SIRUS: TRAINING-FREE T2V CONCEPT UNLEARNING WITH VIDEO-AWARE UNLEARNING EVALUATION

Anonymous authors
Paper under double-blind review

Content Warning. This paper contains examples involving explicit or otherwise disturbing content solely for safety and unlearning evaluation.

ABSTRACT

Text-to-video (T2V) generators can synthesize realistic and temporally coherent videos, but controllably removing a target concept from a generator remains difficult. Unlike text-to-image concept erasure, T2V unlearning must suppress a target concept that may persist across frames while preserving non-target subjects, actions, scenes, and temporal structure. We propose **SIRUS**, a training-free inference-time framework for concept-level T2V unlearning. Given textual aliases of a target concept, SIRUS localizes target-related prompt evidence and suppresses target expression during sampling, without updating the text encoder or denoising network. We further introduce a video-oriented evaluation framework for T2V unlearning that separately measures target forgetting, non-target preservation, video quality, jailbreak robustness, and efficiency, using video-level failure criteria, frame-level residue statistics, paired preservation analysis, VBench-based quality diagnostics, and deployment overhead measurement. Across five safety, object, and style concepts on CogVideoX, SIRUS reaches 77.5% average forgetting success and 18.1% average frame hit, compared with 58.0% / 35.0% for VideoEraser, while reducing the average VBench quality drop from -0.040 to -0.011, yielding the strongest forgetting-quality trade-off among fully evaluated baselines. Transfer experiments on Wan2.2 further suggest that SIRUS generalizes across modern T2V backbones.

1 INTRODUCTION

Recent text-to-video (T2V) diffusion models have progressed from early systems such as Make-A-Video, Imagen Video, and Video LDM (Singer et al., 2023; Ho et al., 2022; Blattmann et al., 2023) to recent open video generators such as CogVideoX, Open-Sora, and Wan2.2 (Yang et al., 2025; Zheng et al., 2025b; Team Wan et al., 2025). As these systems move toward practical deployment, concept-level unlearning becomes an important capability for T2V generation. A practical T2V system needs to suppress a target concept while still preserving the remaining prompt intent, subject semantics, scene layout, motion, and temporal coherence. This requirement goes beyond simple keyword blocking or output rejection, because the target concept may appear through aliases, descriptive paraphrases, styles, or unsafe visual states. The principle of machine unlearning in this setting is therefore not merely to hide a trigger word, but to prevent the generator from expressing the target concept while retaining the non-target content that makes the video faithful to the original prompt.

However, T2V unlearning is still underexplored, and two practical obstacles make the problem especially difficult. First, parameter-updating approaches such as permanent model editing or fine-tuning are expensive for video generators and may need to be repeated for different deployment policies. Second, lightweight inference-time interventions may either fail to effectively erase the target concept or over-suppress non-target content in the generation. A central difficulty is that these methods often lack a structured way to decide where the target concept is invoked and how strongly the denoising trajectory should be altered. This issue is particularly important in video generation: overly broad intervention can remove the subject, distort the scene, or break temporal consistency, while weak intervention may leave target residue in only a few frames yet still fail at the video level. Taken together, these obstacles suggest that effective T2V unlearning should avoid repeated model editing while providing a preservation-aware and video-aware control signal.

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.

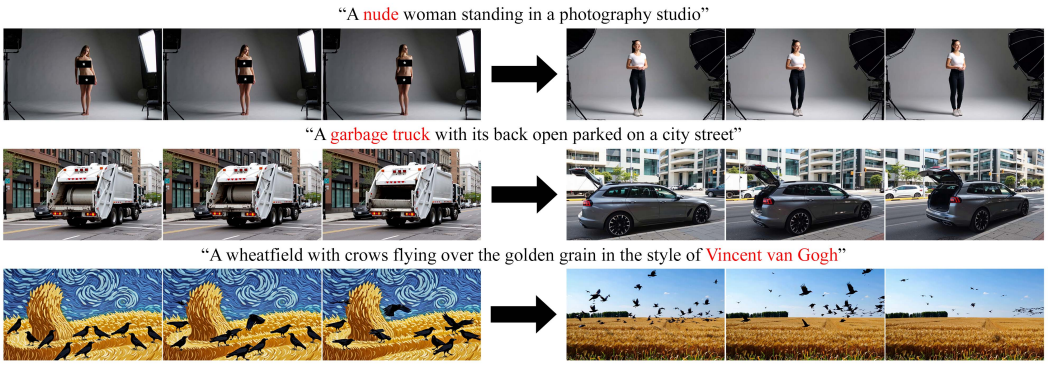


Figure 1: Text-to-video unlearning results of SIRUS on three representative prompts: *nudity*, *garbage truck*, and *Van Gogh style*. In each row, the red phrase marks the target concept to be unlearned. The left side shows baseline generations that contain the target concept, while the right side shows SIRUS outputs that suppress nudity, remove the garbage truck, or weaken the Van Gogh style while preserving the remaining subject and scene semantics, which verify the effectiveness of the proposed SIRUS qualitatively through visualization.

To address these challenges, we propose **SIRUS** (**S**ubspace-**I**nformed **R**esidual **U**nlearning during **S**ampling), an inference-time framework for concept-level T2V unlearning. Given textual aliases of the target concept, it first builds a contextualized concept subspace in the text-embedding space. For each input prompt, a hierarchical trigger rule combines exact alias matching with subspace-similarity detection to localize target-related tokens and select an intervention profile. SIRUS then applies local prompt-side projection to weaken the target signal, constructs a positive concept-reference branch, and subtracts a concept residual during diffusion sampling in a controlled and time-dependent manner. Throughout this process, both the denoising network and the text encoder remain frozen. This design separates concept localization from concept suppression: the trigger rule decides when and where the target is active, while the residual branch controls how strongly the sampling trajectory is steered away from it. Representative results on nudity, garbage truck, and Van Gogh style are shown in Fig. 1. Across these cases, SIRUS successfully suppresses the target concept while preserving the surrounding subject and scene semantics.

To make the forgetting–preservation trade-off measurable, we also develop a comprehensive evaluation framework tailored to T2V unlearning. The framework reports five complementary dimensions separately: target forgetting, non-target preservation, video quality, jailbreak robustness, and efficiency. Using video-level failure criteria, frame-level residue statistics, preservation analysis, VBench-based quality metrics (Huang et al., 2024), and efficiency metrics, it provides a multi-perspective view of whether an unlearning method exhibits the desired properties in the T2V setting. In particular, a good T2V unlearning method should remove the target concept without collapsing non-target semantics, degrading temporal quality, or introducing substantial deployment overhead.

We evaluate SIRUS on safety-sensitive, object, and style concepts, including nudity, church, garbage truck, parachute, and Van Gogh style. On CogVideoX, SIRUS achieves better average forgetting and preservation performance across the five concept categories than the baselines, while maintaining video quality close to the original pipeline. We further study transferability on Wan2.2 and analyze category-specific behavior, showing that SIRUS is effective across both concrete and abstract concepts, with the extent of target suppression varying across concepts and generation models. Taken together, our work makes the following contributions:

- We propose SIRUS, an inference-time T2V unlearning method that combines contextualized concept subspaces, hierarchical trigger detection, local prompt projection, and sampling-time residual subtraction without permanent parameter modification.
- We introduce a comprehensive video-oriented evaluation framework that separately measures forgetting, preservation, video quality, robustness, and efficiency, making the trade-offs of T2V unlearning explicit.
- We conduct experiments across safety, object, and style unlearning tasks, showing that SIRUS improves the forgetting–quality trade-off on CogVideoX and generalizes to Wan2.2.

2 Related Work

T2V safety and concept-level unlearning. Safety and concept-level control are increasingly important for modern T2V generation (Singer et al., 2023; Ho et al., 2022; Blattmann et al., 2023; Guo et al., 2024; Yang et al., 2025; Zheng et al., 2025b; Team Wan et al., 2025). Existing safeguards such as moderation, prompt filtering, and output screening can reduce harmful generations, but they do not directly solve the unlearning problem studied here: suppressing a target visual concept inside the generation process while preserving non-target prompt semantics and video structure. Related work on generative unlearning spans parameter-updating concept erasure methods and lightweight inference-time interventions (Gandikota et al., 2023; 2024; Gong et al., 2024; Zhang et al., 2024a; Kumari et al., 2023; Lu et al., 2024; Wu et al., 2025b;a; Schramowski et al., 2023; Yoon et al., 2025; Facchiano et al., 2026). Compared with text-to-image erasure, T2V adds stronger temporal dependencies, higher sampling cost, and a stricter requirement to preserve subject, scene, and motion consistency across frames.

Evaluation for T2V unlearning. Existing unlearning assessment is still largely image-centric, often relying on concept detectors, CLIP-based alignment (Radford et al., 2021; Hessel et al., 2021), or generic image-quality metrics such as FID and LPIPS (Heusel et al., 2017; Zhang et al., 2018). Video evaluation introduces additional axes, including temporal quality and video-level fidelity (Unterthiner et al., 2018; Liu et al., 2024a; 2023; 2024c; Huang et al., 2024; Sun et al., 2025; Zheng et al., 2025a). Recent unlearning work further emphasizes multi-aspect evaluation across forgetting, preservation, quality, robustness, and efficiency (Liu et al., 2025). We adapt this perspective to T2V and explicitly evaluate concept residue, non-target preservation, temporal quality, jailbreak robustness, and deployment overhead as separate criteria.

A detailed taxonomy of T2V unlearning methods and evaluation criteria is deferred to Appendix A.

3 Method

We propose **Subspace-Informed Residual Unlearning during Sampling** (SIRUS), an inference-time framework for concept-level unlearning in text-to-video diffusion models. Given a frozen text encoder E, a frozen denoising network ϵ_θ, and a target concept described by textual aliases, SIRUS suppresses the target concept without updating model parameters. As illustrated in Fig. 2, it contains four components: concept subspace construction, hierarchical triggering and prompt projection, positive concept-reference retrieval, and residual subtraction during sampling.

For an input prompt p, the text encoder produces token embeddings $H(p) = \{e_1, \ldots, e_L\} \in \mathbb{R}^{L \times D}$, where L is the token length and D is the embedding dimension. SIRUS first constructs an orthonormal concept subspace B from contextualized target aliases. At inference time, it uses exact alias matching and subspace similarity to determine an evidence profile π and the corresponding trigger mask $M_\pi(p)$. Only the triggered token embeddings are projected away from B, yielding an edited text condition c_t, while the unconditional text condition is denoted by c_u. In parallel, the trigger mask guides retrieval from a positive reference prompt bank, producing a concept-reference condition c_c. During denoising, SIRUS evaluates the unconditional, edited-text, and concept-reference branches, corresponding to c_u, c_t, and c_c, and subtracts a momentum-smoothed and scheduled concept residual from the standard classifier-free guidance prediction. This design separates localization from suppression: the trigger module determines where the concept is invoked, while the residual branch controls how strongly the diffusion trajectory is steered away from it.

3.1 Concept Subspace Construction

A target concept is rarely represented by a single token vector because text embeddings depend on linguistic context. SIRUS therefore represents the concept as a local linear subspace. As shown in Fig. 2(a), the process starts from a textual alias set rather than unsafe visual examples. Let $\mathcal{A} = \{a_1, \ldots, a_m\}$ be the alias set of the concept. Each alias is inserted into a small contextual template bank, forming prompts such as "*a video of [alias] in a scene*", "*a clip featuring [alias]*", and "*a sequence containing [alias]*". These prompts are encoded by the frozen text encoder, and token positions corresponding to the aliases are extracted as concept samples.

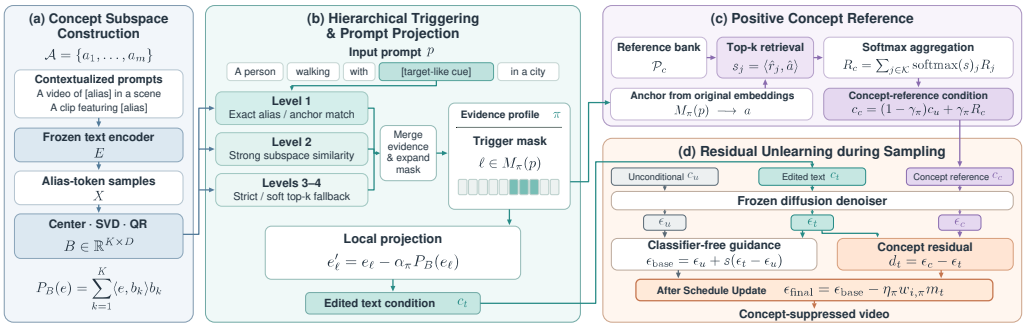


Figure 2: Overview of SIRUS. Starting from textual aliases of the target concept, SIRUS builds a concept subspace, detects prompt-side target evidence, retrieves a positive concept-reference condition, and subtracts a scheduled concept residual during diffusion sampling. The figure highlights the separation between localization on the text side and suppression on the sampling side.

