocity regression in latent space, which encourages the student to approximate an average solution under prediction uncertainty and consequently reduces fine-detail fidelity and the naturalness of the generated videos.

To address this limitation, we introduce Pixel Reconstruction Loss (PRL), which directly anchors the student to the desired removal result in RGB space. Given the predicted velocity, we first recover the clean latent as

$$\widehat{\mathbf{x}}_0 = \mathbf{x}_t - \sigma_t \mathbf{v}_\theta(\mathbf{x}_t, t, c). \tag{9}$$

The reconstructed latent is then decoded by a frozen VAE decoder D_ψ:

$$\widehat{\mathbf{y}} = D_\psi(\widehat{\mathbf{x}}_0)\,, \qquad \mathcal{L}_{\mathrm{PRL}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \|\widehat{\mathbf{y}}_i - \mathbf{y}_i\|_2^2\,, \tag{10}$$

where $\mathbf{y}$ is the ground-truth RGB video and $\mathcal{I}$ contains all spatial locations, color channels, and video frames. Although the decoder parameters ψ remain frozen, gradients are propagated through the decoder to optimize the student.

Unlike VRL, PRL measures errors according to their visible consequences in RGB space. It therefore provides direct supervision for sample-specific removal while encouraging the reconstruction of sharp and natural image structures, as demonstrated in the final column of Fig. 4. The lower NIQE Mittal et al. (2012) achieved by PRL further indicates that its predictions more closely follow the statistics of natural images. The complete student objective is $\mathcal{L}_{\mathrm{student}} = \mathcal{L}_{\mathrm{SDD}} + \mathcal{L}_{\mathrm{PRL}}$.

4 Experiment

4.1 Implementation Details

We train FastEraser on VOR-Train Fu et al. (2026), using 81 frames at a resolution of 832×480. The real score model, fake score model, and student are initialized from the official EffectErase checkpoint. We use separate AdamW optimizers for the student and fake score model, both with a learning rate of 2×10^{-5}. Following the DMD update schedule, the fake score model is updated at every iteration, whereas the student is updated once every five iterations. Training lasts for 1,000 student updates, corresponding to 5,000 fake-score updates. All experiments are conducted on NVIDIA A100 GPUs.

4.2 Benchmarks and Metrics

Benchmarks. We evaluate FastEraser on four effect-aware VOR benchmarks: VOR-Eval (43 videos) and VOR-Wild (195) Fu et al. (2026), as well as PROVE-M (80) and PROVE-H (100) Li et al. (2026). VOR-Eval and PROVE-M provide paired ground-truth clean videos, whereas VOR-Wild and PROVE-H consist of real-world videos without ground truth.

Metrics. Across all four benchmarks, we report RC-S, RC-T Li et al. (2026), and Gemini-3.5-Flash scores. RC-S measures spatial removal coherence (higher is better), while RC-T measures temporal consistency (lower is better). Gemini-3.5-Flash serves as a vision-language-model judge of removal completeness and visible artifacts on the middle frame of each video (higher is better). We use the evaluation prompt from EffectErase Fu et al. (2026), as shown in Section C of the supplementary material. For VOR-Eval and PROVE-M, which provide ground-truth clean videos, we additionally report PSNR and SSIM Wang et al. (2004) to measure reconstruction fidelity. In the ablation study, we further report mask-guided NIQE Mittal et al. (2012) to focus on the naturalness of the reconstructed regions.

4.3 Quantitative Results

We compare FastEraser with existing methods on four challenging effect-aware VOR benchmarks, as reported in Table 1. FastEraser performs strongly across full-reference and no-reference metrics, as well as vision-language model evaluations. It also achieves the shortest DiT denoising time in Table 2. This efficiency stems from SDD, which jointly trains a student for few-step inference and sparse attention. Reducing the denoising steps yields an approximately $10\times$ acceleration, while

Table 1: Quantitative comparison on four VOR benchmarks. Best and second-best results are highlighted in light red and light yellow for each metric on each benchmark, respectively.

