Object-WIPER: Training-Free Object and Associated Effect Removal in Videos (Supplementary Material)

Saksham Singh Kushwaha[1,*], Sayan Nag[2], Yapeng Tian[1], Kuldeep Kulkarni[2]
[1]The University of Texas at Dallas, [2]Adobe Research

Contents

1. WIPER-Bench 1
1.1. Dataset construction details 1
1.2. Examples and statistics 1

2. Implementation details 2
2.1. Object-WIPER model details 2
2.2. Baseline details 2
2.3. Evaluation metric details. 2

3. Limitations 3

4. User studies 4
4.1. Interface and setup 4
4.2. Analysis 4

5. Associated Effects Localization details 4
5.1. Analysis on text tokens 5
5.2. Replacing m^{PRO} with $\mathbf{M}^{obj}$ 5
5.3. Limitation of $\mathbf{M}^{AE}$ using OmnimatteZero . . 5
5.4. Limitation of Concept attention 5
5.5. Ablation on masks. 5

6. More experiments and ablations 5
6.1. Runtime and VRAM comparison 5
6.2. Trade-off between runtime vs. quality 6
6.3. Comparison with more baselines 6
6.4. Sensitivity to backbone 6

1. WIPER-Bench

1.1. Dataset construction details

We collected videos from Pexels [1] and YouTube [2] by searching for keywords such as "*shadow*", "*reflection*", "*mirror*", "*translucent*", "*transparent*", "*animal + shadow/reflection*" and "*object + shadow/reflection*". We avoided videos where a person's face was clearly visible, to maintain privacy and ethical reasons. In addition to simple scenes, we also included complex videos containing disconnected associated effects or multiple co-occurring effects. In total, we manually downloaded 52 candidate videos. From each video, we selected at most two non-overlapping 2-second clips, resulting in 74 candidate samples.

All landscape videos were resized to a resolution of 480×848, and portrait videos were resized to 720×400. We also resampled all videos to 24 fps. For annotation, we manually labeled the object masks frame-by-frame using the SAM2 [20] demo interface. A few videos resulted in huge segmentation errors when SAM2 was applied and were therefore discarded. After balancing category distribution, our final dataset consists of 60 videos.

1.2. Examples and statistics

Given the collected data, the distribution of categories is shown in Fig. A1. These statistics reflect the natural availability of such phenomena in real-world videos. The final dataset includes 25 reflection cases, 14 mirror cases, 11 shadow cases, and 16 translucent associated effects. Additionally, 6 videos contain multiple associated effects, and 12 videos include disconnected associations. Examples of multiple and disconnected associations are shown in Fig. A2.

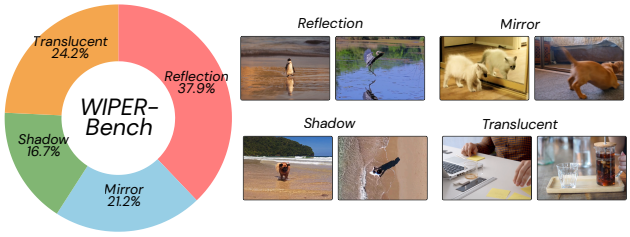

Figure A1. Statistics and example cases from WIPER-bench for evaluating object removal with associated effects.

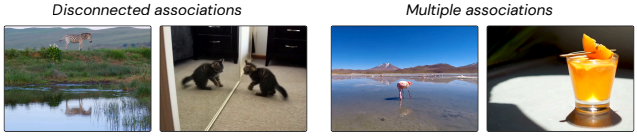

Figure A2. WIPER-bench also includes naturally occurring complex cases, such as disconnected associations and multiple co-occurring associations.