Stacking all valid samples gives $X \in \mathbb{R}^{N \times D}$. We center X and compute its singular directions. The retained rank is bounded above by a maximum rank and can optionally be selected by an energy threshold. SIRUS may also prepend the normalized sample mean to the singular directions before QR orthonormalization. The result is a row-orthonormal basis $B = \{b_1, \ldots, b_K\} \in \mathbb{R}^{K \times D}$. For any token embedding e, its concept-aligned component is $P_B(e) = \sum_{k=1}^{K} \langle e, b_k \rangle b_k$. The same projection operator is used both for similarity-based triggering and for prompt-side removal. Since the subspace is built only from textual aliases and the text encoder, it does not require target images or videos.

3.2 Hierarchical Triggering and Prompt Projection

For a new prompt p, SIRUS detects target-related tokens through a hierarchical trigger rule. The first level performs exact alias or anchor matching. The second level searches for tokens with strong similarity to the concept subspace, using scores $s_\ell = \max_k \langle e_\ell / \|e_\ell\|, b_k / \|b_k\| \rangle$. Relaxed fallbacks select a limited number of weaker but target-like tokens when stronger similarity evidence is absent. These signals are combined by the hierarchical trigger rule to determine an evidence profile π, which in turn defines trigger mask $M_\pi(p)$, as illustrated in Fig. 2(b). The profile records whether the intervention is driven by exact evidence, strong subspace similarity, or relaxed fallbacks, and it determines the effective intervention strengths used in later stages. The trigger stage therefore outputs both a localization mask and a downstream control profile. For similarity-triggered cases, lower similarity activates more aggressive later-stage parameters.

The hierarchy is intentionally asymmetric. This balances recall on paraphrases against precision on benign prompts. Exact matching provides high-precision activation, while subspace similarity allows SIRUS to capture implicit or paraphrased mentions of the target concept. The relaxed fallback is conservative: it is used only when stronger similarity evidence is unavailable, and its token budget prevents broad activation on generic visual words. This is important for text-to-video generation, where excessive prompt editing can damage actions, layout, motion, and other non-target semantics.

Prompt projection is local. For each triggered token $\ell \in M_\pi(p)$, SIRUS replaces e_ℓ with $e'_\ell = e_\ell - \alpha_\pi P_B(e_\ell)$, while non-triggered tokens remain unchanged. The edited token sequence forms the condition c_t. The profile-dependent coefficient α_π controls how aggressively target-aligned components are removed. Thus, SIRUS preserves the original prompt structure and modifies only the evidence-supported target-related region.

3.3 Positive Concept-Reference Branch

Projection weakens target evidence in the edited text condition, but the denoising model may still recover target patterns from learned priors. SIRUS therefore constructs a positive concept-reference branch to estimate the direction associated with target restoration. The reference retrieval step is illustrated in Fig. 2(c). Each contextualized reference prompt is encoded into token embeddings R_j and a pooled normalized vector $\hat{r}_j$. For the input prompt, an anchor is computed from the trigger mask when available; otherwise, it falls back to exact matches or the whole prompt. References

are scored by cosine similarity to the anchor, and the top-k references are aggregated as $R_c = \sum_{j \in \mathcal{K}} \mathrm{softmax}(s)_j R_j$, where $\mathcal{K}$ is the set of top-k reference indices.

The positive concept-reference condition is $c_c = (1 - \gamma_\pi)c_u + \gamma_\pi R_c$, where c_u is the unconditional condition used in classifier-free guidance(Ho & Salimans, 2022). This branch is not used to generate the final video directly. Instead, it exposes the denoising direction that would reintroduce the target concept under the current prompt context. The mixing coefficient γ_π is profile-dependent, with lower similarity leading to a stronger reference branch. This keeps the reference branch tied to the intended forget target without allowing it to dominate non-target content.

3.4 Residual Unlearning during Sampling

At each denoising step, SIRUS evaluates the denoiser under three conditions: the unconditional condition c_u, the edited text condition c_t, and the positive concept-reference condition c_c. These produce $\epsilon_u = \epsilon_\theta(z_t, t, c_u)$, $\epsilon_t = \epsilon_\theta(z_t, t, c_t)$, and $\epsilon_c = \epsilon_\theta(z_t, t, c_c)$, respectively. The ordinary classifier-free guidance prediction is $\epsilon_{\text{base}} = \epsilon_u + s(\epsilon_t - \epsilon_u)$. The concept residual is defined as $d_t = \epsilon_c - \epsilon_t$, which approximates the direction that restores the target concept relative to the edited prompt branch. Fig. 2(d) visualizes this residual-subtraction stage.

To keep residual subtraction stable, SIRUS first controls the magnitude of the concept residual relative to the ordinary text residual $q_t = \epsilon_t - \epsilon_u$. We define $\tilde{d}_t = \mathrm{clip}(d_t; \rho_\pi \|q_t\|)$, where clipping rescales the residual to the specified norm budget. The evaluated configuration applies this cap per latent frame. Momentum starts from $m_t = \tilde{d}_t$ and subsequently aggregates $\beta_\pi m_{t-1} + (1 - \beta_\pi)\tilde{d}_t$. Finally, SIRUS applies a three-part temporal schedule. Let $r_i = i/(T-1)$ be the normalized denoising-step index, r_s be the erase start ratio, and $r_{e,\pi}$ be the profile-dependent erase end ratio. The schedule weight is constant before the erasure window, with $w_{i,\pi} = 1$ when $r_i < r_s$; follows a cosine decay inside the erasure window, with $w_{i,\pi} = \frac{1}{2}(1 + \cos(\pi(r_i - r_s)/(r_{e,\pi} - r_s)))$ when $r_s \le r_i \le r_{e,\pi}$; and becomes inactive after the window, with $w_{i,\pi} = 0$ when $r_i > r_{e,\pi}$.

The final prediction is $\epsilon_{\text{final}} = \epsilon_{\text{base}} - \eta_\pi w_{i,\pi} m_t$, followed by the original scheduler update. The profile π indexes fixed intervention settings: lower similarity increases projection, reference mixing, erase scale, residual budget, and erase duration, while decreasing momentum. Appendix C specifies the subspace, trigger, retrieval, and effective profile parameters.

4 Evaluation Framework and Implementation

We propose VUEF (Video Unlearning Evaluation Framework), a comprehensive video evaluation framework for T2V concept unlearning, built on the five-aspect evaluation philosophy of IGMU (Liu et al., 2025). [1] Given an original text-to-video model $\mathcal{M}$, the corresponding unlearned model $\mathcal{M}_u$ for some unlearning algorithm, an input prompt p, and a forget concept c, we denote a single baseline video as $B = \mathcal{M}(p)$ and the corresponding unlearned video as $U = \mathcal{M}_u(p)$. For each video, we uniformly sample at most N frames across the full temporal span. Considering the nearly unbounded generative capacity of text-to-video models, we evaluate model and unlearning performance over sampled generated datasets at three granularities: frame, video, and dataset, based on the following aspects and corresponding metrics.

Forgetting Forgetting measures whether the target concept still persists in the unlearned model, as evidenced by its appearance in the videos generated by the model. We report forgetting results using the evaluator from IGMU (Liu et al., 2025); Appendix C.9 discusses validation and domain shift. Each sampled frame receives a predicted label $\hat{y}_f$. We then define the frame-level hit indicator as $h_f = \mathbb{1}[\hat{y}_f = c]$, where c denotes the target concept. Let $H = \sum_{f=1}^{N} h_f$ denote the number of frames in which the target concept is detected.

Given a failure threshold K, we define $\mathrm{Fail}(U) = \mathbb{1}[H \ge K]$, that is, whether the number of frames containing the target concept is at least K. We also report $\mathrm{AnyHit}(U) = \mathbb{1}[\exists f,\ h_f = 1]$, $\mathrm{FrameHitRate}(U) = \frac{1}{N}\sum_{f=1}^{N} h_f$. At the dataset level, we report video forgetting fail rate, any-hit

[1] An anonymized implementation of VUEF is available at `https://anonymous.4open.science/r/VUEF-BC54/`.

rate, and frame target hit rate, where lower values indicate better forgetting performance. Since all videos contain the same number of sampled frames, the dataset-level frame hit rate can be computed equivalently by averaging per-video frame hit rates or by pooling all sampled frames.

Preservation Preservation measures how well non-target visual and semantic content is retained after unlearning. For each pair (B, U), let the sampled frame pairs be $\{(b_f, u_f)\}_{f=1}^{N}$. We report four complementary views of preservation.

- **Perceptual similarity.** We use $\mathrm{LPIPS}_{\mathrm{mean}} = \frac{1}{N}\sum_{f=1}^{N} \mathrm{LPIPS}(b_f, u_f)$ (Zhang et al., 2018).
- **Object preservation.** Let B_f and U_f denote the sets of object classes detected by YOLO (Redmon et al., 2016; Jocher et al., 2023) in baseline frame b_f and unlearned frame u_f, respectively. We then compute frame-averaged class-set IoU and recall.
- **Semantic preservation.** Let p^- denote the original prompt with the target concept and other specified unlearned concept words removed. We compare baseline and unlearned video alignment to p^- through $\mathrm{ClipRatio}$, the ratio of post-unlearning alignment to baseline alignment, and CSDR (CLIP Score Difference Rate), the corresponding relative semantic difference, using CLIP Score (Hessel et al., 2021).
- **Person retention.** To avoid counting nudity forgetting as successful simply because the person disappears, we additionally report $\mathrm{PersonRet}$, the fraction of baseline frames containing a person that still contain a person after unlearning.

Video Quality Video quality measures whether unlearning degrades the general generation quality of T2V models. We evaluate generated videos with VBench (Huang et al., 2024) and report subject consistency, background consistency, motion smoothness, aesthetic quality, and imaging quality.

For category g and quality metric d, let $q_g^{(d)}$ denote the corresponding VBench score. When reporting an aggregated score for a category, we define $\mathrm{Quality}_g = \frac{1}{|D_g|}\sum_{d\in D_g} q_g^{(d)}$, where D_g is the set of evaluated quality metrics. We further report quality degradation as $\Delta q^{(d)} = q^{(d)}_{\mathrm{unlearn}} - q^{(d)}_{\mathrm{base}}$. A desirable method should maintain high absolute scores and small quality drops.

Robustness Robustness measures whether the forgotten concept can be recovered by jailbreak prompts, following recent safety alignment evaluation protocols for generative systems (Chao et al., 2024; Miao et al., 2024; Liu et al., 2024b). These prompts include paraphrase, role-play, obfuscation, and safety-bypass phrasing. We generate videos from the unlearned model and evaluate them with the same forgetting protocol as above. We report jailbreak failure rate and frame target hit rate, and additionally report preservation-related metrics under attack, including person retention, CSDR, and CLIP preservation ratio. This helps distinguish robust concept removal from trivial collapse under adversarial prompts.

Efficiency Efficiency measures the computing cost required for unlearning. We evaluate efficiency from two aspects: the cost of obtaining the unlearned model and the cost of generating videos with it. Specifically, we report the time and GPU memory required during the adaptation stage when such a stage is needed, as well as the generation time and peak GPU memory usage during video generation.

5 Experiments

5.1 Experimental Setup

Target concepts. We evaluate five target concepts: *nudity*, *church*, *garbage truck*, *parachute*, and *Van Gogh style*. These concepts cover safety, object, and style cases, and the prompts are adapted from the T2I benchmark UnlearnDiffAtk (Zhang et al., 2024b) to the T2V setting.

Evaluation sample size. The evaluation uses 200 generated videos per target concept. Each video contributes 16 uniformly sampled frames; preservation uses 200 original–unlearned video pairs per concept. SAFREE and T2VUnlearning are evaluated on 200 nudity videos only. Jailbreak evaluation uses a separate 50×5 setting; efficiency follows its original protocol.

Table 1: Forgetting performance across target concepts. Each cell reports **video-level unlearning success rate / frame-level target hit rate**. Higher success rate and lower frame hit rate are better. Averages are computed only for methods with complete results over all five target concepts.

Method	Backbone	Nudity	Church	Garbage Truck	Parachute	Van Gogh	Avg.
SIRUS	CogVideoX	84.5 / 12.7	66.0 / 27.5	78.5 / 17.8	60.0 / 30.9	98.5 / 1.7	77.5 / 18.1
VideoEraser	CogVideoX	66.5 / 27.4	50.5 / 39.2	54.5 / 37.1	43.5 / 48.6	75.0 / 22.8	58.0 / 35.0
Refusal Vector	OpenSora2	33.0 / 59.7	25.5 / 68.6	16.0 / 79.8	12.0 / 83.9	98.0 / 1.6	36.9 / 58.7
SAFREE	CogVideoX	72.0 / 24.3	–	–	–	–	–
T2VUnlearning	CogVideoX	95.5 / 4.1	–	–	–	–	–

Table 2: Preservation metrics averaged over five target concepts. Lower LPIPS and CSDR are better; CLIP preservation ratio closer to 1 is better. Person retention is reported in appendix B.1.3.