Method	PROVE-M					PROVE-H		
	Gemini↑	RC-S↑	RC-T↓	PSNR↑	SSIM↑	Gemini↑	RC-S↑	RC-T↓
MiniMax-Remover	5.10	0.477	0.462	21.28	0.8520	5.02	0.452	0.349
YOSE	5.05	0.476	0.545	21.54	0.8452	5.13	0.469	0.351
ROSE	5.76	0.490	0.668	24.75	0.8633	5.43	0.463	0.461
VOID	5.21	0.482	0.488	22.33	0.8590	5.09	0.471	0.290
Generative Omnimatte	5.63	0.511	0.334	24.38	0.8690	5.40	0.493	0.256
EffectErase	6.38	0.518	0.245	26.81	0.8900	5.99	0.509	0.217
FastEraser (Ours)	6.39	0.523	0.268	26.89	0.8915	6.12	0.511	0.225
Method	VOR-Eval					VOR-Wild		
	Gemini↑	RC-S↑	RC-T↓	PSNR↑	SSIM↑	Gemini↑	RC-S↑	RC-T↓
MiniMax-Remover	5.44	0.449	0.368	21.00	0.7561	5.53	0.429	0.336
YOSE	5.49	0.454	0.466	21.09	0.7569	5.51	0.438	0.377
ROSE	5.72	0.442	0.575	21.85	0.7365	6.02	0.449	0.500
VOID	5.67	0.474	0.346	21.68	0.7597	5.72	0.466	0.272
Generative Omnimatte	5.65	0.489	0.291	21.41	0.7553	5.83	0.484	0.260
EffectErase	6.12	0.513	0.240	22.42	0.7462	5.97	0.473	0.243
FastEraser (Ours)	5.98	0.516	0.252	22.69	0.7410	6.03	0.489	0.262

Table 2: DiT inference time and parameter counts for FastEraser and competing methods.

Method	Steps	Params (B)	Runtime (s)
ROSE	50	1.6	173.5
VOID	50	5.6	107.5
EffectErase	50	1.6	86.0
Generative Omnimatte	50	1.6	80.0
MiniMax-Remover	12	1.1	20.3
YOSE	12	1.1	7.3
FastEraser (Ours)	4	1.6	4.6

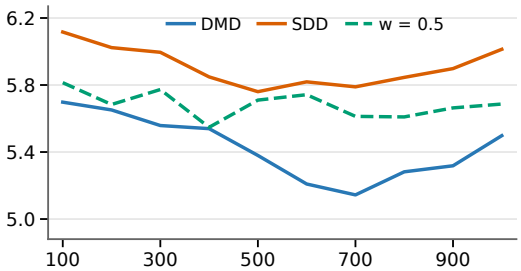


Figure 5: Performance trajectories of SDD and DMD during distillation.

sparse attention provides a further roughly 2× speedup. Together, they deliver a DiT denoising speedup approaching 20×.

4.4 Ablation Study

We first compare SDD with other distillation strategies, including the consistency-based LCM Luo et al. (2023) and the distribution-based DMD Yin et al. (2024b). As shown in Table 3, SDD incurs a smaller performance drop during distillation, supporting the effectiveness of using asymmetric architectures for the real and fake score models.

To explore the contribution of the asymmetric sparsity score models, we perform an interesting experiment. We train a dual-fake variant that maintains both dense and sparse fake score models. We decompose the score signal into a dense-fake DMD term, $\Delta_{\text{DMD}} = s_{\text{fake}}^{\text{dense}} - s_{\text{real}}^{\text{dense}}$, and a sparse-minus-dense fake-score residual, $\Delta_{\text{sp}} = s_{\text{fake}}^{\text{sparse}} - s_{\text{fake}}^{\text{dense}}$. The student update uses $\Delta_{\text{DMD}} + w\Delta_{\text{sp}}$: $w = 0$ gives the DMD score direction, $w = 1$ gives the SDD score direction, and $w = 0.5$ averages the dense and sparse fake scores before subtracting the real score. As shown in Figure 5, the $w = 0.5$ variant generally improves performance over DMD but remains below SDD, suggesting that the full sparse-fake residual provides a more useful learning signal for the few-step student.

We then evaluate PRL in Table 4. Although VRL provides strong reconstruction fidelity, it substantially compromises the naturalness and visual coherence of the generated videos. In contrast, PRL applies supervision in RGB space, compensating for the lack of direct supervision from ground truth

Table 3: Ablation of distillation frameworks.

Setting	Gemini↑	RC-S↑	RC-T↓
Teacher (50 steps)	6.02	0.513	0.240
Teacher (4 steps)	6.07	0.500	0.260
LCM	6.05	0.495	0.265
DMD	5.77	0.505	0.251
SDD	6.19	0.518	0.225

Table 4: Ablation of auxiliary supervision.