2. Implementation details

2.1. Object-WIPER model details

We use pretrained Hunyuan-T2V model [10] as our video-generation model. It consists of $M = 20$ MMDiT and $S = 40$ single blocks. We the use RF-Solver [25] sampler for inversion and denoising that has 25 time steps through the model. We store and copy background feature values for $k = 15$ time steps and last $r = 20$ single (or self-attention). We use classifier free guidance (cfg) value of 1 during inversion and 5 during denoising. We apply adaptive masking for $k = 15$ time steps and using all 40 single blocks. For all the MMDiT and single blocks, we apply attention scaling for 10 steps. We choose $c = 0.8$ and $b = 1.2$. To calculate the associated mask we use MMDiT layers of intermediate time steps $t_i \in \{6, 7, 10\}$. To improve readability, we summarize all symbols and notations used in the paper in Tab. A1.

2.2. Baseline details

Training based methods: We compare our method against several state-of-the-art object removal approaches, including VACE [9], ProPainter [28], ROSE [18], and GenProp [17]. ROSE and GenProp are trained to remove both object and its associated effect, similar to it we want to do that in a training-free way. For VACE, ProPainter, and ROSE, we use the official checkpoints and publicly released implementations. As GenProp is not open-source, we contacted the authors directly and obtained their predicted videos for evaluation.

Training-free methods. Given our training-free approach, we mainly compare our method with previous (open-sourced) training-free approaches, including KV-Edit [29] and Attentive-Eraser. Since these approaches are image-based we implement for the video by running them frame-wise. We extend KV-Edit for videos, as explained next.

KV-Edit-Video KV-Edit [29] demonstrates strong performance on image-based object removal and is originally implemented on the FLUX [11]. However, performing object removal independently on each frame does not account for temporal consistency in videos. Given the architectural similarity between FLUX and Hunyuan, and to ensure a fair comparison, we extend KV-Edit to operate on the Hunyuan video model.

Following their approach, we store all intermediate tokens and (self-attention) video key/value features during inversion. We then reinitialize the tokens corresponding to the object region and, during denoising, replace the tokens and (self-attention) key/value features for the background region with those saved from inversion. Due to CPU memory limitations, we exclude saving and restoring the key/value tensors for the MMDiT blocks.

We illustrate an example of object removal in Fig. A3. The (frame-wise) KV-Edit produces inpainted regions that are temporally inconsistent across frames. Extending KV-Edit to operate on video tokens improves temporal coherence in the inpainted regions. However, KV-Edit-Video still introduces boundary inconsistencies and noticeable artifacts because it copies background tokens and attention features using a fixed mask. In contrast, our method employs a timestep-adaptive masking strategy that refines the fixed mask avoids copying all background tokens, resulting in both temporally and spatially consistent object removal.

Figure A3. Object removal comparison. KV-Edit (frame-wise) produces temporally inconsistent inpainting across frames. Extending the method to video latents, KV-Edit-Video, improves temporal coherence, but this extension still introduces noticeable artifacts along object–background boundaries.

OmnimatteZero OmnimatteZero [21] introduces a training-free approach for generating video omnimattes. One of their intermediate goals involves removing foreground objects to get backgrounds. However, due to the unavailability of public code and insufficient implementation details, we were unable to reproduce their method and therefore could not include it in our comparisons. Moreover, their primary focus is on producing omnimattes and evaluating them on simulated datasets specifically designed for that task.

In contrast, our objective is to remove objects from real-world videos and to evaluate performance directly on such real data. Unlike omnimatte datasets, which provide ground-truth background videos without objects, real videos do not have ground-truth object-free references. To address this gap, we also propose a new evaluation metric, **TokSim**, tailored for assessing object removal quality in real-world videos.

2.3. Evaluation metric details.

TokSim. Due to the lack of appropriate metrics for evaluating object removal in videos, we propose TokenSimilarity, a token-level metric computed using image patch embeddings extracted from DINOv3. For each pair of consecutive frames f and $f+1$, we first compute the union of their object masks. If the object has been successfully removed, the

Table A1. Summary of notations used throughout the paper.

