EffectErase: Joint Video Object Removal and Insertion for High-Quality Effect Erasing Supplementary Material

Yang Fu Yike Zheng Ziyun Dai Henghui Ding✉
Institute of Big Data, College of Computer Science and Artificial Intelligence, Fudan University, China
https://henghuiding.com/EffectErase/

In the supplement, we provide additional dataset details in Sec. 6, further method descriptions in Sec. 7, and more qualitative results in Sec. 8.

6. Details of Dataset Construction

In this section, we provide a detailed description of the captured and rendered components of our **V**ideo **O**bject **R**emoval (**VOR**) dataset used to train EffectErase.

6.1. Real-World Data

Consistent Data Pairs. Each pair consists of one video where the target object is present with its effects and a counterpart where both are absent. To keep the two recordings identical, as shown in Fig. 10, we develop a custom capture app that locks exposure and focus across the entire pair, ensures matched file names and fixed recording durations, enables Bluetooth triggering to avoid screen-touch motion, and uses a tripod to eliminate camera shake.

Diverse Scenes and Objects. We collect data across a wide range of real-world environments, including parks, campuses, and streets, spanning a total of 293 scenes and covering over 45 scene categories. The dataset also features a broad set of objects, ranging from static items such as sports balls and tools to dynamic subjects including children, teenagers, and various vehicles.

Ken Burns Effects. We propose an extended version of the Ken Burns effect that provides fourteen distinct camera-motion patterns. These include basic zoom-in and zoom-out motions; directional motions such as panning left or right and tilting up or down; combined zoom–translation motions; a walk-bob motion that mimics the vertical sway of handheld footage; and a random-combo mode that randomly mixes zoom and translation directions. For each clip, we randomly select five motion types and assign each type a randomized zoom curve and translation intensity. The module then updates a virtual camera center over time and crops the corresponding view to a fixed resolution, producing natural and diverse camera-movement variants that enhance training for the video object removal model.

6.2. Synthesized Data

3D Enviroments. We collect 150 high-quality 3D environment assets from free online resources. These scenes cover a wide range of realistic daily-life settings across both indoor and outdoor domains, *e.g.* city streets, farms, coastal areas, mountains, parking lots, classrooms and forests.

Characters with Animations. We include a diverse set of animated characters and objects, such as dancing humans, walking bears, moving boats, and flying balloons, covering realistic, anime, and game-style visual domains.

Camera Trajectories. Due to the wide variety of camera motions and shooting angles in real scenarios, we aim to cover as many camera movement patterns as possible. To this end, we manually design both realistic camera paths and natural camera motion behaviors such as zoom and pan, thereby ensuring that the synthesized movements closely mimic human-operated filming practices.

6.3. Mask Annotation

We first provide a point prompt to obtain the mask in the first frame and manually verify its quality. The same point prompt is then fed to SAM2 [12] to propagate the mask across the entire sequence. We review all propagated mask sequences and remove those that fail to maintain stable and complete object coverage across all frames.

6.4. Dataset Statics

As shown in Table 4, we provide a detailed comparison between our VOR dataset and existing image- and video-based removal datasets. We summarize the image-based datasets and the video-based datasets. Compared with prior work, VOR offers substantially richer scene diversity,

✉ Corresponding author (henghui.ding@gmail.com).

Table 4. Comparison of object removal datasets. Image-level datasets are listed above the line, and video-level datasets are listed below. "–" denotes unreported or not applicable. Synth. (3D) denotes data generated using a graphics rendering engine, while Synth. (paste) denotes data created by directly pasting cropped foreground objects onto backgrounds.

Dataset	Source	Dynamic Camera	Dynamic Object	Dynamic Background	Scene Types	Object Classes	Image Pairs	Video Pairs	Average Duration (s)	Total Hours
ObjectDrop [16]	Real	✗	✗	✗	–	–	2.5K	–	–	–
Syn4Removal [7]	Synth. (paste)	✗	✗	✗	–	–	1,000K	–	–	–
LayerDecomp [17]	Synth. (paste)	✗	✗	✗	–	–	6.0K	–	–	–
OmniPaint [18]	Real	✗	✗	✗	–	–	3.3K	–	–	–
RORem [8]	Synth. (paste)	✗	✗	✗	–	–	201.1K	–	–	–
RORD [13]	Real	✗	✓	✗	24	76	516.7K	3,106	–	5.98
Video4Removal [15]	Real	✗	✓	✗	6	–	134.3K	–	–	1.55
ROSE [10]	Synth.(3D)	✓	✗	✗	25	102	1,501.0K	16,678	6.00	27.79
VOR (Ours)	Real + Synth.(3D)	✓	✓	✓	**67**	**366**	**12,556.8K**	**60,000**	**8.72**	**145.33**

Figure 10. **Data capture software.** Our app records aligned video pairs by locking exposure and focus, matching file names and durations, enabling reliable Bluetooth triggering for stable control, and using a tripod to remove camera shake.