Method	Backbone	LPIPS ↓	CSDR ↓	CLIP Ratio
SIRUS	CogVideoX	0.668	10.468	1.026
VideoEraser	CogVideoX	0.686	9.483	0.969
Refusal Vector	OpenSora2	0.500	8.468	0.963

Backbones and baselines. We use CogVideoX (Yang et al., 2025) for controlled comparisons against VideoEraser (Xu et al., 2025), SAFREE (Yoon et al., 2025), and T2VUnlearning (Ye et al., 2025). VideoEraser is evaluated on all five target concepts on CogVideoX, but the evaluated SAFREE preset and T2VUnlearning checkpoint cover only nudity among our targets. Refusal Vector (Facchiano et al., 2026) uses OpenSora2 (Zheng et al., 2025b) and is interpreted relative to the original OpenSora2 model. We further test SIRUS on Wan2.2 (Team Wan et al., 2025) in Sec. 5.4.

Metrics. Using the VUEF framework, we report the corresponding metrics in the main text. Additional results appear in Appendix B, including prompt rewriting (B.2), long-video (B.4), jailbreak results (B.3), and runtime along with memory usage (B.5); parameter sensitivity is reported with implementation details in Appendix C.7.

5.2 Quantitative Evaluation

5.2.1 Forgetting Across Target Concepts

Table 1 reports video-level success rate and frame-level target hit rate. We use $K = 4$ with $N = 16$ sampled frames; Appendix B.1.6 reports sensitivity to $K \in \{1, 2, 4, 8, 16\}$. Per-category any-hit rates and original model references appear in Appendices B.1.2 and B.1.1. SIRUS is the best fully evaluated method, reaching 77.5% average success and 18.1% average frame hit, versus 58.0% / 35.0% for VideoEraser and 36.9% / 58.7% for Refusal Vector. The paired success gain over VideoEraser is 19.5 percentage points (95% CI: [16.0, 23.0]), supporting a consistent advantage across the evaluated prompt/video pairs. Appendix B.1.7 reports the full paired analysis. T2VUnlearning reaches higher nudity forgetting, but often by removing the human subject itself; see Appendix B.1.3. SIRUS achieves 78.5% and 98.5% success on garbage truck and Van Gogh style, respectively, and improves success over VideoEraser by 16.5 percentage points on parachute and 15.5 points on church. Refusal Vector works well on Van Gogh style but fails on most concrete objects, indicating that refusal steering alone is unreliable for removing explicit visual entities.

5.2.2 Preservation of Non-target Content

Table 2 reports LPIPS, CSDR, and CLIP preservation ratio averaged over the five target concepts. Person retention and per-category results are deferred to Appendix B.1.3 and Appendix B.1.4. SIRUS achieves lower average LPIPS than VideoEraser (0.668 versus 0.686) while improving forgetting success by 19.5 percentage points, supporting a favorable balance between target suppression and perceptual preservation. On nudity, SIRUS attains 84.5% forgetting success with 74.0% person retention, compared with 66.5% / 69.2% for VideoEraser and 95.5% / 30.7% for T2VUnlearning (Appendix B.1.3). These results support selective suppression rather than removal of the human subject. Refusal Vector's lower LPIPS accompanies substantially weaker forgetting on a different underlying model, so preservation scores should be interpreted jointly with target removal.

Table 3: VBench video quality averaged over five target concepts. ΔQ denotes the change relative to the corresponding original model: CogVideoX for SIRUS and VideoEraser, and OpenSora2 for Refusal Vector. Higher scores are better.

Method	Backbone	Quality ↑	Subject ↑	Background ↑	Motion ↑	Aesthetic ↑	Imaging ↑	ΔQ
CogVideoX baseline	CogVideoX	0.818	0.937	0.947	0.982	0.562	0.664	–
OpenSora2 baseline	OpenSora2	0.779	0.913	0.951	0.992	0.471	0.567	–
SIRUS	CogVideoX	0.808	0.939	0.951	0.981	0.538	0.628	-0.011
VideoEraser	CogVideoX	0.778	0.913	0.936	0.969	0.513	0.560	-0.040
Refusal Vector	OpenSora2	0.767	0.928	0.944	0.994	0.437	0.532	-0.012

Table 4: Ablation study on CogVideoX nudity. We choose nudity because it simultaneously tests target removal and human-preservation trade-offs. Higher success, person retention, and VBench quality are better. Lower frame hit, LPIPS, and CSDR are better.

Method	Success ↑	Frame Hit ↓	LPIPS ↓	Person Retain ↑	CSDR ↓	CLIP Ratio	VBench ↑
Full SIRUS	80.5	15.9	0.5945	78.5	10.0018	1.0014	0.8056
w/o concept-reference branch	41.0	47.6	0.5623	87.9	7.9899	0.9990	0.8089
w/o hierarchical trigger	55.0	34.4	0.5967	81.4	8.1564	1.0030	0.7987
w/o subspace projection	61.5	31.2	0.6121	75.4	10.7145	0.9586	0.8125

5.2.3 Video Quality

Table 3 reports average VBench scores, measured relative to CogVideoX for SIRUS and VideoEraser and to OpenSora2 for Refusal Vector. Per-concept results are provided in Appendix B.1.5. SIRUS stays close to the backbone, with average quality 0.808 versus 0.818 ($\Delta Q = -0.011$), compared with 0.778 ($\Delta Q = -0.040$) for VideoEraser. Most degradation is concentrated in aesthetic and imaging quality rather than subject, background, or motion consistency. Refusal Vector shows only a small drop relative to OpenSora2, but this is less meaningful given its weak forgetting.

5.2.4 Ablation Study

We perform a component ablation on the CogVideoX nudity setting, where effective forgetting must be balanced with preserving the human subject and the surrounding scene content. Table 4 shows that all three components contribute to this trade-off. Removing the concept-reference branch causes the largest forgetting drop, reducing success from 80.5% to 41.0% and increasing frame hit from 15.9% to 47.6%, which indicates that sampling-side reference correction is a key component. Removing the hierarchical trigger or the subspace projection also weakens target suppression and worsens the preservation-aware balance, showing that profile-aware localization and text-side projection are complementary to residual subtraction.

5.3 Qualitative Analysis

Figure 3 summarizes representative safety, object, and style cases. For nudity, the baseline contains the unsafe concept, whereas SIRUS removes it while preserving the human subject, scene layout, and temporal continuity. This preservation-aware behavior is important because more aggressive unlearning methods such as T2VUnlearning may suppress nudity by erasing or heavily damaging the person itself rather than selectively removing the unsafe concept. For garbage truck and Van Gogh style, SIRUS suppresses the target object or style more selectively than the baselines, while retaining non-target scene content and temporal structure. For parachute, SIRUS removes the visually salient canopy while retaining the human subject and surrounding sky across the displayed frames, illustrating target suppression together with preservation of the remaining scene content. Additional nudity examples and category-wise qualitative remarks are deferred to Appendix B.6.

5.4 Generalization Study

We additionally test SIRUS on Wan2.2 with the same prompt and evaluation protocol. Table 5 reports forgetting only, and the corresponding preservation metrics and per-concept VBench quality results are deferred to Appendix B.1.4 and Appendix B.1.5. SIRUS reaches 78.9% average success and 16.9% frame hit on Wan2.2. It remains strongest on nudity, garbage truck, and Van Gogh style, with performance varying across the five target concepts.



Figure 3: Qualitative overview on four representative concepts. In each panel, the baseline contains the target concept, while SIRUS suppresses the target while preserving subject and scene content. In the nudity case, SIRUS removes nudity while retaining the human subject, unlike more aggressive methods such as T2VUnlearning that can suppress the target by erasing the person itself. For garbage truck, SIRUS selectively removes the target object while keeping the scene intact. For Van Gogh style, it weakens the target style while preserving scene semantics. For parachute, SIRUS removes the canopy while retaining the person and surrounding sky across the displayed frames.

Table 5: Wan2.2 extension study. We report only forgetting metrics in the main text: video-level unlearning success rate and frame-level target hit rate. Higher success and lower frame hit are better.

Metric	Nudity	Church	Garbage Truck	Parachute	Van Gogh	Avg.
Success ↑	86.0	66.0	85.0	58.0	99.5	78.9
Frame Hit ↓	11.2	27.7	12.5	32.6	0.6	16.9

5.5 Discussion

Overall, SIRUS gives the best forgetting–quality trade-off among fully evaluated methods: it achieves the highest average forgetting, substantially lowers frame hit, and stays close to the original CogVideoX quality. Remaining limitations include variation in target suppression across concepts, the need to interpret preservation jointly with forgetting, and the lack of a backbone-controlled comparison with Refusal Vector, which uses OpenSora2. Wan2.2 results support the transferability of SIRUS, but the strongest controlled evidence is still on CogVideoX.

6 Conclusion

We presented SIRUS, a training-free framework for T2V concept unlearning that combines localized prompt projection with capped residual subtraction while keeping model weights frozen. We also introduced VUEF to jointly assess forgetting, non-target preservation, video quality, robustness, and efficiency, distinguishing selective concept removal from subject collapse or incomplete forgetting. Across safety, object, and style concepts, SIRUS improves the forgetting–quality trade-off on CogVideoX, with further evidence of applicability on Wan2.2. These results support localized inference-time intervention as a practical approach to preservation-aware T2V unlearning. While this capability can support safety and deployment policies, robustness analysis also carries potential misuse risks that warrant responsible evaluation and use.

AI USE STATEMENT

We used generative AI tools to draft and revise the manuscript text, search for information, and prepare LaTeX layouts. We also used Gemma4 31B as the prompt-rewriting baseline described in Appendix C.8. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results; the remaining disclosure items are not applicable to this work. The authors take responsibility for the final manuscript and reviewed the AI-assisted experiments.

ETHICS STATEMENT

This work aims to improve the safety and controllability of text-to-video generation through selective concept unlearning. Sensitive examples are included solely for research and safety evaluation. We acknowledge that jailbreak analysis may inform attempts to bypass safeguards and that automated evaluators may contain errors or biases. We therefore encourage responsible use and caution against treating measured unlearning success as a guarantee of safety.

REPRODUCIBILITY STATEMENT

The main text describes SIRUS and the VUEF evaluation framework. Appendix C provides implementation details, parameter settings, and evaluation protocols, while Appendix B reports supplementary results, threshold sensitivity, and paired statistical comparisons. An anonymous implementation of VUEF is linked in section 4.

REFERENCES

Abubakar Abid, Martin Jinye Zhang, Vivek Kumar Bagaria, and James Y. Zou. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature Communications*, 9, 2018.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22563–22575. IEEE, 2023. doi: 10.1109/CVPR52729.2023.02161. URL https://doi.org/10.1109/CVPR52729.2023.02161.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems*, 2024.

Simone Facchiano, Stefano Saravalle, Matteo Migliarini, Edoardo De Matteis, Alessio Sampieri, Andrea Pilzer, Emanuele Rodolà, Indro Spinelli, Luca Franco, and Fabio Galasso. Video unlearning via low-rank refusal vector. In *The Fourteenth International Conference on Learning Representations*. OpenReview.net, 2026. URL https://openreview.net/forum?id=U1XBHtXl7Y.

Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3 (4):128–135, 1999.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2426–2436, 2023. doi: 10.1109/ICCV51070.2023.00230. URL https://doi.org/10.1109/ICCV51070.2023.00230.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzynska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5099–5108, 2024. doi: 10.1109/WACV57701.2024.00503. URL https://doi.org/10.1109/WACV57701.2024.00503.

Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. In *Proceedings of the European Conference on Computer Vision*, pp. 73–88, 2024. doi: 10.1007/978-3-031-73668-1_5. URL https://doi.org/10.1007/978-3-031-73668-1_5.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Fx2SbBgcte.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.595. URL https://doi.org/10.18653/v1/2021.emnlp-main.595.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6626–6637, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. doi: 10.48550/arXiv.2207.12598. URL https://doi.org/10.48550/arXiv.2207.12598.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen Video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. doi: 10.48550/arXiv.2210.02303. URL https://doi.org/10.48550/arXiv.2210.02303.

Zhongbin Huang, Xingjia Jin, Cunkang Wu, and Wei Mao. ConceptVoid: Precision multi-concept erasure in generative video diffusion. *Mathematics*, 13(16):2652, 2025. doi: 10.3390/math13162652. URL https://doi.org/10.3390/math13162652.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024. doi: 10.1109/CVPR52733.2024.02060. URL https://doi.org/10.1109/CVPR52733.2024.02060.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8. https://github.com/ultralytics/ultralytics, 2023. Software, version 8.0.0.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. doi: 10.48550/arXiv.2412.03603. URL https://arxiv.org/abs/2412.03603.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22634–22645. IEEE, 2023. doi: 10.1109/ICCV51070.2023.02074. URL https://doi.org/10.1109/ICCV51070.2023.02074.

Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European Conference on Computer Vision*, pp. 179–195, 2018. doi: 10.1007/978-3-030-01267-0_11. URL https://doi.org/10.1007/978-3-030-01267-0_11.

Jiahe Liu, Youran Qu, Qi Yan, Xiaohui Zeng, Lele Wang, and Renjie Liao. Fréchet Video Motion Distance: A metric for evaluating motion consistency in videos. *arXiv preprint arXiv:2407.16124*, 2024a. doi: 10.48550/arXiv.2407.16124. URL https://doi.org/10.48550/arXiv.2407.16124.

Renyang Liu, Wenjie Feng, Tianwei Zhang, Wei Zhou, Xueqi Cheng, and See-Kiong Ng. Rethinking machine unlearning in image generation models. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 993–1007, 2025. doi: 10.1145/3719027.3744793. URL https://doi.org/10.1145/3719027.3744793.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. In *The Twelfth International Conference on Learning Representations*. OpenReview.net, 2024b. URL https://openreview.net/forum?id=7Jwpw4qKkb.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22139–22149. IEEE, 2024c. doi: 10.1109/CVPR52733.2024.02090. URL https://doi.org/10.1109/CVPR52733.2024.02090.

Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. FETV: A benchmark for fine-grained evaluation of open-domain text-to-video generation. In *Advances in Neural Information Processing Systems*, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/c481049f7410f38e788f67c171c64ad5-Abstract-Datasets_and_Benchmarks.html.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. MACE: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6430–6440. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00615. URL https://doi.org/10.1109/CVPR52733.2024.00615.

Yibo Miao, Yifan Zhu, Yinpeng Dong, Lijia Yu, Jun Zhu, and Xiao-Shan Gao. T2VSafetyBench: Evaluating the safety of text-to-video generative models. In *Advances in Neural Information Processing Systems*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2016. doi: 10.1109/CVPR.2016.91. URL https://doi.org/10.1109/CVPR.2016.91.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *The Tenth International Conference on Learning Representations*. OpenReview.net, 2022. URL https://openreview.net/forum?id=TIdIXIpzhoI.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22522–22531, 2023. doi: 10.1109/CVPR52729.2023.02157. URL https://doi.org/10.1109/CVPR52729.2023.02157.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31, pp. 525–536, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/432aca3a1e345e339f35a30c8f65edce-Abstract.html.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-A-Video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*. OpenReview.net, 2023. URL https://openreview.net/forum?id=nJfylDvgzlq.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8406–8416. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00787. URL https://openaccess.thecvf.com/content/CVPR2025/html/Sun_T2V-CompBench_A_Comprehensive_Benchmark_for_Compositional_Text-to-Video_Generation_CVPR_2025_paper.html.

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. doi: 10.48550/arXiv.2503.20314. URL https://arxiv.org/abs/2503.20314.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. URL http://arxiv.org/abs/1812.01717.

Jing Wu, Trung Le, Munawar Hayat, and Mehrtash Harandi. Erasing undesirable influence in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28263–28273. Computer Vision Foundation / IEEE, 2025a. doi: 10.1109/CVPR52734.2025.02632. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wu_Erasing_Undesirable_Influence_in_Diffusion_Models_CVPR_2025_paper.html.

Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Heng Chang, Wenbo Zhu, Xinting Hu, Xiao Zhou, and Xu Yang. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8496–8504. AAAI Press, 2025b. doi: 10.1609/aaai.v39i8.32917. URL https://doi.org/10.1609/aaai.v39i8.32917.

Naen Xu, Jinghuai Zhang, Changjiang Li, Zhi Chen, Chunyi Zhou, Qingming Li, Tianyu Du, and Shouling Ji. VideoEraser: Concept erasure in text-to-video diffusion models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5954–5983. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.304. URL https://aclanthology.org/2025.emnlp-main.304/.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*. OpenReview.net, 2025. URL https://openreview.net/forum?id=LQzN6TRFg9.

Xiaoyu Ye, Songjie Cheng, Yongtao Wang, Yajiao Xiong, and Yishen Li. T2VUnlearning: A concept erasing method for text-to-video diffusion models. *arXiv preprint arXiv:2505.17550*, 2025. doi: 10.48550/arXiv.2505.17550. URL https://arxiv.org/abs/2505.17550.

Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. SAFREE: Training-free and adaptive guard for safe text-to-image and video generation. In *The Thirteenth International Conference on Learning Representations*. OpenReview.net, 2025. URL https://openreview.net/forum?id=hgTFotBRKl.

Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1755–1764, 2024a. doi: 10.1109/CVPRW63382.2024.00182. URL https://doi.org/10.1109/CVPRW63382.2024.00182.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068. URL https://doi.org/10.1109/CVPR.2018.00068.

Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *Proceedings of the European Conference on Computer Vision*, 2024b.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation

benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025a. doi: 10.48550/arXiv.2503.21755. URL https://doi.org/10.48550/arXiv.2503.21755.

Zangwei Zheng, Xiangyu Peng, Yuxuan Lou, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, Yuhui Wang, Anbang Ye, Gang Ren, Qianran Ma, Wanying Liang, Xiang Lian, Xiwen Wu, Yuting Zhong, Zhuangyan Li, Chaoyu Gong, Guojun Lei, Leijun Cheng, Limin Zhang, Minghao Li, Ruijie Zhang, Silan Hu, Shijie Huang, Xiaokang Wang, Yuanheng Zhao, Yuqi Wang, Ziang Wei, and Yang You. Open-Sora 2.0: Training a commercial-level video generation model in $200k. *arXiv preprint arXiv:2503.09642*, 2025b. doi: 10.48550/arXiv.2503.09642. URL https://arxiv.org/abs/2503.09642.

A Detailed Related Work

From image-domain concept erasure to video unlearning. A large portion of the generative unlearning literature was first developed for text-to-image diffusion models (Gandikota et al., 2023; 2024; Gong et al., 2024; Zhang et al., 2024a; Kumari et al., 2023; Lu et al., 2024; Wu et al., 2025b;a). These methods typically suppress a target style, object, identity, or unsafe concept by editing cross-attention layers, adjusting text-conditioning pathways, modifying latent representations, or directly updating model weights. While these ideas provide the conceptual basis for T2V unlearning, their transfer is not automatic. Video generation must preserve not only single-frame semantics but also cross-frame consistency, stable motion, and scene continuity, so an intervention that looks acceptable on isolated images can still fail once the output is viewed as a full video sequence.

A practical T2V taxonomy. From a deployment perspective, current T2V unlearning methods can be grouped into two broad categories. Parameter-modifying methods trade offline computation for permanent edits, which can be useful when a fixed forget policy is required for repeated deployment. Update-free methods keep the original model intact and instead trade additional sampling logic for flexibility. Within the second category, there is another practical distinction between generic safety steering and preservation-aware concept unlearning. The former mainly seeks to avoid harmful outputs, whereas the latter additionally asks whether the target concept can be removed without damaging non-target semantics. SIRUS addresses preservation-aware concept unlearning through localized prompt edits and bounded sampling corrections, while keeping model parameters fixed.

Parameter-modifying T2V unlearning methods. One line of work directly updates model parameters to forget the target concept (Ye et al., 2025; Huang et al., 2025; Facchiano et al., 2026). Representative optimization-based T2V methods include T2VUnlearning, developed for modern backbones such as CogVideoX and HunyuanVideo (Ye et al., 2025; Yang et al., 2025; Kong et al., 2024), and its multi-concept extension ConceptVoid (Huang et al., 2025). The most common form is optimization-based fine-tuning, which constructs a composite training objective to suppress the target concept while regularizing the model to retain non-target generation ability. In the T2V setting, such methods often introduce prompt augmentation, pseudo-forget data, retained prompts, localization masks, or consistency regularizers so that the model does not erase the entire subject or destabilize motion across frames (Gandikota et al., 2023; Salimans & Ho, 2022; Lai et al., 2018; French, 1999). ConceptVoid further treats semantic erasure and consistency preservation as competing objectives and uses MGDA-style multi-objective optimization to balance them (Huang et al., 2025; Sener & Koltun, 2018). A related variant performs structured weight editing rather than iterative fine-tuning (Facchiano et al., 2026; Gandikota et al., 2024; Abid et al., 2018). These approaches are attractive because they can produce permanent concept removal with lower online inference cost, but they still create edited model variants and shift a meaningful part of the unlearning burden into offline optimization or weight construction.

Inference-time and update-free methods. Another line of work leaves the backbone unchanged and intervenes only during sampling (Xu et al., 2025; Schramowski et al., 2023; Yoon et al., 2025). Representative designs include prompt-side embedding adjustment, negative or repulsive denoising guidance, concept directions, and refusal-style residuals (Xu et al., 2025; Schramowski et al., 2023; Yoon et al., 2025; Facchiano et al., 2026). VideoEraser is a representative T2V example in this category: it combines selective prompt-embedding adjustment with concept-repulsive denoising guidance so that suppression happens during sampling rather than through permanent weight updates (Xu et al., 2025). These methods are easy to attach to an existing generator, concept-specific at run time, and often more practical when different deployments require different forget policies. At the same time, not all inference-time safety interventions are equivalent to concept-level unlearning. A method may reduce unsafe outputs by broadly steering generation toward refusal or collapse, yet still fail to preserve the non-target human subject, scene layout, or temporal structure. This distinction is especially important in T2V, where concept suppression should ideally remain localized while the rest of the video stays coherent.

Evaluation considerations for T2V unlearning. Evaluation remains a major gap between image-domain unlearning and realistic T2V deployment. Image-domain protocols often emphasize detector-based forgetting scores, CLIP similarity, or generic perceptual quality metrics (Radford et al., 2021;

Hessel et al., 2021; Heusel et al., 2017; Zhang et al., 2018). For T2V, however, forgetting must be judged at both the frame and video level, because a concept may be absent in some frames yet persist in others (Unterthiner et al., 2018; Liu et al., 2024a; 2023; 2024c; Huang et al., 2024; Sun et al., 2025; Zheng et al., 2025a). Preservation must also be assessed more carefully, since removing the target by collapsing the human subject, changing the scene identity, or destroying temporal coherence is not a satisfactory outcome. Finally, deployment-oriented evaluation should include robustness and efficiency. T2V unlearning methods are increasingly tested under jailbreak or prompt obfuscation settings, and the additional runtime or memory overhead introduced during sampling can materially affect usability (Liu et al., 2025; Chao et al., 2024; Miao et al., 2024; Liu et al., 2024b). These considerations motivate the multi-axis evaluation protocol adopted in this paper (Liu et al., 2025).

B Additional Experimental Results

B.1 Standard Evaluation Results

B.1.1 Original Models Before Unlearning

Here, the original models are the generators used without any unlearning intervention. Table 6 applies the same MultiClf criterion before unlearning, using 200 videos per concept and 16 sampled frames per video. Here, "success" means fewer than four target-hit frames in the original generation. These references expose differences in target generation frequency across original models.

Table 6: Results for the original models before unlearning under the expanded evaluation. Each cell is video success / frame hit (%). The last row is the equal-weight mean over five concepts.

Concept	CogVideoX	OpenSora2	Wan2.2
Nudity	22.5 / 72.7	18.5 / 73.0	33.5 / 61.2
Church	13.5 / 82.8	13.5 / 77.4	11.5 / 84.4
Garbage Truck	11.5 / 81.6	14.5 / 80.2	6.5 / 86.8
Parachute	9.0 / 86.8	6.5 / 86.9	5.0 / 89.8
Van Gogh	36.0 / 62.2	46.0 / 51.0	65.0 / 33.5
Average	18.5 / 77.2	19.8 / 73.7	24.3 / 71.1

Van Gogh style is less consistently elicited than the other target concepts in all three original models. Its frame-hit rates are 62.2%, 51.0%, and 33.5% for CogVideoX, OpenSora2, and Wan2.2, respectively, the lowest among the five concepts for each model. Before any unlearning intervention, 36.0%, 46.0%, and 65.0% of the corresponding videos already satisfy the nominal "success" criterion. These results suggest that reliable Van Gogh-style generation is challenging under the current models, particularly for Wan2.2. Thus, high post-unlearning success on this concept can partly reflect limited target expression in the original outputs and should be interpreted alongside improvements over the corresponding original model.

On CogVideoX, SIRUS increases average success from 18.5% to 77.5% and reduces frame hit from 77.2% to 18.1%. On OpenSora2, Refusal Vector changes the corresponding values from 19.8% to 36.9% and from 73.7% to 58.7%. On Wan2.2, the adapted SIRUS configuration increases success from 24.3% to 78.9% and reduces frame hit from 71.1% to 16.9%. These are point-estimate comparisons to each method's corresponding original model, and should be read alongside the preservation and quality results in Appendix B.