Variable	Value	Dimension	Description
$\mathcal{I}_k$	-	$3 \times (F+1) \times H \times W$	Input pixel video frames
$\hat{\mathcal{I}}_k$	-	$3 \times (F+1) \times H \times W$	Predicted pixel video frames
$\mathbf{Z}_t$	-	$16 \times (F/4+1) \times H/8 \times W/8$	Video latent at timestep t during inversion
$\tilde{\mathbf{Z}}_t$	-	$16 \times (F/4+1) \times H/8 \times W/8$	Video latent at timestep t during denoising
$\mathbf{Z}(j)$	-	$16 \times 1 \times H/8 \times W/8$	j^{th} video latent frame during inversion
$\mathbf{M}^{obj}$	-	$1 \times (F+1) \times H \times W$	User provided binary object pixel mask
	-	$16 \times (F/4+1) \times H/8 \times W/8$	Max-pool & repeat to align with video latent
	-	$1 \times (F/4+1) \times H/16 \times W/16$	Max-pool to align with video tokens
$\mathbf{M}^{AE}$	-	$1 \times (F/4+1) \times H/16 \times W/16$	Estimated associated mask aligned with video tokens
	-	$16 \times (F/4+1) \times H/8 \times W/8$	Upsampled & repeat to align with video latent
	-	$1 \times (F+1) \times H \times W$	Upsampled binary mask to align with pixel video
$\hat{\mathbf{M}}_t^{obj}$	-	$1 \times (F/4+1) \times H/16 \times W/16$	Estimated adaptive mask at timestep t aligned with video tokens
m^{PRO}	-	$1 \times (F/4+1) \times H/16 \times W/16$	Estimated Proposal mask at timestep aligned with video tokens
P_s; P_t	-	-	Input source and target text prompts, respectively
$\mathbf{f}_T$; $\mathbf{f}_I$	-	-	Video and text feature embeddings, before attention
N_I; N_T	-	-	Number of video patches and text tokens
d_T; d_I; d	-	-	Video feature dimension; Text feature dimension; Shared dimension
$(\mathbf{Q}_T, \mathbf{K}_T, \mathbf{V}_T)$; $(\mathbf{Q}_I, \mathbf{K}_I, \mathbf{V}_I)$	-	$N_T \times d$; $N_I \times d$	Query, Key & Values for Video and Text tokens
$\mathbf{A}^{X \to Y}$	-	$N_X \times N_Y$	Attention maps from X to Y ($X, Y \in \{I, T\}$)
$RS(j)$	-	$1 \times (F/4+1) \times H/16 \times W/16$	Object response score for j^{th} frame aligned with video token
c	0.8	-	Attention scaling factor for background to object attention
b	1.2	-	Attention scaling factor for object to background attention
k	15	-	Number of timesteps for value feature saving (inversion) and copying (denoising)
r	20	-	Number of last single blocks for which value saving and copying happens

union of the masks defines the object-token region, which should now resemble the surrounding background tokens and remain consistent with the corresponding region in the next frame.

For the tokens within the object region, we measure their embedding distance to the corresponding tokens in the ground-truth frame f, as well as their similarity to tokens in frame $f+1$. Additionally, we compare these object-region tokens with nearby background patches f_{bg}) within a 24-pixel neighbourhood outside the union mask. These comparisons collectively quantify how well the removed region integrates with its temporal and spatial context.

BG-PSNR. We evaluate background preservation by computing the PSNR (Peak Signal-to-Noise Ratio) over the unmasked regions of the video.

FG-flickering. Temporal flickering was introduced in VBench [8] to assess the temporal quality of generated videos. Building on this idea, we compute the L1 difference between consecutive frames, but restrict the evaluation to the object region. For each pair of consecutive frames, we take the union of their object masks and compute the L1 distance only within this region. By focusing on the former object area, FG-flickering isolates the temporal stability of the inpainted region, making it significantly more sensitive to object-removal inconsistencies than global flicker metrics.

