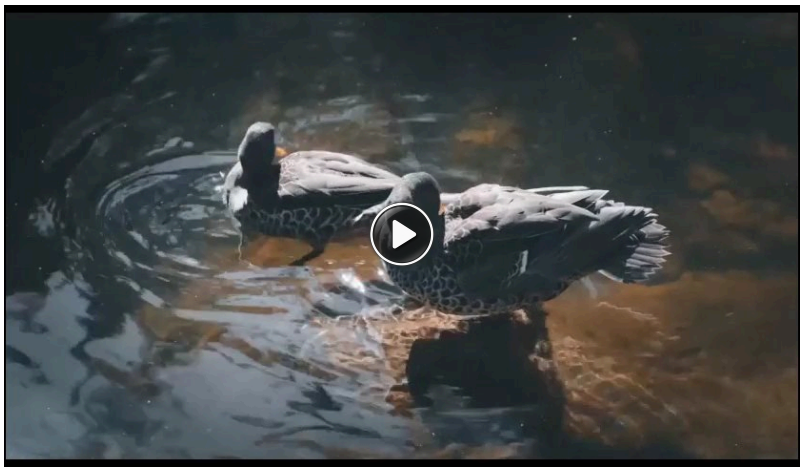
Erasing Objects and People From Video with AI

Originally published at https://www.unite.ai/erasing-objects-and-people-from-video-with-ai/

No, the kid doesn't stay in the picture, if AI has anything to do with it.

Removing people and objects from images and video is a popular sub-strand of research in VFX-centric AI literature, with a growing number of dedicated datasets and frameworks addressing the challenge. The latest of these, from the Institute of Big Data at China's Fudan University, is *EffectErase*, an 'effect-aware' video object removal system that, the authors contend, improves notably on the state of the art in tests:



CLICK TO PLAY VIDEO ON SITE

Assembled from material at the project website, examples of the EffectErase method (please note that while we provide a link, the source site contains so many hi-res and non-optimized autoplay videos that it may affect the stability of your web browser. TThe accompanying YouTube video is an easier and fuller reference, and is embedded at the end of this article). Source

The new work involved the creation/curation of a semi-novel dataset comprising nearly 350 original real-world and also synthesized scenes (making use of public repositories*), either captured with dedicated equipment or sourced and re-purposed into a workflow built around the open source Blender 3D framework.

The hybrid Video Object Removal (VOR) dataset forms the basis for the *EffectErase* application itself, which is built over the Wan2.1 video-generation system. The system also defines two new related benchmarks: *VOR Eval* and *VOR Wild* – respectively, for samples with and without ground truth.

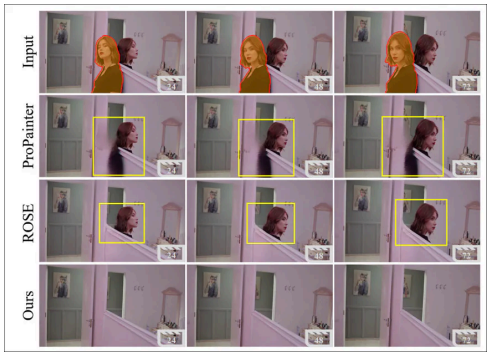
(Though the paper has an accompanying project site, it's rather over-burdened with multiple hi-res videos, and hard to load; so please refer to the excerpts I have curated in the embedded video above, if you find the project site difficult to use)

Dataset	Source	Dynamic Camera	Dynamic Object	Dynamic Background	Scene Types	Object Classes	Image Pairs	Video Pairs	Average Duration (s)	Total Hours
RORD [31]	Real	✗	✓	✗	24	76	516.7K	3,106	–	5.98
Video4Removal [38]	Real	✗	✓	✗	6	–	134.3K	–	–	1.55
ROSE [26]	Synth.	✓	✗	✗	25	102	1,501.0K	16,678	6.00	27.79
VOR (Ours)	Real + Synth.	✓	✓	✓	**67**	**366**	**12,556.8K**	**60,000**	**8.72**	**145.33**

A comparison of quantities across comparable prior datasets, with respect to the new offering. Source

The researchers claim that their approach yields state-of-the-art performance, both in quantitative metrics and in qualitative results as adjudicated through a human study.