Setting	PSNR↑	SSIM↑	NIQE↓
Teacher (50 steps)	22.42	0.7462	9.09
Teacher (4 steps)	23.00	0.7629	9.74
SDD	21.34	0.7023	8.71
SDD + VRL	23.72	0.7643	10.55
SDD + PRL	22.52	0.7410	8.75

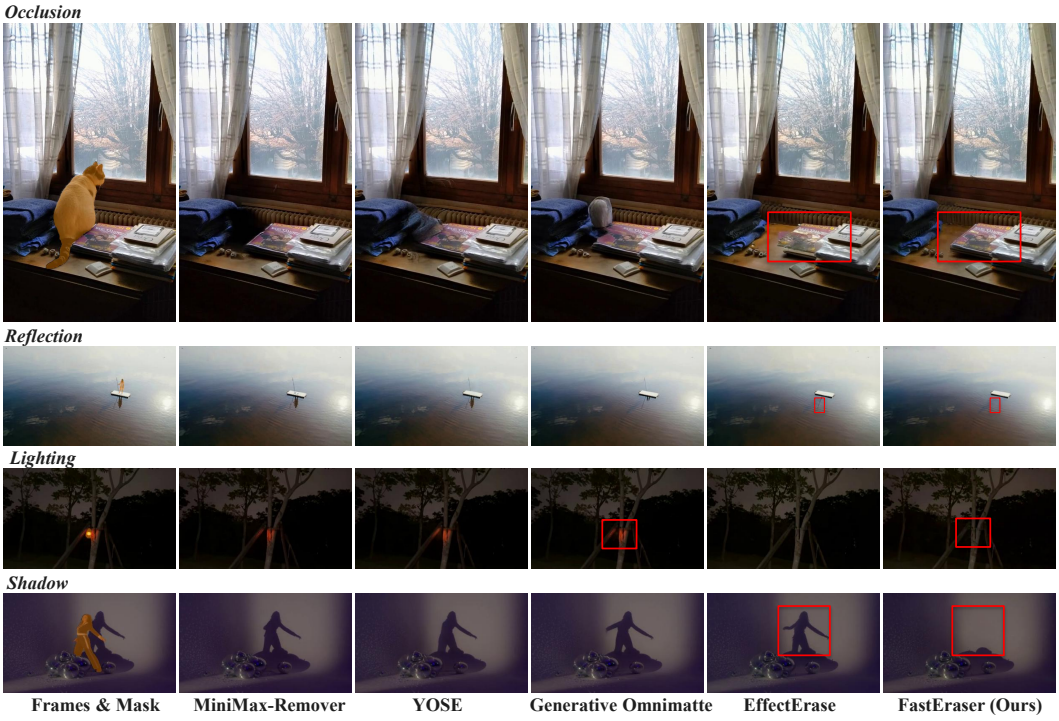


Figure 6: Qualitative comparison of effect-aware video object removal. Red boxes highlight differences in reconstruction quality and residual effects.

videos while improving perceptual naturalness. Additional comparisons with the LPIPS and GAN auxiliary losses used in prior DMD Yin et al. (2024b;a) are reported in Section A.2.

4.5 Visualization

We present representative visual comparisons between FastEraser and competing methods in Figure 6. The first row illustrates conventional reconstruction of occluded regions, while the second through fourth rows show diverse object-associated effects, including reflections, lighting, and shadows. A case involving physical interactions is also shown in Figure 3(c). We also provide visual results as MP4 videos in the supplementary material.

5 Conclusion

We present FastEraser, an efficient framework for effect-aware video object removal. Sparsity-Driven Distillation introduces dense–sparse attention asymmetry to guide CFG-free models toward few-step inference, while preserving the ability to remove objects and their associated effects beyond the input mask. Pixel Reconstruction Loss complements distillation with direct supervision in RGB space, reinforcing removal capability and mitigating the over-smoothing associated with latent regression. Experiments on four benchmarks demonstrate that FastEraser achieves competitive removal quality and reconstruction fidelity with only four denoising steps, substantially reducing inference cost relative to multi-step baselines.

AI USE STATEMENT

We used generative AI tools to aid and polish the writing of this paper, including translation and language editing, and to draft selected passages. During the experimental work, these tools assisted us in implementing parts of the experimental code, running experiments, and computing evaluation metrics. The authors reviewed the AI-assisted material against the experimental records and take responsibility for the final manuscript and its claims.

REFERENCES

Yang Fu, Yike Zheng, Ziyun Dai, and Henghui Ding. Effecterase: Joint video object removal and insertion for high-quality effect erasing. *arXiv preprint arXiv:2603.19224*, 2026.