Text-alignment. We compute the cosine similarity between the CLIP [19] embeddings of the output video frame and the target text prompt.

Quality. We use DOVER [26] to measure overall video quality. However, we observe that this global metric does not reliably reflect the quality of object removal.

For videos containing associated effects, we expand the original object mask by taking its union with the upsampled (calculated) associated-effect masks. This augmented mask more accurately separates the object (+ associated effect) region from the background for evaluation.

3. Limitations

While our method is particularly impressive in identifying the associated effects and removing them, we note that the inherent nature of the training-free paradigm in which our method operates in introduces several limitations. Specifically, background preservation ability of our method is limited by the reconstruction ability of the RF-Solver Edit [25]. For example, the background PSNR of the inversion–denoising reconstruction on the DAVIS dataset is only 25.44 dB. This indicates that even RF-Solver Edit alone can introduce undesirable artifacts in the background region during inversion and denoising.

Our approach is further bounded by the capacity of the underlying video diffusion model and its VAE reconstruction. The video model may struggle with highly complex or previously unseen cases, leading to degraded results. Notably, the background PSNR of the Video-VAE reconstruction on DAVIS (30.27 dB) is 3.7 dB lower than that of the Image-VAE reconstruction (34.05 dB), highlighting a gap in reconstruction quality that directly impacts background

preservation of our approach.

4. User studies

4.1. Interface and setup

We conduct human evaluation study to show the efficacy of our method in the training-free regime as well as the effectiveness of TokSimin estimating the object removal ability of different methods. Specifically, we do 15 pairwise comparisons between our result and a baseline result randomly selected from one of the three training-free algorithms, KV-edit [29], KV-edit-video [29] and Attentive Eraser [22], for three separate questions, 'Video Quality', 'Object Removal' and 'Background Preservation'. We show the interface for user-study in Fig. A4. For the video

Figure A4. User study interface. We ask the users three types of questions related to video quality, object removal quality and background preservation quality.

quality assessment, we show only the results from our approach and one of the baseline approaches and ask the question, 'Which of the two videos has better video quality?'. For the object removal assessment, we show the input video with the mask for the object to be removed overlayed and the two results, and ask the question, 'Given the input video, which of the two results have better object removal?'. For the background preservation study, we show the input video with mask for the object to removed overlayed and the results, and ask the question, 'Given the input video, which of the two results have better background preservation with respect to input video?'.

4.2. Analysis

Human Preferences: In total we collected responses from 10 users across 45 pairwise comparisons, making it a total of 450 responses. For video quality, our method was preferred 96.67% of the times. For the object removal, our method was preferred 90.67% of the videos, and for background preservation, our method was preferred 77.33% of the times. As shown through metrics in the main paper, it is expected that our method performs betters in terms of video quality, object removal as opposed to background preservation.

TokSim and Human Preference Agreement: We also obtained TokSim for each of the videos in the pairwise comparisons and determined which video was preferred if we strictly assume higher TokSim scores is akin to better object removal. We dub these as 'TokSim Preferences'. For each of the 15 pairwise comparisons, we compare the TokSim preferences with perferences of 10 users and found that TokSim preferences is 83.64% accurate with respect to human. This clearly shows the value of using the metric proposed in being a strong replacement of human evaluation.

Inter-rater Agreement: Inter-rater reliability was assessed using Fleiss κ which is appropriate for evaluating consistency among more than two raters who assign categorical judgments [4]. The observed κ value of 0.72 indicates substantial agreement amongst the raters suggesting that they demonstrated a high level of concordance in their evaluations and that the ratings are sufficiently consistent to support subsequent analyses [12].

5. Associated Effects Localization details

Since only the object mask is provided and the associated effects also need to be removed, we leverage the model's prior knowledge encoded in the unified text–video token space within the joint-attention (MMDiT) layers. For reflection and shadow cases, we use text tokens corresponding to both the *object* and its *effect* to guide the removal process. For mirror cases, where the reflected object is visually real object, we found that using only the *object*-related text tokens yields better localization.