They note that prior works have not always succeeded in removing adjunct effects of an object, such as shadows and reflections, and that their dataset has been carefully created to amend this shortcoming:

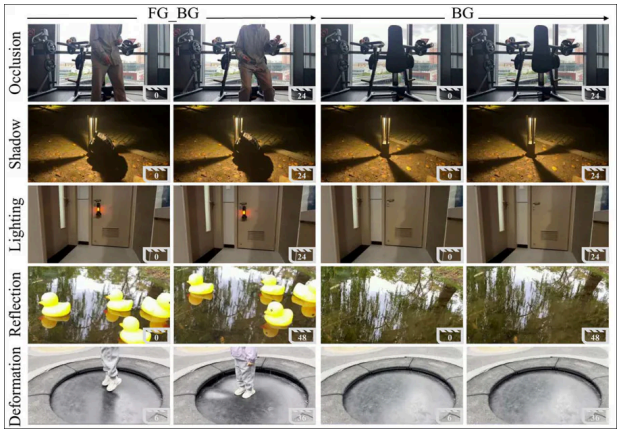


Examples of prior approaches' failure to look beyond the object sought for removal, to secondary indications, such as reflections and shadows.

The new paper is titled *EffectErase: Joint Video Object Removal and Insertion for High-Quality Effect Erasing*, and comes from four researchers at Fudan University's College of Computer Science and Artificial Intelligence.

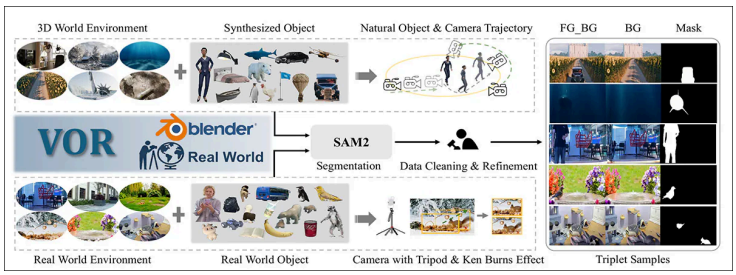
Method

The hybrid VOR dataset was designed to encompass a wide enough gamut of scenarios as to cover all the implications of attempting to remove a person or object from video:



Paired frames from the VOR dataset illustrate how object removal must extend beyond the visible subject to its induced effects, with examples showing occlusion, shadow, lighting shifts, reflections, and physical deformation, each presented as input (object present) alongside the corresponding clean background after removal. For further examples, see the accompanying YouTube video embedded at the end of this article.

The five representative types of 'interference' to be addressed are defined by the authors as *occlusion*, including various types of glass and smoke occlusion; *shadows*; *lighting* (for instance, when an object to be removed creates or alters the path of light); *reflection*; and *deformation* (for instance, the imprint of a user on a cushion, which should not survive the person's removal).



Dataset construction pipeline for VOR, combining Blender-generated synthetic scenes with real-world captures, where synthetic data is built from curated 3D envi objects, and camera trajectories, and real footage recorded across diverse scenes, augmented with Ken Burns motion. SAM2 segmentation and manual refineme produce aligned foreground and background video triplets with corresponding masks.

For the real-world original data, the researchers used fixed cameras to record 'with' and 'without' scenes covering a wide range of environments, the time of day, and weather conditions.