Youping Gu, Xiaolong Li, Yuhao Hu, Minqi Chen, and Bohan Zhuang. Blade: Block-sparse attention meets step distillation for efficient video generation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/5bcb807ae43ad0851a6ba6162a866404-Paper-Conference.pdf.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Yao-Chih Lee, Erika Lu, Sarah Rumbley, Michal Geyer, Jia-Bin Huang, Tali Dekel, and Forrester Cole. Generative omnimatte: Learning to decompose video into layers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12522–12532. IEEE, 2025.

Fuhao Li, Shaofeng You, Jiagao Hu, Yu Liu, Yuxuan Chen, Zepeng Wang, Fei Wang, Daiguo Zhou, and Jian Luan. Prove: A perceptual removal coherence benchmark for visual media. In *Proceedings of the 34th ACM International Conference on Multimedia*, 2026.

Xiaowen Li, Haolan Xue, Peiran Ren, and Liefeng Bo. Diffueraser: A diffusion model for video inpainting. *arXiv preprint arXiv:2501.10018*, 2025.

Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17541–17550. IEEE, 2022.

Geng Lin, Chen Gao, Jia-Bin Huang, Changil Kim, Yipeng Wang, Matthias Zwicker, and Ayush Saraf. Omnimatterf: Robust omnimatte with 3d background modeling. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23414–23423. IEEE, 2023.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Dongyang Liu, Peng Gao, David Liu, Ruoyi Du, Zhen Li, Qilong Wu, Xin Jin, Sihan Cao, Shifeng Zhang, Steven HOI, and Hongsheng Li. Decoupled dmd: Cfg augmentation as the spear, distribution matching as the shield. In *International Conference on Learning Representations*, volume 2026, pp. 140643–140666, 2026.

Enzhe Lu, Zhejun Jiang, Jingyuan Liu, Yulun Du, Tao Jiang, Chao Hong, Shaowei Liu, Weiran He, Enming Yuan, Yuzhi Wang, et al. Moba: Mixture of block attention for long-context llms. *Advances in Neural Information Processing Systems*, 38:17790–17815, 2026.

Erika Lu, Forrester Cole, Tali Dekel, Andrew Zisserman, William T Freeman, and Michael Rubinstein. Omnimatte: Associating objects and their effects in video. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4505–4513. IEEE, 2021.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Chenxuan Miao, Yutong Feng, Jianshu Zeng, Zixiang Gao, Hantang Liu, Yunfeng Yan, Donglian Qi, Xi Chen, Bin Wang, and Hengshuang Zhao. Rose: Remove objects with side effects in videos. *Advances in Neural Information Processing Systems*, 38:149140–149162, 2026.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

Saman Motamed, William Harvey, Benjamin Klein, Luc Van Gool, Zhuoning Yuan, and Ta-Ying Cheng. Void: Video object and interaction deletion. *arXiv preprint arXiv:2604.02296*, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Mohammed Suhail, Erika Lu, Zhengqi Li, Noah Snavely, Leonid Sigal, and Forrester Cole. Omnimatte3d: Associating objects and their effects in unconstrained monocular video. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 630–639. IEEE, 2023.

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

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Chenyang Wu, Lina Lei, Fan Li, Chunle Guo, Dehong Kong, Xinran Qin, Zhixin Wang, Mingming Cheng, and Chongyi Li. Yose: You only select essential tokens for efficient dit-based video object removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32926–32935, 2026a.

Jianzong Wu, Liang Hou, Haotian Yang, Ye Tian, Pengfei Wan, Di ZHANG, and Yunhai Tong. Vmoba: Mixture-of-block attention for video diffusion models. In *International Conference on Learning Representations*, volume 2026, pp. 132534–132548, 2026b.

Haocheng Xi, Shuo Yang, Yilong Zhao, Chenfeng Xu, Muyang Li, Xiuyu Li, Yujun Lin, Han Cai, Jintao Zhang, Dacheng Li, et al. Sparse videogen: Accelerating video diffusion transformers with spatial-temporal sparsity. *arXiv preprint arXiv:2502.01776*, 2025.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. IEEE, 2024b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974. IEEE, 2025.

Jintao Zhang, Haoxu Wang, Kai Jiang, Shuo Yang, Kaiwen Zheng, Haocheng Xi, Ziteng Wang, Hongzhou Zhu, Min Zhao, Ion Stoica, et al. Sla: Beyond sparsity in diffusion transformers via fine-tunable sparse–linear attention. In *International Conference on Learning Representations*, volume 2026, pp. 138288–138305, 2026a.