5.1. Analysis on text tokens

For shadow and reflection associated effects, we empirically find that using only *object*-text tokens or only *effect*-text tokens fails to capture the full object–effect region. For example, as shown in Fig. A5, using only "duck" text tokens highlights only the object, while using only "reflection" tokens produces incorrect and overly spread localizations. Therefore, we jointly use both token types, which yields a compact and accurate localization of the object and its associated effect.

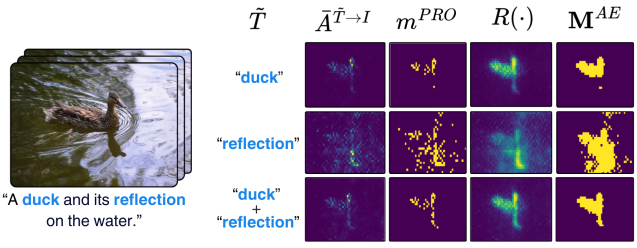

Figure A5. Effect of text tokens on localization. Using only *object*-text tokens or only *effect*-text tokens leads to incorrect localization, whereas combining both yields accurate object–effect masks.

5.2. Replacing m^{PRO} with $\mathbf{M}^{obj}$

We analyse whether the proposal mask m^{PRO} must be computed using text guidance, or if the user-provided object mask alone can serve as an adequate proposal. As shown in Fig. A7, skipping the proposal-mask estimation step results in masks that fail to capture the associated effects. This highlights the importance of the text-guided proposal stage for associated effect localization.

5.3. Limitation of $\mathbf{M}^{AE}$ using OmnimatteZero

Generative-Omnimatte [13] and OmnimatteZero [21] estimates the associated-effect regions by selecting per-frame high-response tokens conditioned on the user-provided object mask $\mathbf{M}^{obj}$. However, as shown in Fig. A7, this strategy fails to correctly identify the associated-effect regions.

5.4. Limitation of Concept attention

We observe that text-to-image approaches [6, 7], which use text prompts to localize concepts in images, struggle to achieve the level of spatial precision required to distinguish the object, its associated effects, and the background. As shown in Fig. A6, concept attention often produces coarse or ambiguous activations that fail to correctly isolate both the object and its associated effects. This makes it non-trivial to leverage such methods for accurate object (+associated effect) separation from background. In contrast, our text-to-video–based approach provides significantly sharper and more consistent localization, enabling reliable identification of both the object and its associated effects.

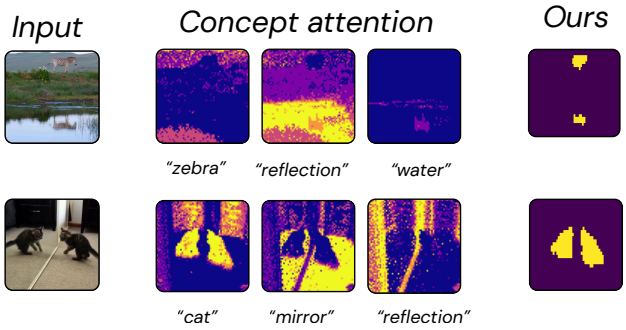

Figure A6. Comparison with concept attention [6]. We show the (left) input image and (middle) activations for text concepts using Concept attention and (right) our estimated object-associated effect. We observe that concept-attention struggle to precisely localize the object and its associated effects, while our text-to-video approach provides accurate localization.

5.5. Ablation on masks.

We compare how would the combination of different masking strategy helps. In Tab. A2, we compare on the subset of DAVIS with associated effects. We observe that our strategy of Adpative masking on M^{obj} and adding M^{AE} outperforms any other combination of masking for object removal.