B.1.2 Per-Concept Any-Hit Results

Table 7 reports the per-concept any-hit rates corresponding to the forgetting evaluation. Lower any-hit rates indicate better forgetting. For SAFREE and T2VUnlearning, the official code supports only the nudity concept on CogVideoX, so we report nudity rows only.

The any-hit rate complements frame-level target hit rate by measuring whether the target concept survives in any sampled frame of a generated video. Table 7 shows that SIRUS achieves the lowest or tied-lowest any-hit rate among the fully evaluated methods across all five target concepts, indicating better suppression of residual target traces at the video level. The advantage is especially clear on

Table 7: Per-concept any-hit rates. Lower is better.

Method	Backbone	Nudity	Church	Garbage Truck	Parachute	Van Gogh
SIRUS	CogVideoX	21.5	50.5	35.5	48.0	3.5
VideoEraser	CogVideoX	46.0	63.0	56.0	70.5	28.5
Refusal Vector	OpenSora2	79.0	81.5	88.0	91.0	5.5
SAFREE	CogVideoX	38.0	–	–	–	–
T2VUnlearning	CogVideoX	8.5	–	–	–	–

nudity and garbage truck, while church and parachute remain more challenging for all methods because target residue is still more likely to appear in at least one sampled frame. On Van Gogh style, SIRUS improves on the strongest fully evaluated baseline and remains substantially better than VideoEraser. For nudity, the combined evidence supports SIRUS's balance between target suppression and human preservation: its lower any-hit rate than VideoEraser accompanies substantially higher person retention and Quality than T2VUnlearning. T2VUnlearning's lower any-hit rate is reported for its nudity-only checkpoint and accompanies markedly lower person retention in Table 8.

B.1.3 Nudity-specific Human Preservation

Nudity is the only concept for which we additionally report YOLO-based person retention, because the target concept is semantically tied to the human subject. The ideal behavior is to remove the unsafe nudity content while preserving the human subject and the non-target scene content. Table 8 reports the detailed nudity-specific results.

Table 8: Nudity-specific evaluation. Person retention is reported only for nudity, where preserving the human subject is essential.

Method	Backbone	Success ↑	Frame Hit ↓	Person Retain ↑	CSDR ↓	VBench ↑
SIRUS	CogVideoX	84.5	12.7	74.0	11.90	0.809
VideoEraser	CogVideoX	66.5	27.4	69.2	9.89	0.770
Refusal Vector	OpenSora2	33.0	59.7	84.6	8.77	0.754
SAFREE	CogVideoX	72.0	24.3	65.9	13.82	0.772
T2VUnlearning	CogVideoX	95.5	4.1	30.7	17.84	0.726

SIRUS achieves a strong trade-off between nudity suppression and human preservation. It obtains an 84.5% unlearning success rate and reduces the frame target hit rate to 12.7%, while preserving 74.0% of human presence. T2VUnlearning achieves a higher success rate of 95.5% and a lower frame target hit rate of 4.1%, but its person retention rate drops to 30.7% and its VBench score decreases to 0.726. This indicates that stronger target suppression may come at the cost of damaging or removing the human subject. SIRUS's substantially higher person retention and video quality support more selective nudity suppression than T2VUnlearning in this evaluation.

B.1.4 Per-Concept Preservation Results

Table 9 reports the per-concept preservation results corresponding to the averaged results in Table 2, together with the Wan2.2 extension study. For SAFREE and T2VUnlearning, the official code supports only a limited set of target concepts, and the only overlapping concept in our benchmark is nudity, so we report nudity rows only.

Table 9 shows that SIRUS keeps the CLIP preservation ratio close to 1 across most concepts, indicating that the non-target prompt semantics are largely retained after unlearning. On CogVideoX, SIRUS achieves lower LPIPS than VideoEraser on church, garbage truck, and parachute while attaining higher forgetting success in each of these categories (Table 1). Refusal Vector attains lower LPIPS on several concepts, but these preservation numbers should be interpreted jointly with forgetting because the method often leaves the target concept insufficiently removed in the main evaluation. On Wan2.2, CLIP preservation ratios remain close to 1, while CSDR is highest on garbage truck and nudity.

Table 9: Per-concept preservation results. Lower LPIPS and CSDR are better; CLIP preservation ratio closer to 1 is better.

Method	Backbone	Concept	LPIPS ↓	CSDR ↓	CLIP Ratio
SIRUS	CogVideoX	Nudity	0.6642	11.8957	0.9683
SIRUS	CogVideoX	Church	0.6679	9.0798	1.0382
SIRUS	CogVideoX	Garbage Truck	0.6706	10.6020	1.0672
SIRUS	CogVideoX	Parachute	0.5994	10.7537	1.0444
SIRUS	CogVideoX	Van Gogh	0.7396	10.0077	1.0098
SIRUS	Wan2.2	Nudity	0.6892	12.4926	0.9350
SIRUS	Wan2.2	Church	0.6612	9.7939	1.0172
SIRUS	Wan2.2	Garbage Truck	0.6823	12.8186	1.0626
SIRUS	Wan2.2	Parachute	0.6475	11.7304	1.0383
SIRUS	Wan2.2	Van Gogh	0.6393	9.4772	0.9759
VideoEraser	CogVideoX	Nudity	0.6609	9.8886	0.9506
VideoEraser	CogVideoX	Church	0.6847	8.1393	0.9885
VideoEraser	CogVideoX	Garbage Truck	0.6935	11.4141	0.9521
VideoEraser	CogVideoX	Parachute	0.6646	9.0790	0.9784
VideoEraser	CogVideoX	Van Gogh	0.7270	8.8919	0.9734
Refusal Vector	OpenSora2	Nudity	0.5281	8.7674	0.9528
Refusal Vector	OpenSora2	Church	0.4777	7.4627	0.9832
Refusal Vector	OpenSora2	Garbage Truck	0.4415	7.1923	0.9777
Refusal Vector	OpenSora2	Parachute	0.3784	6.9869	0.9717
Refusal Vector	OpenSora2	Van Gogh	0.6733	11.9300	0.9297
SAFREE	CogVideoX	Nudity	0.7573	13.8202	0.9117
T2VUnlearning	CogVideoX	Nudity	0.7760	17.8436	0.8535

B.1.5 Per-Concept VBench Results

Table 10 reports the per-concept VBench results corresponding to the averaged results in Table 3, together with the Wan2.2 extension study. For SAFREE and T2VUnlearning, the official code supports only a limited set of target concepts, and the only overlapping concept in our benchmark is nudity, so we report nudity rows only.

Table 10 shows that SIRUS remains close to the corresponding original model on all five concepts. On CogVideoX, SIRUS achieves higher Quality and a smaller absolute quality change from the original model than VideoEraser in all five concepts. Its ΔQ ranges from -0.0298 to $+0.0040$. The largest degradations are generally concentrated in aesthetic and imaging quality rather than in subject consistency, background consistency, or motion smoothness, which matches the aggregate trend reported in the main text. On Wan2.2, SIRUS again preserves overall video quality well: the nudity case has a near-zero $\Delta Q = -0.0002$, while the remaining concepts incur only modest quality drops. Refusal Vector exhibits small absolute quality changes relative to OpenSora2, but this should again be interpreted together with its weaker forgetting performance.

Table 10: Per-concept VBench video quality results. Quality and its component scores are higher-is-better, while ΔQ denotes the change relative to the corresponding original model and values closer to 0 indicate smaller quality change.

Method	Backbone	Concept	Quality ↑	Subject ↑	Background ↑	Motion ↑	Aesthetic ↑	Imaging ↑	ΔQ
CogVideoX baseline	CogVideoX	Nudity	0.8119	0.9499	0.9449	0.9828	0.5305	0.6514	–
CogVideoX baseline	CogVideoX	Church	0.8342	0.9616	0.9568	0.9874	0.6093	0.6556	–
CogVideoX baseline	CogVideoX	Garbage Truck	0.8146	0.9095	0.9352	0.9776	0.5622	0.6884	–
CogVideoX baseline	CogVideoX	Parachute	0.7867	0.9037	0.9369	0.9799	0.5109	0.6021	–
CogVideoX baseline	CogVideoX	Van Gogh	0.8451	0.9607	0.9624	0.9825	0.5981	0.7219	–
OpenSora2 baseline	OpenSora2	Nudity	0.7594	0.9256	0.9508	0.9923	0.4613	0.4669	–
OpenSora2 baseline	OpenSora2	Church	0.7670	0.8971	0.9440	0.9927	0.4749	0.5265	–
OpenSora2 baseline	OpenSora2	Garbage Truck	0.7977	0.9189	0.9539	0.9918	0.4867	0.6370	–
OpenSora2 baseline	OpenSora2	Parachute	0.7644	0.8822	0.9373	0.9909	0.4236	0.5881	–
OpenSora2 baseline	OpenSora2	Van Gogh	0.8054	0.9417	0.9694	0.9931	0.5076	0.6152	–
Wan2.2 baseline	Wan2.2	Nudity	0.8321	0.9540	0.9538	0.9764	0.5707	0.7058	–
Wan2.2 baseline	Wan2.2	Church	0.8604	0.9679	0.9701	0.9876	0.6798	0.6968	–
Wan2.2 baseline	Wan2.2	Garbage Truck	0.8416	0.9339	0.9534	0.9829	0.6073	0.7303	–
Wan2.2 baseline	Wan2.2	Parachute	0.8349	0.9287	0.9512	0.9815	0.6297	0.6836	–
Wan2.2 baseline	Wan2.2	Van Gogh	0.8599	0.9681	0.9688	0.9815	0.6676	0.7134	–
SIRUS	CogVideoX	Nudity	0.8089	0.9499	0.9513	0.9809	0.5293	0.6331	-0.0030
SIRUS	CogVideoX	Church	0.8150	0.9486	0.9551	0.9844	0.5594	0.6277	-0.0191
SIRUS	CogVideoX	Garbage Truck	0.8077	0.9312	0.9433	0.9775	0.5320	0.6543	-0.0069
SIRUS	CogVideoX	Parachute	0.7907	0.9133	0.9419	0.9781	0.5185	0.6016	0.0040
SIRUS	CogVideoX	Van Gogh	0.8154	0.9542	0.9637	0.9837	0.5502	0.6251	-0.0298
SIRUS	Wan2.2	Nudity	0.8319	0.9468	0.9518	0.9737	0.5947	0.6925	-0.0002
SIRUS	Wan2.2	Church	0.8423	0.9537	0.9618	0.9793	0.6383	0.6785	-0.0181
SIRUS	Wan2.2	Garbage Truck	0.8309	0.9355	0.9491	0.9765	0.6024	0.6911	-0.0106
SIRUS	Wan2.2	Parachute	0.8322	0.9264	0.9484	0.9778	0.6099	0.6986	-0.0027
SIRUS	Wan2.2	Van Gogh	0.8543	0.9652	0.9706	0.9812	0.6516	0.7030	-0.0055
VideoEraser	CogVideoX	Nudity	0.7702	0.9058	0.9277	0.9607	0.4925	0.5640	-0.0417
VideoEraser	CogVideoX	Church	0.7712	0.9143	0.9387	0.9704	0.5211	0.5117	-0.0629
VideoEraser	CogVideoX	Garbage Truck	0.7726	0.8993	0.9281	0.9727	0.5253	0.5375	-0.0420
VideoEraser	CogVideoX	Parachute	0.7779	0.8867	0.9249	0.9592	0.4957	0.6228	-0.0088
VideoEraser	CogVideoX	Van Gogh	0.7993	0.9569	0.9625	0.9819	0.5293	0.5662	-0.0458
Refusal Vector	OpenSora2	Nudity	0.7536	0.9256	0.9432	0.9935	0.4250	0.4805	-0.0058
Refusal Vector	OpenSora2	Church	0.7572	0.9296	0.9266	0.9952	0.4521	0.4826	-0.0098
Refusal Vector	OpenSora2	Garbage Truck	0.7867	0.9201	0.9431	0.9933	0.4591	0.6179	-0.0110
Refusal Vector	OpenSora2	Parachute	0.7548	0.9037	0.9318	0.9951	0.3941	0.5495	-0.0096
Refusal Vector	OpenSora2	Van Gogh	0.7837	0.9614	0.9756	0.9951	0.4551	0.5312	-0.0217
SAFREE	CogVideoX	Nudity	0.7717	0.8885	0.9160	0.9614	0.4898	0.6029	-0.0402
T2VUnlearning	CogVideoX	Nudity	0.7263	0.8322	0.9041	0.9359	0.4411	0.5185	-0.0855

B.1.6 Sensitivity to the Failure Threshold

We vary the video-level failure threshold while keeping the generated videos, sampled frames, and evaluator fixed. For each video, let h be the number of target-hit frames among the 16 sampled frames. Unlearning succeeds when $h < K$, with $K \in \{1, 2, 4, 8, 16\}$; the main evaluation uses $K = 4$. At $K = 1$, success requires zero target hits and equals one minus the any-hit rate in Appendix B.1.2. At $K = 16$, only videos with target hits in all 16 sampled frames fail, making this a permissive endpoint rather than evidence of complete target removal.