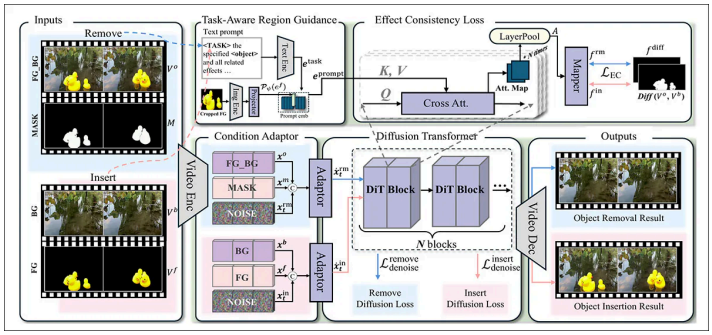
For the synthesized data, multiple viewpoints were rendered, and multi-object scenarios created, featuring deliberately complex and challenging types of camera movement, such as might occur in real-world footage; and the researchers observe that this approach is more sophisticated and effortful than that used for the otherwise similar *Remove Objects with Side Effects in Videos* (ROSE) dataset.

To increase motion diversity, the Ken Burns effect was applied to camera-captured pairs, adding controlled pans, zooms, and slight handheld movement under fourteen predefined rules, with five motion patterns sampled per pair while keeping the crop inside the original frame.

Scale and diversity were further expanded by combining synthetic objects with multiple camera setups, Masks were generated by placing manual point prompts on key frames, propagating segmentation with Segment Anything 2 (SAM2), cleaning and refining results, and assembling validated foreground, background, and mask triplets for training.

The final collection runs to 145 hours of video across 60,000 paired videos, real and synthetic, covering 366 object classes in 443 scenes.

The EffectErase network itself ingests material via a Variational Auto-Encoder (VAE†), with the latent denoising handled by Wan2.1. Over this backbone, EffectErase operates *Removal-Insertion Joint Learning*, which trains both tasks together on the same regions; *Task-Aware Region Guidance* (TARG), which uses object and task tokens with cross-attention to model spatiotemporal links between objects and their effects and allow task switching; and *Effect Consistency Loss*, which aligns effect regions aligned across removal and insertion tasks:



Schema for the EffectErase framework. During training, paired videos are encoded into a shared latent space, fused with noise, and processed by a diffusion trans guided by task-aware cross attention, while an effect consistency loss aligns removal and insertion regions so both tasks focus on the same area.

In themselves, the removal and insertion processes are trained together, using a shared diffusion backbone, so the model learns to focus on the same affected regions and structural cues.

Videos with objects, background-only videos, and masks, are first encoded into a latent space; noise is then added for diffusion training, and the model learns to recover clean representations under task-specific guidance. A lightweight adapter then fuses the noisy features with removal or insertion conditions, allowing both tasks to share supervision, while remaining controllable.

Task-Aware Region Guidance creates a task-specific signal by combining language tokens with visual features extracted from the foreground object, using CLIP, replacing a generic object token with an embedding derived from the actual image content. This fused representation is injected into the backbone through cross attention, allowing the model to track how an object and its visual effects evolve over space and time, while enabling flexible switching between removal and insertion.

Effect Consistency Loss forces removal and insertion processes to focus on the same changed areas, since both tasks deal with the same object and its visual effects. Attention maps from each branch are then combined into soft region maps, and aligned with a difference map computed from the object and background videos, so that subtle changes like lighting and shadows are preserved. This extra loss helps insertion guide removal and keeps both tasks consistent.

Data and Tests

The researchers tested their approach against various inpainting, video inpainting, and object removal methods: *OmniPaint*; *ObjectClear*; *VACE*; *DiffuEraser*; *ProPainter*; *ROSE*; and *MiniMax-Remover*.

Wan2.1 was fine-tuned with LoRA†† using the VOR dataset at a resolution of 832x480px. 81 consecutive frames (the effective limit for WAN, beyond which errors tend to occur) were randomly sampled for training, which took place for 129,000 iterations at a batch size of 8, on eight H100 GPUs, each with 80GB of VRAM. The learning rate was set to 1×10^{2}, and the LoRA rank to 256.

The ROSE-Benchmark synthetic collection was the only external dataset tested; the other two were *VOR-Eval*, the *VOR* dataset test split; and VOR-Wild, a test set comprising 195 real videos scraped from the internet, featuring 'dynamic objects'.

Metrics used were Peak Signal-to-Noise Ratio (PSNR); Structural Similarity Index (SSIM); Learned Perceptual Image Patch Similarity (LPIPS); and Fréchet Video Distance (FVD). A user study of 195 generated videos from VOR-Wild was also considered, with averaged ratings from 20 volunteers taken into account.