Masking Strategy	TokSim↑	BG-PSNR↑	Text-align($\times 10^2$)↑
$\mathbf{M}^{obj}$	27.69	<u>22.11</u>	25.69
$\mathbf{M}^{AE}$	26.75	21.69	25.16
$\mathbf{M}^{obj} \cup \mathbf{M}^{AE}$	<u>28.66</u>	21.63	25.84
$\hat{\mathbf{M}}_t^{obj}$	27.19	**22.37**	25.56
$\hat{\mathbf{M}}_t^{AE}$	28.52	21.99	<u>26.20</u>
$(\widehat{\mathbf{M}^{obj} \cup \mathbf{M}^{AE}})_t$	28.49	21.64	26.00
$\hat{\mathbf{M}}_t^{obj} \cup \mathbf{M}^{AE}$ **(Ours)**	**29.32**	21.64	**26.54**

Table A2. Ablation on DAVIS subset with associated effects. $\mathbf{M}^{obj}$, $\mathbf{M}^{AE}$, $\hat{\mathbf{M}}_t(\cdot)$ are the object, associated and time adapted mask, respectively.

6. More experiments and ablations

6.1. Runtime and VRAM comparison

We compare the runtime and peak VRAM of our method against training-free baselines. For fairness, we exclude model-loading and I/O overheads (image/video loading and saving) and report only the inference time. The results are averaged over 10 runs on videos of size $25 \times 480 \times 848$ (Frames $\times$ Height $\times$ Width) and shown in Tab. A3. As shown in Tab. A3, our method achieves inference time comparable to existing training-free approaches, while surpass-

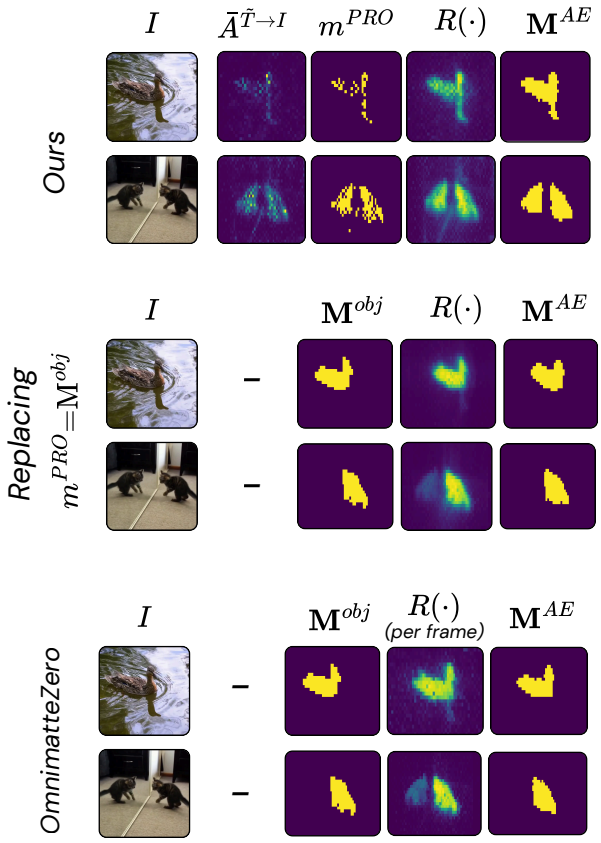

Figure A7. Comparison of associated-effect mask localization. **Top:** Our method accurately localizes both the object and its associated effects. **Middle:** Replacing m^{PRO} with the user-provided $\mathbf{M}^{obj}$ (i.e., skipping the proposal-mask estimation) results in masks that fail to capture the associated effects. **Bottom:** Approaches used by OmnimatteZero [21] and Generative-Omnimatte [13] are unable to correctly localize the associated-effect regions.

ing them in object-removal quality. We note that VRAM usage largely depends on the memory requirements of the underlying image or video generation model. Compared with the video-based baseline, KV-Edit-Video, our approach uses the same amount of memory while achieving superior object removal.