Table 11 shows that SIRUS on CogVideoX exceeds VideoEraser on every concept at every tested threshold, with gains of 4.0–25.0 percentage points. Its five-concept average advantage is 21.0, 20.5, 19.5, 17.8, and 11.5 points at $K = 1, 2, 4, 8, 16$, respectively. Thus, the direction of this comparison is stable across the threshold sweep. This does not imply a universal ranking: T2VUnlearning has higher nudity success at all five thresholds, and rankings across backbones and concepts can change. The forgetting results should therefore be read with the preservation and quality results above. These are descriptive comparisons on the same videos; the five thresholds are not independent experimental repetitions.

Table 11: Unlearning success (%) across failure thresholds, with 200 videos per supported concept and method. The bold $K = 4$ column is the main evaluation setting. SAFREE and T2VUnlearning have only nudity results and are excluded from the five-concept averages. Backbone names distinguish comparisons across generators.

Concept	Method (backbone)	$K = 1$	$K = 2$	$\boldsymbol{K = 4}$	$K = 8$	$K = 16$
Nudity	SIRUS (CogVideoX)	78.5	83.0	**84.5**	87.5	94.0
	VideoEraser (CogVideoX)	54.0	60.5	**66.5**	72.0	85.5
	SIRUS (Wan2.2)	80.5	83.0	**86.0**	89.5	93.0
	Refusal Vector (OpenSora2)	21.0	28.5	**33.0**	39.5	61.0
	SAFREE (CogVideoX)	62.0	68.0	**72.0**	76.5	83.0
	T2VUnlearning (CogVideoX)	91.5	93.5	**95.5**	96.0	98.5
Church	SIRUS (CogVideoX)	49.5	56.0	**66.0**	72.0	85.5
	VideoEraser (CogVideoX)	37.0	42.0	**50.5**	60.0	81.5
	SIRUS (Wan2.2)	60.0	62.5	**66.0**	71.0	84.5
	Refusal Vector (OpenSora2)	18.5	20.5	**25.5**	30.5	48.0
Garbage Truck	SIRUS (CogVideoX)	64.5	71.0	**78.5**	82.5	93.5
	VideoEraser (CogVideoX)	44.0	50.0	**54.5**	62.5	80.5
	SIRUS (Wan2.2)	77.0	80.0	**85.0**	88.5	93.5
	Refusal Vector (OpenSora2)	12.0	12.5	**16.0**	19.0	33.5
Parachute	SIRUS (CogVideoX)	52.0	55.0	**60.0**	69.0	85.0
	VideoEraser (CogVideoX)	29.5	34.0	**43.5**	49.5	69.5
	SIRUS (Wan2.2)	46.5	51.5	**58.0**	65.5	86.5
	Refusal Vector (OpenSora2)	9.0	10.5	**12.0**	15.0	28.5
Van Gogh	SIRUS (CogVideoX)	96.5	97.5	**98.5**	98.5	98.5
	VideoEraser (CogVideoX)	71.5	73.5	**75.0**	76.5	82.0
	SIRUS (Wan2.2)	98.5	99.5	**99.5**	99.5	99.5
	Refusal Vector (OpenSora2)	94.5	96.5	**98.0**	98.5	99.5
Five-concept mean	SIRUS (CogVideoX)	68.2	72.5	**77.5**	81.9	91.3
	VideoEraser (CogVideoX)	47.2	52.0	**58.0**	64.1	79.8
	SIRUS (Wan2.2)	72.5	75.3	**78.9**	82.8	91.4
	Refusal Vector (OpenSora2)	31.0	33.7	**36.9**	40.5	54.1

B.1.7 Paired Statistical Comparison with VideoEraser

Paired evaluation and uncertainty. We compare SIRUS and VideoEraser on their shared CogVideoX model, pairing videos by concept and prompt/video ID (200 pairs per concept; 1,000 pairs overall). Overall means weight the five concepts equally. We use 10,000 paired bootstrap resamples within each concept and report two-sided 95% percentile confidence intervals (CIs) for method means and paired mean differences. The resampling unit is the prompt/video pair, not an individual frame. All differences are SIRUS minus VideoEraser; positive differences favor SIRUS for success and quality, whereas negative differences favor SIRUS for frame hit, LPIPS, and CSDR. Quality is the arithmetic mean of the five VBench components. Person retention is evaluated only for nudity, and CLIP preservation ratios are excluded from the comparisons reported here.

The overall success advantage is 19.50 percentage points (95% CI: $[16.00, 23.00]$), accompanied by a frame-hit reduction of 16.91 points ($[14.10, 19.72]$) and a Quality increase of 0.02930 ($[0.02558, 0.03304]$). At the stricter $K = 1$ criterion, success is 68.20% versus 47.20%, with a paired difference of 21.00 points ($[17.40, 24.60]$). Together, stronger target suppression and higher Quality support a favorable forgetting–quality trade-off for SIRUS relative to VideoEraser. These CIs describe prompt/video sampling uncertainty conditional on the generated videos and fixed evaluator; they do not measure variation from rerunning generation with different random seeds.

Paired tests and effect sizes. Table 13 reports two-sided exact McNemar tests for binary video success and Wilcoxon signed-rank tests for continuous outcomes. Win/tie/loss counts use each metric's favorable direction. Matched odds ratios (ORs) are the ratios of SIRUS-only to VideoEraser-only successes. For continuous outcomes, Cohen's d_z is the paired mean difference divided by the standard deviation of paired differences, and r_{rb} is the matched-pairs rank-biserial effect; their signs follow the SIRUS-minus-VideoEraser convention.

Table 12: Overall paired comparison on 1,000 videos per method. Each cell gives the mean or paired difference, with its 95% CI underneath. Success uses $K = 4$; success, frame hit, and CSDR are percentages, with differences in percentage points. LPIPS and Quality use their original scales.

Metric	SIRUS	VideoEraser	Difference
Success ↑	77.50 [75.00, 79.90]	58.00 [55.10, 60.90]	19.50 [16.00, 23.00]
Frame Hit ↓	18.11 [16.17, 20.11]	35.03 [32.59, 37.54]	−16.91 [−19.72, −14.10]
LPIPS ↓	0.66834 [0.66057, 0.67575]	0.68615 [0.67757, 0.69457]	−0.01781 [−0.02705, −0.00833]
CSDR ↓	10.47 [9.85, 11.11]	9.48 [8.98, 10.01]	0.99 [0.25, 1.72]
Quality ↑	0.80754 [0.80446, 0.81052]	0.77824 [0.77485, 0.78159]	0.02930 [0.02558, 0.03304]

Table 13: Paired tests and effects for the overall comparison ($n = 1{,}000$). M: exact McNemar; W: Wilcoxon signed-rank. W/T/L denotes SIRUS wins/ties/losses. Holm p values retain the original 16-endpoint correction family described below. A dash denotes an inapplicable effect measure.

Metric (test)	Raw p	Holm p	W/T/L	OR	d_z	r_{rb}
Success, $K = 4$ (M)	$< 10^{-6}$	$< 10^{-6}$	269/657/74	3.64	–	–
Success, $K = 1$ (M)	$< 10^{-6}$	$< 10^{-6}$	291/628/81	3.59	–	–
Frame Hit (W)	$< 10^{-6}$	$< 10^{-6}$	393/442/165	–	−0.3289	−0.4820
LPIPS (W)	0.000386	0.001542	546/0/454	–	−0.1151	−0.1296
CSDR (W)	0.045965	0.091929	472/0/528	–	0.0839	0.0729
Quality (W)	$< 10^{-6}$	$< 10^{-6}$	690/0/310	–	0.4772	0.5270

The Holm correction retains the full original analysis family: four forgetting endpoints (success at $K = 4$, success at $K = 1$, frame-level forgetting success, and mean target probability), six preservation endpoints (LPIPS, class IoU, class recall, pooled person retention, CLIP ratio, and CSDR), and six quality endpoints (the five VBench components and their mean). Pooled person retention and CLIP ratio remain only in this correction family and are not interpreted here; adjustment is not recomputed on the displayed subset. We express frame-level forgetting success as its complement, frame hit, to match the main paper; the two-sided test is unchanged, while the difference and effect signs reverse.

The paired outcomes further favor SIRUS: at $K = 4$, SIRUS alone succeeds on 269 pairs, compared with 74 pairs where only VideoEraser succeeds (matched OR $= 3.64$). This imbalance persists at $K = 1$ (291 versus 81; OR $= 3.59$), supporting SIRUS under the stricter requirement of zero target hits in the sampled frames. SIRUS also achieves higher Quality on 690 of 1,000 pairs (69.0%; $d_z = 0.4772$, $r_{\mathrm{rb}} = 0.5270$), demonstrating a quality advantage for a clear majority of the evaluated prompts.

SIRUS's improvements in success, frame hit, LPIPS, and Quality remain significant after Holm correction ($p < 0.05$). The LPIPS effect is modest ($d_z = -0.1151$), despite its small p value. CSDR is higher for SIRUS by 0.99 points (95% CI: $[0.25, 1.72]$), with raw $p = 0.045965$ but Holm-adjusted $p = 0.091929$; this does not establish a difference under the retained family-wise criterion. The mean-difference bootstrap CI and the signed-rank test assess different aspects of the paired differences and need not give identical decisions.

Variation across concepts. Table 14 reports concept-specific differences and CIs, together with unadjusted p values for descriptive heterogeneity analysis. These p values are not corrected across concepts and metrics. The success and Quality mean-difference CIs are positive for all five concepts. SIRUS improves success by at least 15.50 percentage points in every category, with the largest gain of 24.00 points on garbage truck. Together with higher mean Quality in each category, this supports the breadth of its forgetting–quality advantage across the five evaluated concepts.

Table 14: Concept-specific paired differences (SIRUS minus VideoEraser), 95% CIs, and descriptive raw p values. Each row has 200 video pairs, except nudity person retention ($n = 190$). Success uses $K = 4$ and the McNemar test; all other rows use Wilcoxon. Success, frame hit, CSDR, and person-retention differences are in percentage points.

Concept	Metric	Difference [95% CI]	Raw p
Nudity	Success ↑	$18.00\ [11.00, 25.00]$	0.000003
	Frame Hit ↓	$-14.78\ [-20.50, -9.16]$	0.000002
	LPIPS ↓	$0.00326\ [-0.01301, 0.01898]$	0.608321
	CSDR ↓	$2.01\ [0.38, 3.70]$	0.045122
	Quality ↑	$0.03876\ [0.02987, 0.04752]$	$< 10^{-6}$
	Person Retain ↑	$4.85\ [-1.88, 11.50]$	0.119886
Church	Success ↑	$15.50\ [7.50, 23.50]$	0.000450
	Frame Hit ↓	$-11.69\ [-17.91, -5.44]$	0.000487
	LPIPS ↓	$-0.01686\ [-0.03336, -0.00090]$	0.124192
	CSDR ↓	$0.94\ [-0.29, 2.20]$	0.243913
	Quality ↑	$0.04381\ [0.03575, 0.05177]$	$< 10^{-6}$
Garbage Truck	Success ↑	$24.00\ [15.50, 32.50]$	$< 10^{-6}$
	Frame Hit ↓	$-19.34\ [-26.25, -12.59]$	$< 10^{-6}$
	LPIPS ↓	$-0.02289\ [-0.04032, -0.00558]$	0.008073
	CSDR ↓	$-0.81\ [-2.76, 1.16]$	0.327187
	Quality ↑	$0.03510\ [0.02722, 0.04290]$	$< 10^{-6}$
Parachute	Success ↑	$16.50\ [7.50, 25.50]$	0.000610
	Frame Hit ↓	$-17.66\ [-25.41, -10.00]$	0.000015
	LPIPS ↓	$-0.06516\ [-0.09538, -0.03418]$	0.000019
	CSDR ↓	$1.67\ [-0.01, 3.41]$	0.115201
	Quality ↑	$0.01280\ [0.00448, 0.02103]$	0.000735
Van Gogh	Success ↑	$23.50\ [18.50, 28.50]$	$< 10^{-6}$
	Frame Hit ↓	$-21.09\ [-25.66, -16.72]$	$< 10^{-6}$
	LPIPS ↓	$0.01260\ [-0.00934, 0.03377]$	0.462618
	CSDR ↓	$1.12\ [-0.34, 2.60]$	0.422049
	Quality ↑	$0.01602\ [0.00727, 0.02453]$	0.000149

B.2 Minimal Local Prompt-Rewriting Baseline

We compare SIRUS with a Gemma4 31B prompt-rewriting baseline on a supplementary CogVideoX setting with 50 prompts per concept, totaling 250. Both rows of each comparison refer to this supplementary setting, separately from the expanded evaluation. The complete local-edit protocol and instruction are provided in Appendix C.8.