Outdoor (41)

Urban (13)
street/road; downtown; market; plaza; alley; bridge; parking lot; construction site; harbor / dock; garden; amusement park; outdoor cafe; night market.

Natural (19)
park; forest; grassland; mountain; river; lake; waterfall; beach; sea; island; cave; snowfield; garden; bamboo grove; hot spring; canyon; hill; farm; greenhouse.

Transportation (3)
highway; subway/bus station; gas station.

Sports / Athletic (6)
tennis court; basketball court; soccer field; table tennis; volleyball; running tack.

Indoor(26)

Home spaces(9)
living room; bedroom; dining room; kitchen; bathroom; balcony; restroom; garage; home office.

Commercial / Public (17)
store; corridor; cafe; restaurant; gym; hotel; office; meeting room; classroom; lab; dormitory; Library; exhibition hall; warehouse; shopping mall; zoo; factory/workshop.

Figure 11. **Scene category hierarchy.** Our taxonomy organizes 67 scene types into structured outdoor and indoor groups.

broader object coverage, longer video durations, and a significantly larger number of paired sequences.

Since no unified scene taxonomy exists across datasets, we introduce our own categorization scheme to standardize all scene types in Fig. 11, covering both indoor and outdoor environments with a total of 67 comprehensive categories. Specifically, for RORD [13], its original scene labels are merged into our taxonomy; for Video4Removal [15], scene types are assigned based on the descriptions in the paper and aligned with our scheme; and for ROSE [10], we manually inspect every scene in the raw data and annotate them according to our proposed categorization.

For video pair counts, the numbers for RORD [13] are obtained by counting the lowest-level folders in the dataset structure. The total video hours of RORD [13] and Video4Removal [15] are estimated by converting the total frame count to duration using 24 fps.

7. Method Details

7.1. Details of the Proposed modules

Adaptor Details. The adaptor is implemented as a 3D convolutional layer with a kernel size of $1 \times 2 \times 2$ and a stride of $1 \times 2 \times 2$. To improve convergence, the first sixteen input channels of its weights are copied from the convolution used in the original patch-embedding module, while the remaining channels are initialized with Xavier uniform initialization [3]. All bias terms are zero-initialized.

Projector Details. The projector maps the object-image features extracted by the image encoder into the latent space required by our model. It is composed of two sequential MLP blocks: the first transforms the input embedding dimension to the output dimension, and the second further refines the representation with a residual MLP. Each block applies LayerNorm [1], a linear projection, a GELU activation [5], and a second linear projection, while the second block includes a residual connection. A final LayerNorm is applied to stabilize the projected token.

Mapper Details. The mapper predicts an effect-area distribution map from the fused cross-attention features. We aggregate the cross-attention maps from all DiT layers [11] and apply max-pooling across layers to obtain a compact feature volume. This volume is then processed by the mapper, implemented as a lightweight per-pixel MLP operating on the channel dimension. The module applies a linear projection, a GELU activation [5], and a second linear projection to produce a logit map for each frame.

7.2. Training details

Similar to previous work [10], the backbone model is a controllable generation variant of Wan2.1 1.3B [14]. We optimize the network with AdamW [9] using a learning rate of 1×10^{-4} and a batch size of 1 through gradient accumulation. Training is conducted for up to 120K iterations. To adapt the base model to the video object-removal task, we apply LoRA [6] to the attention projections q, k, v, o and the feed-forward layers `ffn.0` and `ffn.2`, with all LoRA weights initialized using Kaiming initialization [4].

7.3. Inference details

During inference, the model supports both removal and insertion. For removal, we provide the input video together with a mask video, and the model outputs the object-removed result. For insertion, we provide the background video and an object video, and the model generates the inserted output. All denoising steps are set to 50.

7.4. Metric details

QScore. To further assess the removal quality, we use the Qwen-VL model [2] to evaluate each removed video with a designed prompt as shown in Fig. 12. The evaluation considers both removal completeness and visual artifacts, and the final QScore is obtained by averaging the results.

User Study. We conduct a user study with 20 volunteers, where each participant scores 195 generated videos from VOR-Wild, and the final score is obtained by averaging all individual ratings across participants.

8. More Results

8.1. Effect-region Erasing Evaluation

As shown in the Tab. 5, EffectErase effectively removes effect regions outside the object mask, with evaluation metrics computed only over the corresponding effect regions.