Method	**Run-time↓ (sec)**	**VRAM↓ (GB)**	**Removal Quality↑ (Toksim)**
KV-Edit [29]	323.85	22.47	23.17
Attentive-Eraser	305.52	10.52	30.82
KV-Edit-Video	551.35	30.53	28.68
Object-WIPER (ours)	354.69	30.53	32.80

Table A3. Run-time comparison. Our method achieves comparable inference time

6.2. Trade-off between runtime vs. quality

The Tab. A4 details our method (Hunyuan backbone) on Davis ($480 \times 848 \times 25$). Varying the number of steps reveals a clear tradeoff between removal quality and inference time, while peak VRAM remains constant.

#steps	10	20	25	50	100
Speed (s)	192.03	312.01	354.69	626.83	1134.04
Removal Quality (TokSim)	26.62	27.74	28.45	30.85	36.51
VRAM (GB)	30.53	30.53	30.53	30.53	30.53

Table A4. Run-time vs. quality trade-off

6.3. Comparison with more baselines

Beyond the main table, we also compare against several training-based and training-free approaches i.e. FloED [5], Diffueraser [15], Minimax-Remover [31], InsViE [27], Senorita-2M [30], Lucy-Edit [23], Ditto [3], ICVE [16], Rorem [14]. We observe that our approach outperforms these methods in object removal.

Data	Metric	FloED	Diffueraser	Minimax-Remover	InsViE	Senorita-2M	Lucy-Edit	Ditto	ICVE	Rorem	**Object-WIPER**
DAVIS	TokSim	29.17	29.57	30.36	10.98	23.08	5.79	11.11	27.32	30.89	**32.80**
	BG-PSNR	20.41	**26.13**	21.54	13.66	17.75	20.69	12.42	18.47	24.71	23.02
	Text-align	25.81	26.23	26.16	23.44	24.26	22.86	21.91	225.20	26.12	**26.63**
WIPER-Bench	TokSim	9.90	23.27	25.59	12.12	18.47	6.68	17.22	24.52	26.10	**33.09**
	BG-PSNR	19.88	**37.39**	24.30	16.16	19.69	22.04	12.60	31.34	27.54	27.53
	Text-align	24.70	25.86	26.09	24.79	25.04	25.14	23.77	26.38	26.16	**26.91**

Table A5. We compare our approach to more baselines.

6.4. Sensitivity to backbone

We observe consistent performance on DAVIS across two different backbones (see Tab. A6 and Fig. A8), demonstrating that our method is backbone-agnostic. Note that for Wan2.2 (TI2V-5B) [24], we reused just the associated masks from Hunyuan.

	TokSim	BG-PSNR	FG-flicker	Text-align	Qual.
Ours (Hunyuan)	32.80	23.02	16.37	26.63	61.62
Ours (Wan2.2)	30.23	23.08	21.80	26.84	60.57

Table A6. Quantitative comparison of our approach using the Hunyuan and Wan2.2 models. We observe that our method achieves consistent performance across different backbones.

Figure A8. Qualitative comparison of our approach using the Hunyuan and Wan2.2 backbones. Our method, implemented on both backbones, cleanly removes the object and its associated effects.

References

[1] https://www.pexels.com/. 1

[2] https://www.youtube.com/. 1

[3] Qingyan Bai, Qiuyu Wang, Hao Ouyang, Yue Yu, Hanlin Wang, Wen Wang, Ka Leong Cheng, Shuailei Ma, Yanhong Zeng, Zichen Liu, Yinghao Xu, Yujun Shen, and Qifeng Chen. Scaling instruction-based video editing with a high-quality synthetic dataset. arXiv preprint arXiv:2510.15742, 2025. 6

[4] Joseph L Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and psychological measurement, 33(3):613–619, 1973. 4

[5] Bohai Gu, Hao Luo, Song Guo, and Peiran Dong. Advanced video inpainting using optical flow-guided efficient diffusion. arXiv preprint arXiv:2412.00857, 2024. 6