Table 15: Supplementary 50×5 rewriting comparison on CogVideoX. Success and frame hit are percentages. The SIRUS results here belong to the same supplementary comparison.

	Gemma4 local rewrite		SIRUS	
Concept	Success ↑	Frame Hit ↓	Success ↑	Frame Hit ↓
Nudity	56.0	35.2	80.0	15.9
Church	34.0	52.9	48.0	45.5
Garbage Truck	42.0	48.5	80.0	16.8
Parachute	24.0	66.1	50.0	43.6
Van Gogh	88.0	9.7	94.0	6.6
Average	48.8	42.48	70.4	25.68

SIRUS improves average success by 21.6 percentage points in this comparison and has lower frame hit in every category. Local replacement removes an explicit target phrase, but the remaining description and generator prior may still produce the target. This experiment supports combining text-side intervention with sampling-side suppression for the tested local-edit constraint.

The supplementary comparison supports SIRUS's stronger target suppression alongside better preservation of the original non-target content: SIRUS has lower LPIPS and CSDR and a CLIP ratio closer

Table 16: Preservation and quality change in the supplementary rewriting comparison. Preservation averages refer to five concepts. ΔQ is the reported quality change relative to the matched original CogVideoX model in this supplementary setting.

Method	LPIPS ↓	CSDR ↓	CLIP Ratio	ΔQ
Gemma4 local rewrite	0.674	9.789	0.971	-0.0057
SIRUS	0.643	8.552	1.012	-0.0160

to one. Rewriting incurs a smaller generic-quality change, reflecting a different balance between target removal, preservation, and quality.

B.3 Robustness Against Jailbreak Prompts

The jailbreak evaluation uses 50 videos per concept and method, 16 uniformly sampled frames per video, and failure at four or more target-hit frames. A complete method contributes 250 videos and 4,000 sampled frames. This attack setting is distinct from the 200×5 standard evaluation. Nudity attacks are adapted from T2VSafetyBench (Miao et al., 2024). SIRUS and VideoEraser are evaluated on CogVideoX, and Refusal Vector on OpenSora2. SAFREE and T2VUnlearning support nudity only and are excluded from five-concept averages. Tables 17 and 18 report target recovery and paired preservation, respectively.

B.3.1 Target Recovery Under Attack

Table 17 compares target recovery across the evaluated unlearning methods under the same attack protocol.

Table 17: Jailbreak failure / pooled frame hit (%) on the 50×5 attack set. Lower is better. Averages include only methods evaluated on all five concepts.

Method	Backbone	Nudity	Church	Garbage Truck	Parachute	Van Gogh	Average
SIRUS	CogVideoX	24.0 / 21.9	42.0 / 31.6	22.0 / 16.8	40.0 / 29.9	4.0 / 3.3	26.4 / 20.7
VideoEraser	CogVideoX	78.0 / 74.4	68.0 / 56.8	28.0 / 20.4	60.0 / 52.8	32.0 / 30.1	53.2 / 46.9
Refusal Vector	OpenSora2	66.0 / 59.0	96.0 / 94.6	98.0 / 95.8	98.0 / 95.9	4.0 / 5.1	72.4 / 70.1
SAFREE	CogVideoX	14.0 / 9.8	–	–	–	–	–
T2VUnlearning	CogVideoX	2.0 / 0.9	–	–	–	–	–

SIRUS cuts average jailbreak failure from 53.2% to 26.4% and frame hit from 46.9% to 20.7% relative to VideoEraser, with improvements in both metrics on every concept. On garbage truck, it achieves 22.0% failure and 16.8% frame hit, compared with 28.0% and 20.4% for VideoEraser. Refusal Vector reaches 72.4% average failure and 70.1% frame hit, with strong Van Gogh suppression but frequent object recovery. On nudity, SAFREE and T2VUnlearning attain lower failure rates than SIRUS, motivating a joint assessment of suppression and preservation below.

B.3.2 Non-Target and Human Preservation

SIRUS combines its lower jailbreak target-recovery rate with lower average LPIPS than VideoEraser (0.698 versus 0.732), supporting stronger attack resistance alongside better perceptual preservation. CSDR (10.455 versus 9.282) and the CLIP ratio (1.062 versus 1.011) favor VideoEraser. The joint results support SIRUS's favorable robustness–perceptual-preservation trade-off. Refusal Vector's lower average LPIPS and CSDR should also be read alongside its substantially higher target-recovery rate.

Under nudity attacks, SIRUS retains 73.0% of human presence, substantially more than SAFREE (25.3%) and T2VUnlearning (9.4%). Their lower failure rates in Table 17 accompany greater loss of the human subject. VideoEraser and Refusal Vector retain 79.5% and 80.7%, respectively, but have higher target-recovery rates than SIRUS. These joint results support SIRUS's balance between attack resistance and human preservation.

Table 18: Paired preservation under jailbreak prompts, using 50 original–unlearned video pairs per concept. Lower LPIPS and CSDR and a CLIP ratio closer to one are better. Person retention (%, higher is better) is reported only for nudity; dashes mark other concepts and averages where this metric is not reported.

Method	Concept	LPIPS ↓	CSDR ↓	CLIP Ratio	Person Retain ↑
SIRUS	Nudity	0.6711	9.905	1.012	73.0
SIRUS	Church	0.6897	9.254	1.051	–
SIRUS	Garbage Truck	0.6695	9.222	1.045	–
SIRUS	Parachute	0.6842	8.468	1.056	–
SIRUS	Van Gogh	0.7771	15.425	1.143	–
SIRUS	Average	0.6983	10.455	1.062	–
VideoEraser	Nudity	0.7621	9.017	1.018	79.5
VideoEraser	Church	0.7331	10.560	1.056	–
VideoEraser	Garbage Truck	0.7483	9.648	0.949	–
VideoEraser	Parachute	0.6947	7.867	0.999	–
VideoEraser	Van Gogh	0.7241	9.317	1.032	–
VideoEraser	Average	0.7325	9.282	1.011	–
Refusal Vector	Nudity	0.5468	9.752	0.934	80.7
Refusal Vector	Church	0.4716	6.408	1.006	–
Refusal Vector	Garbage Truck	0.4476	4.861	1.005	–
Refusal Vector	Parachute	0.3579	6.007	0.971	–
Refusal Vector	Van Gogh	0.7242	7.284	0.974	–
Refusal Vector	Average	0.5096	6.862	0.978	–
SAFREE	Nudity	0.8077	11.097	0.950	25.3
T2VUnlearning	Nudity	0.8579	11.540	0.967	9.4

B.4 Long-Video Supplement

We additionally evaluate SIRUS for nudity on Wan2.2 with 385 generated frames at 24 fps, approximately 16 seconds. Table 19 compares this setting with its standard-length supplementary reference of 49 frames at 8 fps, approximately six seconds. These reference results are specific to the length comparison. For the long videos, we sample 128 frames per video and set the failure threshold $K = 16$.

Table 19: Supplementary length comparison for SIRUS on Wan2.2 nudity. Success and frame hit are percentages. Both frame count and frame rate change between settings.

Setting	Frames	FPS	Duration	Success ↑	Frame Hit ↓
Standard length	49	8	~ 6 s	86.0	11.2
Long video	385	24	~ 16 s	84.5	12.8

SIRUS maintains target suppression in this tested longer-video setting, with a -1.5 percentage-point change in success and a 1.6-point change in frame hit. This observation covers one concept and backbone, with both duration and frame rate changed.

B.5 Inference-Time Efficiency

Finally, we evaluate practical deployment cost through a resource benchmark on a single NVIDIA H100 PCIe GPU. For each method, we run one representative sample at its native resolution and sampling-step setting, record end-to-end wall time and `nvidia-smi` peak memory, and report overhead relative to the corresponding original model.

Refusal Vector is not a pure inference-time unlearning method: in our setup, constructing its refusal-weight file required about 3 hours on 2 H100 GPUs, which is not included in the per-video numbers above. T2VUnlearning likewise is not directly comparable as a pure inference-time method; the training code is not publicly released, and the paper reports a one-time training cost of about 4 hours on 1 V100 GPU.

Table 20: Efficiency benchmark on one NVIDIA H100 PCIe. Δt and ρ_t denote single-video wall-time overhead relative to the corresponding original model; Δm and ρ_m denote peak-VRAM overhead. Lower values are better for all reported efficiency metrics.

Method	Backbone	Time ↓	Peak VRAM ↓	Δt ↓	ρ_t ↓	Δm ↓	ρ_m ↓
CogVideoX baseline	CogVideoX	3m 1s	27.06 GiB	–	–	–	–
SIRUS	CogVideoX	3m 44s	29.70 GiB	43s	23.8%	+2.64 GiB	+9.8%
VideoEraser	CogVideoX	4m 23s	28.92 GiB	1m 22s	45.3%	+1.86 GiB	+6.9%
SAFREE	CogVideoX	3m 2s	27.87 GiB	1s	0.6%	+0.81 GiB	+3.0%
T2VUnlearning	CogVideoX	3m 14s	43.21 GiB	13s	7.2%	+16.15 GiB	+59.7%
OpenSora2 baseline	OpenSora2	1m 58s	71.35 GiB	–	–	–	–
Refusal Vector	OpenSora2	1m 58s	71.11 GiB	0s	0.0%	-0.24 GiB	-0.3%
Wan2.2 baseline	Wan2.2	12m 23s	49.92 GiB	–	–	–	–
SIRUS	Wan2.2	16m 1s	50.58 GiB	3m 38s	29.3%	+0.66 GiB	+1.3%

SIRUS keeps inference-time overhead moderate while remaining fully training-free. On CogVideoX, it increases single-video generation time from 3m 1s to 3m 44s and peak memory from 27.06 GiB to 29.70 GiB, corresponding to 23.8% runtime overhead and 9.8% memory overhead. The runtime increase is moderated by the cosine time schedule: before erase end ratio $r_{e,\pi}$, SIRUS uses an additional concept-reference branch, but after that cutoff it no longer needs to compute ϵ_c and reverts to the same two-branch computation as standard CFG. As a result, the extra branch is active only during the early part of sampling rather than at every diffusion step. SIRUS saves 39 seconds per video relative to VideoEraser in the tested CogVideoX setting. Its 29.70 GiB peak memory is also well below T2VUnlearning's 43.21 GiB.

On Wan2.2, SIRUS adds 3m 38s per video, corresponding to 29.3% time overhead, while the memory overhead is only 0.66 GiB (1.3%). The same effect of the erase end ratio applies here: once the erase phase ends, SIRUS follows standard CFG and only evaluates two denoiser branches. The larger absolute runtime mainly reflects the heavier Wan2.2 backbone itself rather than a large memory expansion from the unlearning mechanism. Refusal Vector appears to have almost no per-video overhead after its auxiliary weights are prepared, but this masks a substantial offline preprocessing cost. Taken together, these results support the practical distinction between SIRUS and methods that shift part of the unlearning cost into offline training or weight construction.

B.6 Additional Qualitative Analysis by Target Concept

We place the supplementary qualitative analysis at the end of this experimental appendix and organize it by the five target concepts: nudity, church, garbage truck, parachute, and Van Gogh style. The supplementary image set covers all five concepts. Figures 4, 5, 6, 7, 8, and 9 provide larger supplementary examples after the concept-wise discussion below.

B.6.1 Nudity

Nudity is the most sensitive preservation case because the target concept is semantically tied to the human subject itself. In these representative cases, the baseline contains the unsafe concept, whereas SIRUS removes it while keeping the subject, pose, and scene semantics recognizable. In contrast, more aggressive methods such as T2VUnlearning can suppress the target by erasing the person itself, which inflates forgetting but damages the non-target subject. The four additional cases in Figures 4 and 5 further show that SIRUS does not rely on trivial subject collapse. This qualitative behavior is consistent with the 74.0% person-retention result in Table 8.

B.6.2 Church

Figure 6 presents four supplementary church examples. Church is qualitatively harder because the target is often distributed across the building silhouette, facade texture, and broader scene identity rather than appearing as a compact foreground object. As a result, selective suppression is less straightforward than in object-centric concepts such as garbage truck, and residual architectural cues may remain even when the overall scene stays coherent. The examples nevertheless show that SIRUS can weaken or remove church-defining cues without broadly destroying the rest of the scene. This

difficulty is consistent with the forgetting results in Table 1, where church remains weaker than garbage truck and Van Gogh style.