Additionally the authors devised *QScore*, a metric leveraging the Qwen-VL multimodal model, in order to evaluate the quality of object-removed video output, in terms of remnant artifacts or missed environmental removals, such as shadows and lighting effects:

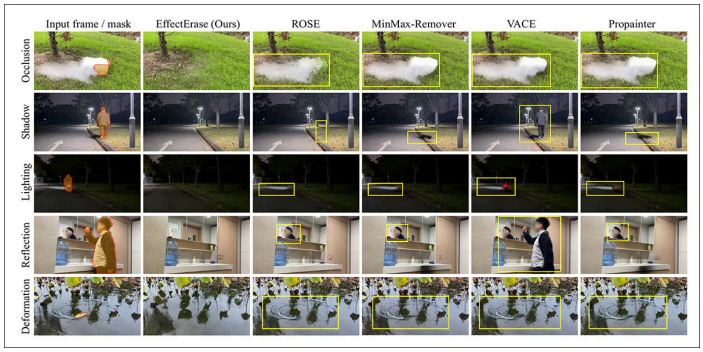
Method	ROSE-Benchmark (with GT)				VOR-Eval (with GT)				VOR-Wild (without GT)	
	PSNR↑	SSIM↑	LPIPS↓	FVD↓	PSNR↑	SSIM↑	LPIPS↓	FVD↓	QScore↑	User↑
ObjectClear [46]	29.535	0.920	0.076	742.829	22.583	0.787	0.190	1391.858	8.979	4.75
OmniPaint [43]	27.569	0.910	0.085	809.645	21.511	0.781	0.201	1439.867	8.942	4.38
Propainter [47]	27.200	0.915	0.095	171.020	21.975	0.800	0.225	589.012	8.860	4.88
DiffuEraser [21]	26.502	0.898	0.128	167.483	21.946	0.802	0.214	559.497	9.113	5.50
VACE [17]	20.805	0.694	0.174	254.117	17.677	0.591	0.294	806.476	8.229	1.50
MinMax-Remover [48]	26.770	0.905	0.099	137.840	21.963	0.802	0.217	539.427	8.984	5.90
ROSE [26]	31.122	0.917	0.077	72.177	22.966	0.792	0.203	383.084	9.240	6.38
EffectErase (Ours)	**32.161**	**0.948**	**0.039**	**55.578**	**23.750**	**0.806**	**0.170**	**342.871**	**9.280**	**7.20**

Quantitative comparison on ROSE and VOR benchmarks, with best and second-best results shown in bold and underlined, respectively.

Regarding these results, the authors note:

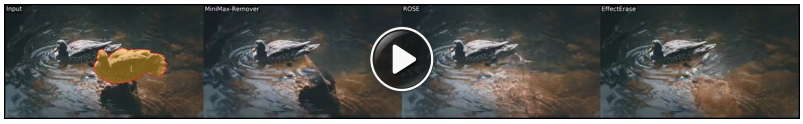
> *'[Current] image inpainting methods operate on individual frames using 2D models without temporal modeling, and therefore fail to maintain temporal consistency in videos.*
>
> *Recent video inpainting [methods] do not explicitly model object side effects, resulting in unnatural removal outcomes. Existing video object removal [methods] lack spatiotemporal correlation modeling between the object and its side effects, and consequently often produce artifacts and residual traces of the removed objects.*
>
> *'Overall, EffectErase achieves state-of-the-art performance across all datasets and evaluation metrics. It obtains the best scores on the video quality metric FVD, demonstrating superior temporal smoothness and consistency of the generated videos.*
>
> *'Our method also achieves the highest QScore and user feedback ratings, further demonstrating its effectiveness in producing visually convincing removal results.'*

For the qualitative evaluation, static results are offered in the paper (shown) directly below, as well as moving results being available in the project site and the accompanying YouTube video presentation:



Qualitative comparison on VOR-Eval across occlusion, shadow, lighting, reflection, and deformation cases. Inpainting methods struggle to remove effects outside mask, while removal models often leave visible artifacts. EffectErase removes both the target objects and their associated effects more cleanly. Please refer to sou paper for better resolution, and to project site for video examples.