[6] Alec Helbling, Tuna Han Salih Meral, Ben Hoover, Pinar Yanardag, and Duen Horng Chau. Conceptattention: Diffusion transformers learn highly interpretable features, 2025. 5

[7] Yihan Hu, Jianing Peng, Yiheng Lin, Ting Liu, Xiaochao Qu, Luoqi Liu, Yao Zhao, and Yunchao Wei. Dcedit: Dual-level controlled image editing via precisely localized semantics. arXiv preprint arXiv:2503.16795, 2025. 5

[8] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 21807–21818, 2024. 3

[9] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. arXiv preprint arXiv:2503.07598, 2025. 2

[10] 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. 2

[11] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2

[12] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. biometrics, pages 159–174, 1977. 4

[13] 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 Proceedings of the Computer Vision and Pattern Recognition Conference, pages 12522–12532, 2025. 5, 6

[14] Ruibin Li, Yang Tao, Guo Song, and Zhang Lei. Rorem: Training a robust object remover with human-in-the-loop. 2025. 6

[15] Xiaowen Li, Haolan Xue, Peiran Ren, and Liefeng Bo. Diffueraser: A diffusion model for video inpainting, 2025. 6

[16] Xinyao Liao, Xianfang Zeng, Ziye Song, Zhoujie Fu, Gang Yu, and Guosheng Lin. In-context learning with unpaired clips for instruction-based video editing. arXiv preprint arXiv:2510.14648, 2025. 6

[17] Shaoteng Liu, Tianyu Wang, Jui-Hsien Wang, Qing Liu, Zhifei Zhang, Joon-Young Lee, Yijun Li, Bei Yu, Zhe Lin, Soo Ye Kim, et al. Generative video propagation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 17712–17722, 2025. 2

[18] 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. arXiv preprint arXiv:2508.18633, 2025. 2

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PmLR, 2021. 3

[20] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714, 2024. 1

[21] Dvir Samuel, Matan Levy, Nir Darshan, Gal Chechik, and Rami Ben-Ari. Omnimattezero: Fast training-free omnimatte with pre-trained video diffusion models. In SIGGRAPH Asia 2025 Conference Papers, 2025. 2, 5, 6

[22] Wenhao Sun, Xue-Mei Dong, Benlei Cui, and Jingqun Tang. Attentive eraser: Unleashing diffusion model's object removal potential via self-attention redirection guidance. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pages 20734–20742, 2025. 4

[23] DecartAI Team. Lucy edit: Open-weight text-guided video editing, 2025. 6

[24] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314, 2025. 6

[25] Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. arXiv preprint arXiv:2411.04746, 2024. 2, 3

[26] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou Hou, Annan Wang, Wenxiu Sun Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In International Conference on Computer Vision (ICCV), 2023. 3

[27] Yuhui Wu, Liyi Chen, Ruibin Li, Shihao Wang, Chenxi Xie, and Lei Zhang. Insvie-1m: Effective instruction-based video editing with elaborate dataset construction. arXiv preprint arXiv:2503.20287, 2025. 6

[28] 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, pages 10477–10486, 2023. 2

[29] Tianrui Zhu, Shiyi Zhang, Jiawei Shao, and Yansong Tang. Kv-edit: Training-free image editing for precise background preservation. arXiv preprint arXiv:2502.17363, 2025. 2, 4, 6

[30] Bojia Zi, Penghui Ruan, Marco Chen, Xianbiao Qi, Shaozhe Hao, Shihao Zhao, Youze Huang, Bin Liang, Rong Xiao, and Kam-Fai Wong. Señorita-2m: A high-quality instruction-based dataset for general video editing by video specialists. arXiv preprint arXiv:2502.06734, 2025. 6

[31] Bojia Zi and Jianan Wang Shihao Zhao Rong Xiao Kam-Fai Wong Weixuan Peng, Xianbiao Qi. Minimax-remover: Taming bad noise helps video object removal. arXiv preprint arXiv:2505.24873, 2025. 6