Table 5. Effect-region erasing evaluation.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓
ROSE	30.267	0.930	0.084	135.013
EffectErase	**32.747**	**0.939**	**0.069**	**98.266**

8.2. More Results of the Insertion Task

Please refer to Fig. 13 for additional results of EffectErase applied to the insertion task.

8.3. More Results of EffectErase

Please refer to Fig. 14 for additional results of EffectErase on in-the-wild data.

8.4. More Comparison with SOTA Methods

Please refer to Fig. 15 for additional qualitative comparisons with state-of-the-art methods.

8.5. Failure Cases and Analysis

Failure cases mainly arise when it is ambiguous whether effects or accessories belong to the target object. As shown in the Fig. 16, 1) the residual lighting may originate from other light sources, yet remains visually natural after removal; 2) parts of the dog's shadow are heavily entangled with the person's shadow, and the leash cannot be clearly assigned to either the dog or the person.

Figure 16. Failure cases when effects or accessories cannot be clearly attributed to the target object.

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 first 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>

Figure 12. **Prompt used for QScore evaluation.** The prompt guides Qwen-VL to assess removal completeness and visual artifacts.

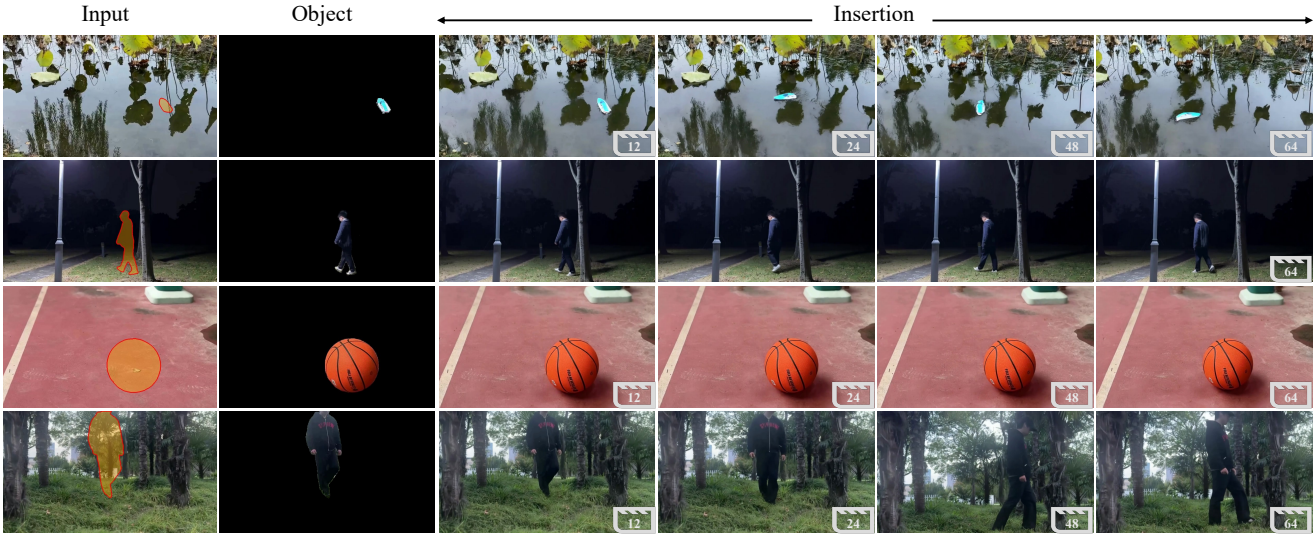

Figure 13. More insertion results of EffectErase.