We also refer the reader to diverse related examples at the project site, previewed below, as well as the official YouTube video embedded at the end of this article:



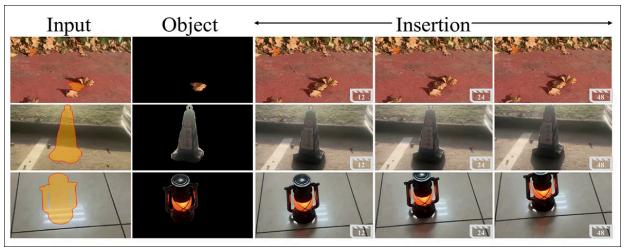
CLICK TO PLAY VIDEO ON SITE

Click to play. *A sample comparison from EffectErase's project site. Please refer to the site for better resolution (with the aforementioned caveats) and for further examples.*

The authors comment:

> *'Video inpainting [methods] often produce artifacts in masked regions and fail to completely remove the side effects caused by the removed objects. Previous object removal approaches, such as [ROSE] and [MinMax-Remover], perform well in removing target objects but still struggle with side effects, especially in occlusion, shadow, lighting, reflection and deformation scenarios.*
>
> *'In contrast, EffectErase effectively removes both target objects and their associated effects, resulting in clean, coherent, and high-quality outcomes.'*

In closing, the researchers observe that their method can also be adapted towards insertion rather than removal tasks, without the need for additional training:



Video object insertion results. EffectErase inserts objects while preserving background content and generating consistent object-induced effects such as shadows and reflections across frames.

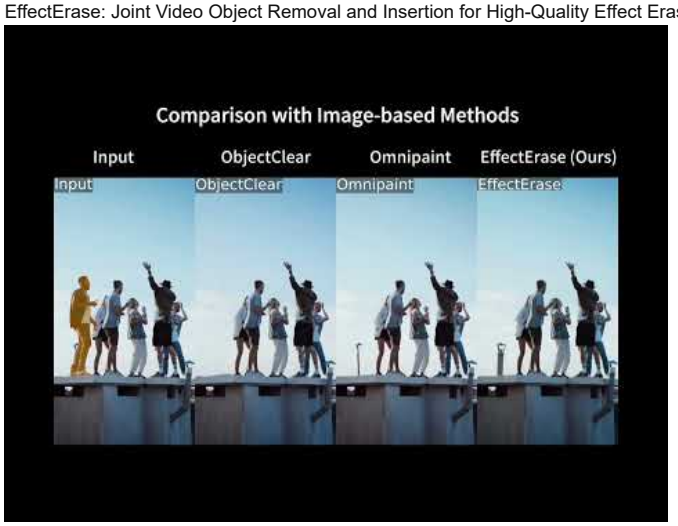
Video results for the insertion task can be seen in the (time-specific) YouTube video examples (also embedded without time-stamps at end of article).

Conclusion

A look at similar projects across the literature reveals that many are still hoping that general-purpose VFX models will eventually be able to fold this kind of functionality into a general 'toolkit' model designed for a range of effects, rather than just this specific task.

However, on the 'jack of all trades' principle, it seems reasonable to assume that dedicated systems like EffectErase will continue to maintain an edge over more general approaches; with the caveat that the gap may eventually contract enough to make the difference not worth the extra effort of training a discrete model.

EffectErase: Joint Video Object Removal and Insertion for High-Quality Effect Erasing (CVPR2026)



* *One would hope, with growing concerns around IP-provenance issue, that all such sources would be cited; but if the available materials from the new work list the source of the 3D models, I was not able to locate this reference.*

† The reference provided appears to be a generic explanatory text from 2013, with the specific VAE not detailed.

†† *Taken from the paper, this is a semantically unclear description, since fine-tuning and LoRA are different processes with very different demands.*

First published Saturday, March 21, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More