Peiyuan Zhang, Yongqi Chen, Haofeng Huang, Will Lin, Zhengzhong Liu, Ion Stoica, Eric Xing, and Hao Zhang. Faster video diffusion with trainable sparse attention. *Advances in Neural Information Processing Systems*, 38:152509–152534, 2026b.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. URL `https://richzhang.github.io/PerceptualSimilarity/`.

Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10477–10486, 2023.

Bojia Zi, Weixuan Peng, Xianbiao Qi, Jianan Wang, Shihao Zhao, Rong Xiao, and Kam-Fai Wong. Minimax-remover: Taming bad noise helps video object removal. *Advances in Neural Information Processing Systems*, 38:75518–75547, 2026.

A Additional Ablation Studies

A.1 Joint versus Sequential Optimization

As illustrated in Figure 2, SDD produces a student that combines few-step inference with sparse attention in a single training process. The compatibility of sparse attention with step distillation was explored in VSA Zhang et al. (2026b), while BLADE Gu et al. (2026) explicitly jointly optimizes both forms of acceleration for video generation. BLADE motivates joint training in part through its data-free design: a subsequent sparsity fine-tuning stage would introduce external video data and potentially affect model quality through dataset bias. We examine whether joint optimization remains beneficial when training videos are available, as in our effect-aware VOR setting.

Table 5 compares joint optimization with both sequential orders. All sparse students use an SLA top-k ratio of 10%, four-step inference. For sparsification followed by distillation, we fine-tune the 50-step EffectErase model with sparse attention, then initialize the student and both score models from this checkpoint and run step distillation. For step distillation followed by sparsification, we distill a dense student and then fine-tune it with sparse attention . Joint optimization trains the sparse student for 1,000 updates with a dense real score model and a sparse fake score model.

Table 5: Comparison of joint and sequential Optimization.

Method	RC-S ↑	RC-T ↓	PSNR ↑	NIQE ↓	PIQE ↓
EffectErase (50 steps)	0.513	0.240	22.42	9.09	52.56
Sparsification → Step Distillation	0.462	0.230	20.22	8.96	52.16
Step Distillation → Sparsification	0.507	0.266	23.37	9.63	56.45
Joint Optimization	0.518	0.225	21.34	8.71	51.49

Among the four-step sparse students, joint optimization achieves the highest RC-S and the lowest RC-T, NIQE, and PIQE. Sparsification followed by step distillation reduces RC-S, possibly because the real score model inherits a weaker reference distribution from the sparse checkpoint. The reverse order degrades naturalness after data-supervised sparse fine-tuning, consistent with drifting away from the high-quality modes found during distillation. Using the distilled model to supervise that second stage would instead constitute another distillation stage. These results show that joint optimization remains advantageous even when task-specific training data are available.

BLADE does not isolate the effect of using a dense versus sparse fake score model. Its text-conditioned video generation setting admits CFG, whereas many effect-aware VOR models are CFG-free. This makes the fake score model's attention pattern a consequential design choice in our setting, which we study in SDD.

A.2 Auxiliary Supervision

Table 6 compares SDD alone with variants using GAN, LPIPS Zhang et al. (2018), and PRL as auxiliary losses. PRL achieves the highest RC-S, suggesting that effect-aware VOR benefits from auxiliary supervision that directly constrains the student's editing behavior.

Table 6: Auxiliary-supervision ablation.

Auxiliary loss	RC-S ↑	RC-T ↓	PSNR ↑	SSIM ↑	NIQE ↓
None	0.5177	0.2909	25.57	0.8785	10.93
GAN	0.5037	0.3059	25.73	0.8638	10.16
LPIPS	0.5189	0.2887	24.86	0.8789	10.51
PRL	0.5233	0.2681	26.89	0.8915	10.68

A.3 ODE-BASED INITIALIZATION

Table 7 examines whether ODE-based initialization benefits distillation. Following the ODE-trajectory initialization in CausVid Yin et al. (2025), we compare direct EffectErase initialization with initializing the student and fake score models from a checkpoint trained for 2,000 ODE updates with a noise-schedule shift of 5. The frozen real score model retains the original teacher weights in both cases. ODE initialization yields only modest, mixed changes while requiring an additional training stage; we therefore use direct EffectErase initialization.