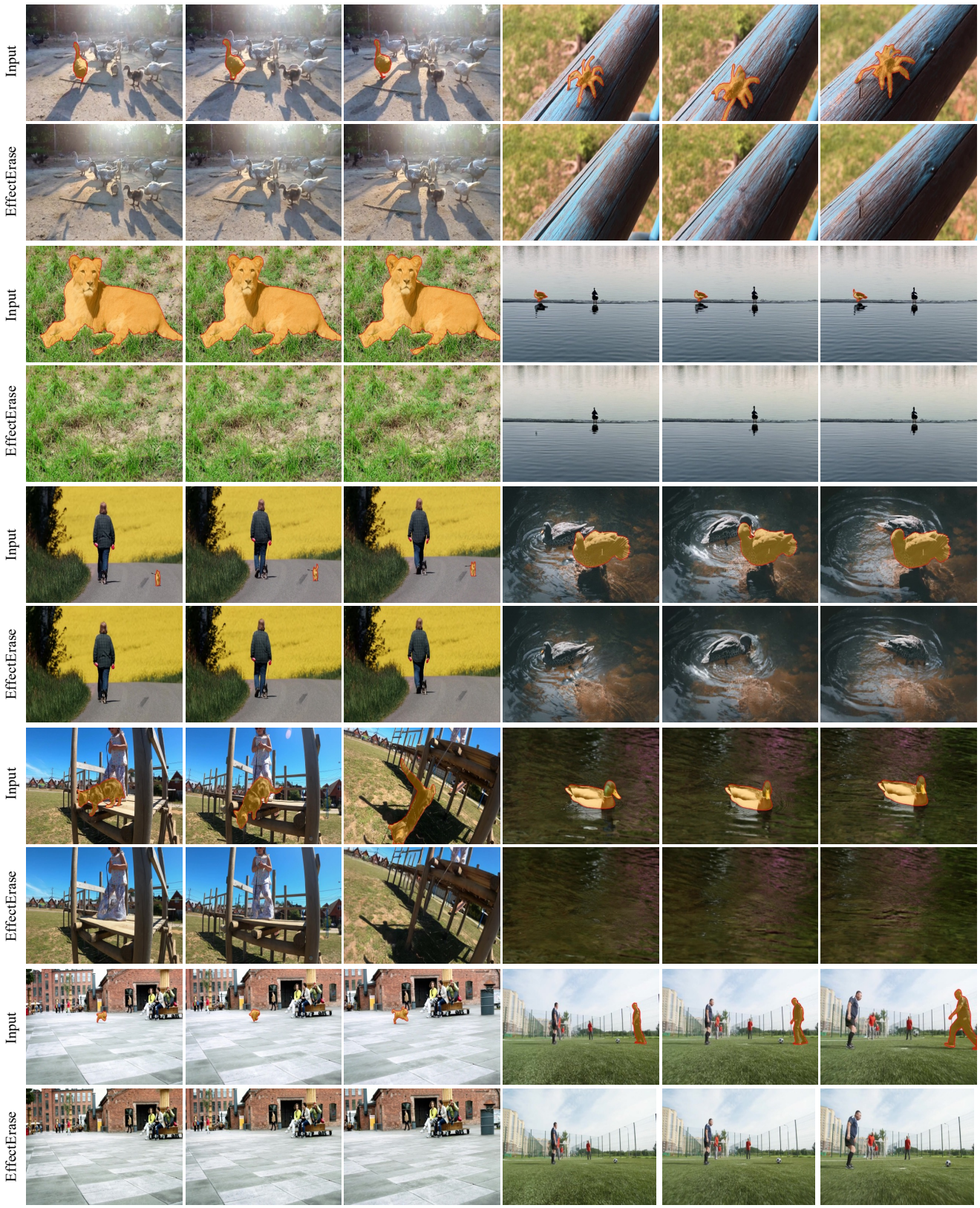

Figure 14. More removal results of EffectErase.

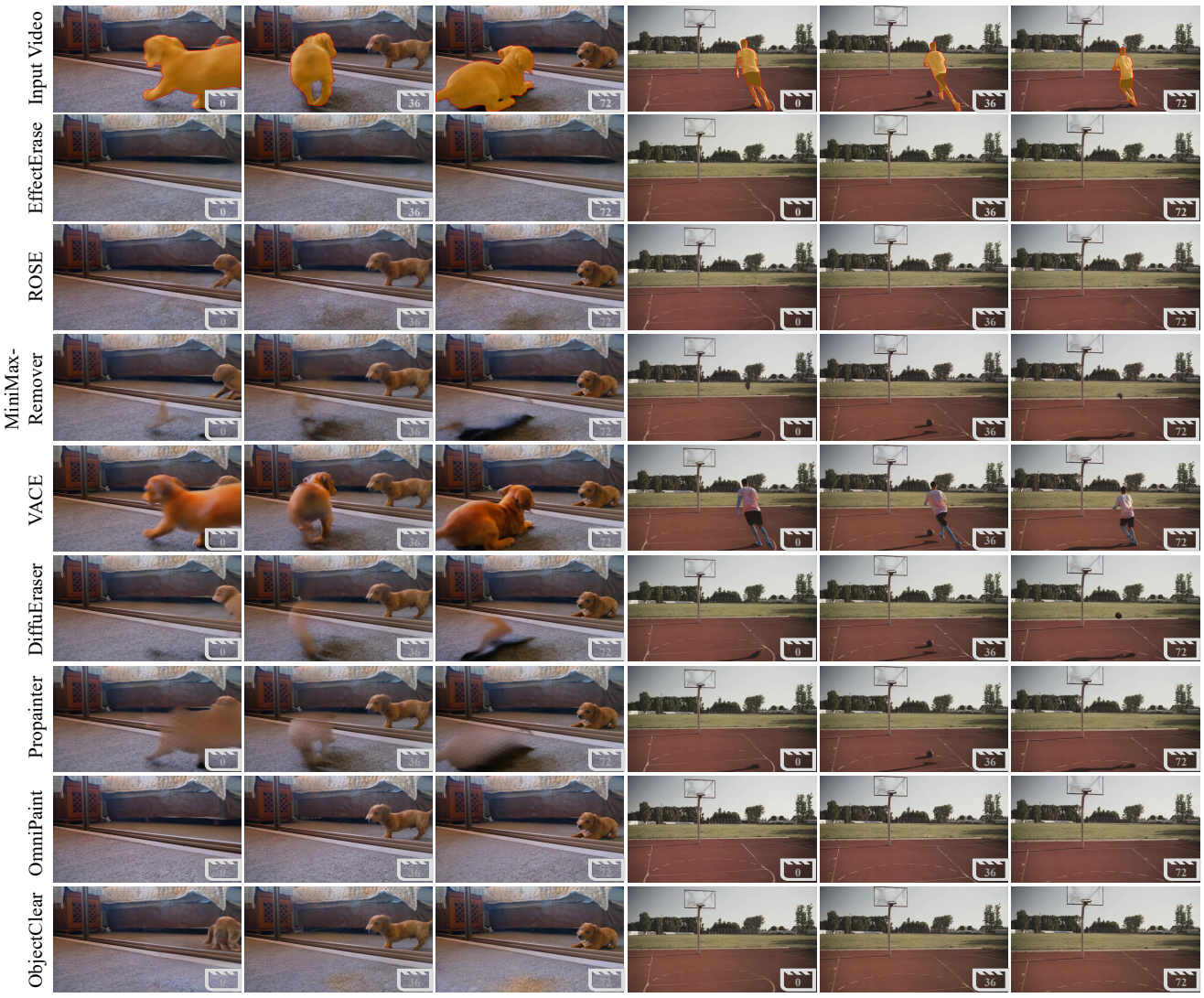

Figure 15. More comparison with state-of-the-art methods.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3

[3] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, pages 249–256, 2010. 3

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, pages 1026–1034, 2015. 3

[5] D Hendrycks. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 3

[6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 3

[7] Longtao Jiang, Zhendong Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Lei Shi, Dong Chen, and Houqiang Li. Smarteraser: Remove anything from images using masked-region guidance. In *CVPR*, 2025. 2

[8] Ruibin Li, Tao Yang, Song Guo, and Lei Zhang. Rorem: Training a robust object remover with human-in-the-loop. In *CVPR*, 2025. 2

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 3

[10] Chenxuan Miao, Yutong Feng, Jianshu Zeng, Zixiang Gao, Liu Hantang, Yunfeng Yan, Donglian Qi, Xi Chen, Bin Wang, and Hengshuang Zhao. ROSE: Remove objects with side effects in videos. In *NeurIPS*, 2025. 2, 3

[11] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 3

[12] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. In *ICLR*, 2025. 1

[13] Min-Cheol Sagong, Yoon-Jae Yeo, Seung-Won Jung, and Sung-Jea Ko. Rord: A real-world object removal dataset. In *BMVC*, 2022. 2

[14] 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. 3

[15] Runpu Wei, Zijin Yin, Shuo Zhang, Lanxiang Zhou, Xueyi Wang, Chao Ban, Tianwei Cao, Hao Sun, Zhongjiang He, Kongming Liang, et al. Omnieraser: Remove objects and their effects in images with paired video-frame data. *arXiv preprint arXiv:2501.07397*, 2025. 2

[16] Daniel Winter, Matan Cohen, Shlomi Fruchter, Yael Pritch, Alex Rav-Acha, and Yedid Hoshen. Objectdrop: Bootstrapping counterfactuals for photorealistic object removal and insertion. In *ECCV*, 2024. 2

[17] Jinrui Yang, Qing Liu, Yijun Li, Soo Ye Kim, Daniil Pakhomov, Mengwei Ren, Jianming Zhang, Zhe Lin, Cihang Xie, and Yuyin Zhou. Generative image layer decomposition with visual effects. In *CVPR*, 2025. 2

[18] Yongsheng Yu, Ziyun Zeng, Haitian Zheng, and Jiebo Luo. Omnipaint: Mastering object-oriented editing via disentangled insertion-removal inpainting. *arXiv preprint arXiv:2503.08677*, 2025. 2