Table 7: ODE-initialization ablation.

Initialization	RC-S ↑	RC-T ↓	PSNR ↑	SSIM ↑	NIQE ↓
EffectErase (Ours)	0.5233	0.2681	26.89	0.8915	10.68
ODE	0.5215	0.2651	27.20	0.8913	10.83

A.4 EFFECT OF ATTENTION SPARSITY

We further evaluate the effect of attention sparsity on SDD in Table 8, using key-block retention ratios of $k \in \{5\%, 10\%, 20\%\}$. For example, $k = 10\%$ means that each query block attends to the most relevant 10% of key blocks. At $k = 5\%$, removal performance declines, consistent with overly aggressive sparsification limiting the model's representational capacity. Increasing k to 20% raises attention computation but does not improve removal performance. One possible explanation is that weaker sparsification reduces the dense–sparse score discrepancy on which SDD relies, making the optimization more similar to vanilla DMD. Among the tested settings, $k = 10\%$ achieves the best removal performance.

Table 8: Attention sparsity ablation.

k	Gemini↑	RC-S↑	RC-T↓	PSNR↑	SSIM↑
5%	5.51	0.489	0.295	20.63	0.6986
10%	6.19	0.518	0.225	21.34	0.7023
20%	5.77	0.494	0.240	20.92	0.7035

B LIMITATIONS

FastEraser reduces the teacher's 50 denoising steps to four but does not yet achieve single-step inference. We hypothesize that removing both the target object and its associated effects while reconstructing a coherent background leads to a complex denoising trajectory that is difficult to compress into one step. Achieving single-step inference without compromising removal quality remains challenging.

Our acceleration targets DiT denoising and does not optimize VAE encoding or decoding. As denoising latency decreases, the unchanged VAE cost accounts for a larger fraction of end-to-end inference time, limiting the practical speedup. Improving VAE efficiency is therefore an important direction for future work.

C GEMINI-BASED EVALUATION PROTOCOL AND PROMPT

We use Gemini-3.5-Flash to assess perceptual removal quality. For each test case, we select the temporally aligned middle frame from the input and output videos. The first image presented to Gemini is the input frame with the target mask overlaid in orange, and the second image is the corresponding frame after removal.

```
You are an expert evaluator for video object removal quality.
You will be shown 2 images:
- The FIRST image is the reference frame before removal, with an
ORANGE MASK marking the object/person to be removed
- The SECOND image is the aligned middle frame of the video after
removal

Please evaluate the removal quality based on the following criteria
(0-10 scale, BE STRICT):
1. Completeness: Is the target marked by the orange mask completely
removed?
2. Artifacts: Are there any blur, distortion, or other artifacts in
the removal area?
3. Secondary Effects: Are shadows, reflections, lighting effects,
etc. eliminated?
4. Background Quality: Does the inpainted background blend
naturally?

Scoring Guidelines:
- 10: Perfect - target completely invisible, zero artifacts, no
secondary effects
- 9: Nearly perfect - only extremely minor edge artifacts
- 8: Target completely removed, slight edge artifacts, no secondary
effects
- 7: Target removed with visible artifacts but good overall quality
- 6: Target removed, noticeable artifacts OR minor secondary effects
- 5: Target removed, obvious artifacts AND secondary effects
- 4: Mostly removed, significant artifacts OR secondary effects
- 3: Removed but severe artifacts
- 2: Not fully removed OR extremely severe artifacts
- 1: Barely removed
- 0: Complete failure

CRITICAL RULES:
- Score 10 is EXTREMELY RARE
- ANY visible secondary effects (shadow/light/reflection) = score 6
or below
- ANY noticeable artifacts = score 6 or below
- Most results should fall in 4-7 range

Please respond in the following format:
<think>
[Brief analysis: what is marked by the orange mask in reference
image, completeness of removal, artifacts, secondary effects,
background quality]
</think>
<result>[score 0-10]</result>
```

D Ethics Statement

Video object removal supports legitimate editing and post-production, but it can also be used to alter footage in misleading ways. Edited videos should be identified as such when their authenticity matters. Our study uses existing video object removal datasets and does not involve a new human-subject study. We encourage users of this technology to respect the privacy and usage rights associated with source videos.

E Reproducibility Statement

The model configurations and objectives are described in Sections 3.2 and 3.3. The Implementation Details and Benchmarks and Metrics subsections specify the training schedule, input resolution, evaluation datasets, and metrics. We plan to publicly release the code and checkpoints in the future.