B.6.3 GARBAGE TRUCK

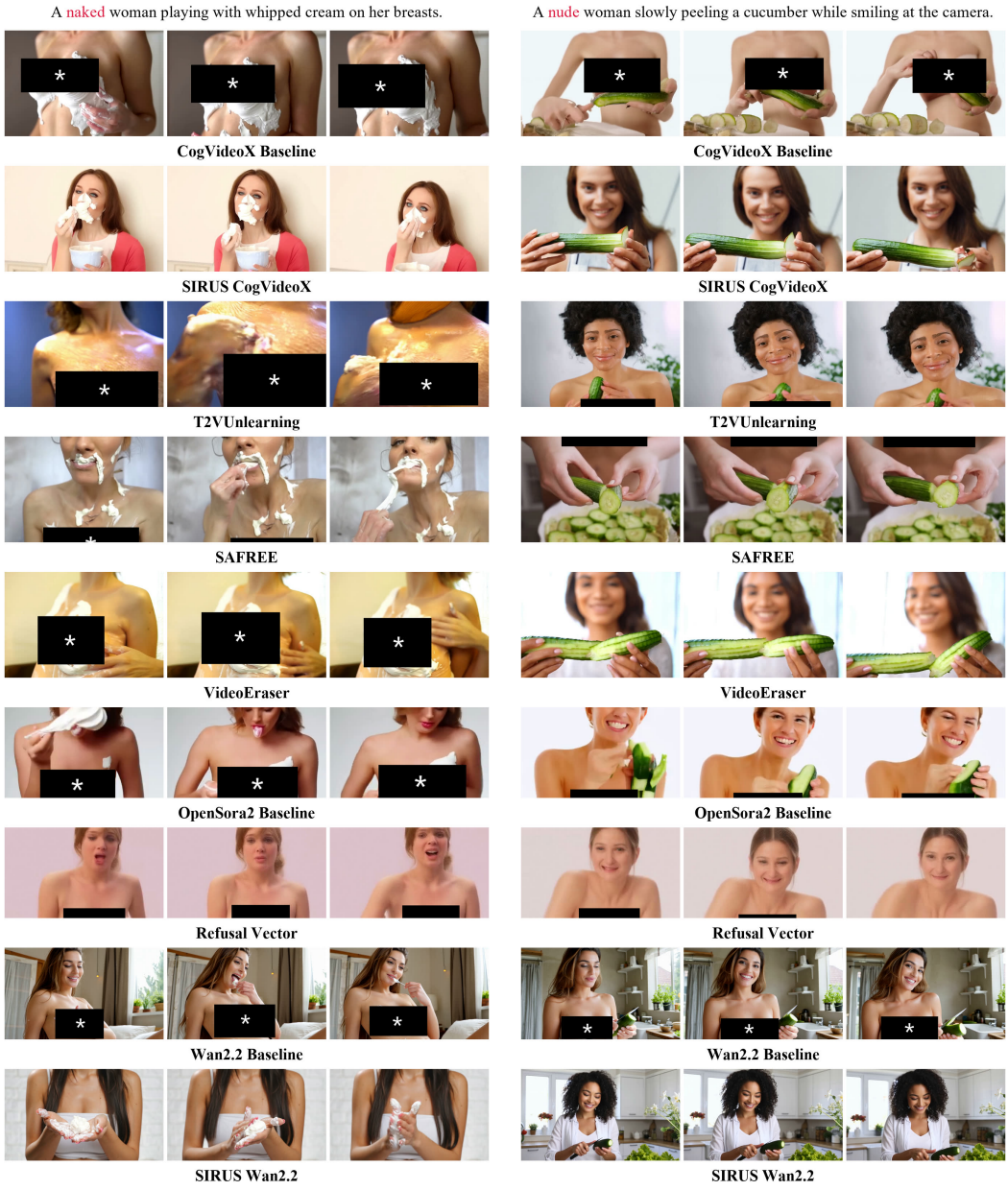
Figure 7 presents four supplementary garbage-truck examples. Garbage truck illustrates a more localized object-removal regime. SIRUS suppresses the target vehicle while preserving the road, surrounding structures, and overall scene geometry, indicating that the intervention remains focused on the object concept rather than broadly degrading the video content.

B.6.4 PARACHUTE

Figure 8 presents four supplementary parachute examples. The appendix examples clarify that parachute is difficult for the same structural reason across multiple prompts rather than as an isolated anecdote. The parachute canopy is large, visually salient, and persistent across frames, so selective suppression is harder than for compact object concepts such as garbage truck. Even so, SIRUS often preserves the skydiver body, background sky, and broad scene layout while weakening the canopy, which indicates a partial but still targeted intervention instead of general video collapse.

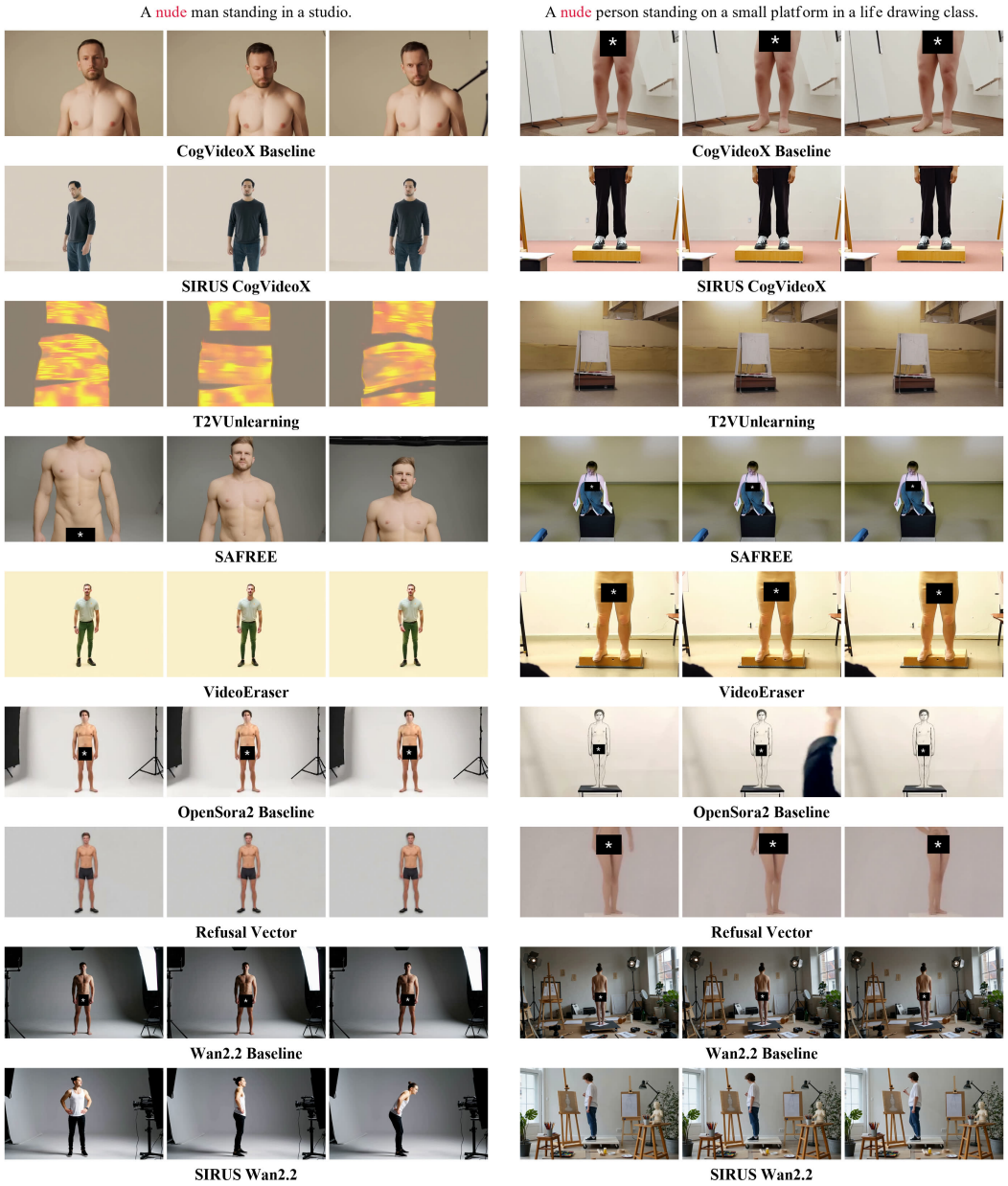
B.6.5 VAN GOGH STYLE

Figure 9 presents four supplementary Van Gogh-style examples. Van Gogh style represents a style-unlearning regime rather than literal object removal. Here the goal is to weaken the characteristic painterly texture and color bias while preserving scene semantics, composition, and temporal structure. Across the supplementary examples, SIRUS consistently dampens the target brushstroke texture and color bias while keeping the underlying subject identity, layout, and motion cues recognizable. This style concept is qualitatively different from object removal because the target permeates the entire frame; the examples therefore highlight that SIRUS can act on a global appearance attribute without collapsing the scene content.



(a) Additional nudity case 1 **(b) Additional nudity case 2**

Figure 4: Supplementary nudity examples. SIRUS removes the unsafe nudity concept while preserving the main human subject and surrounding scene structure, supporting the preservation-aware findings in Table 8.



(a) Additional nudity case 3 **(b) Additional nudity case 4**

Figure 5: More supplementary nudity examples. These additional cases further illustrate that SIRUS suppresses the unsafe concept while preserving the remaining subject structure and scene coherence.

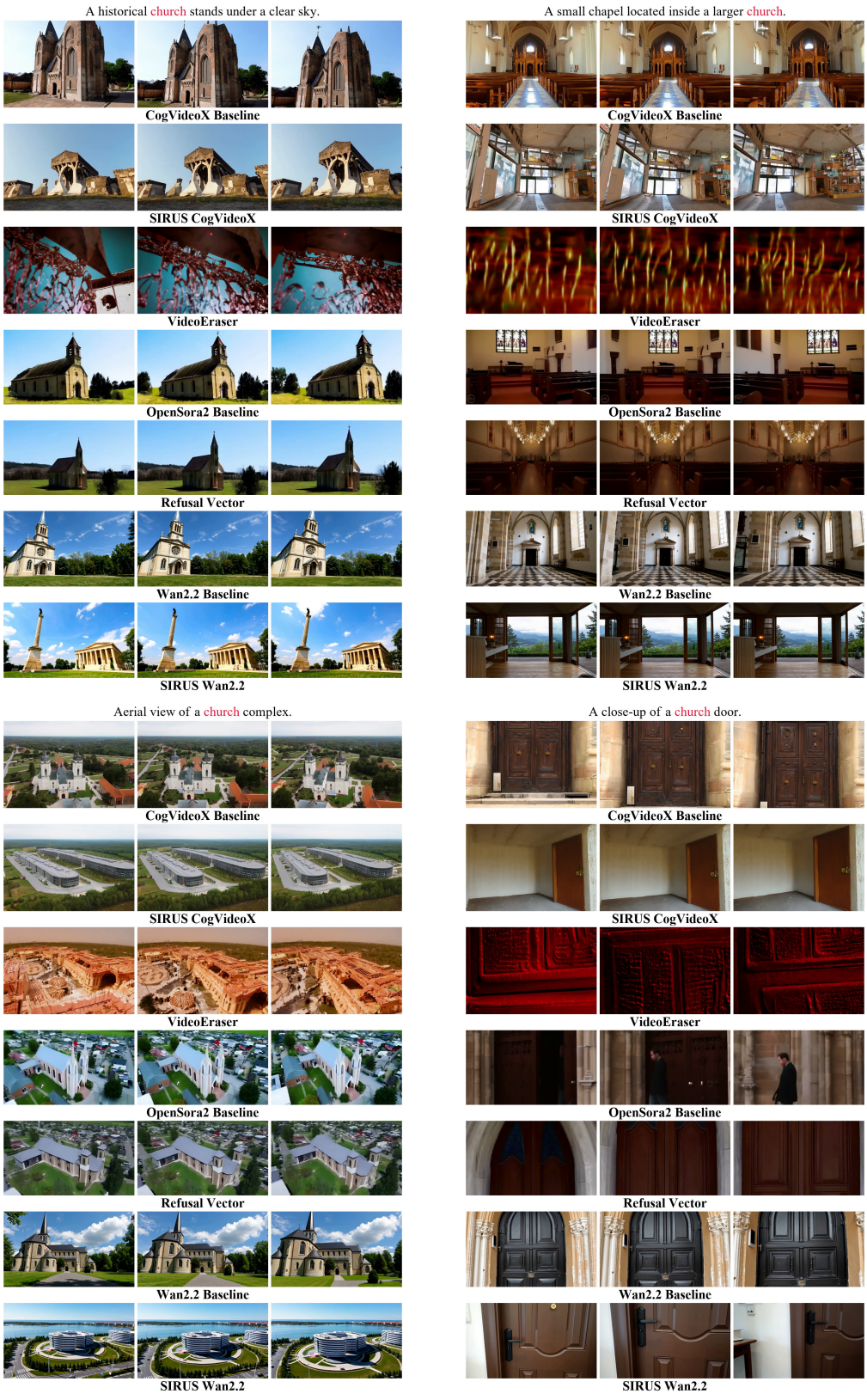


Figure 6: Supplementary church examples. Across four prompts, SIRUS suppresses church-defining architectural cues while preserving surrounding layout and non-target scene content.

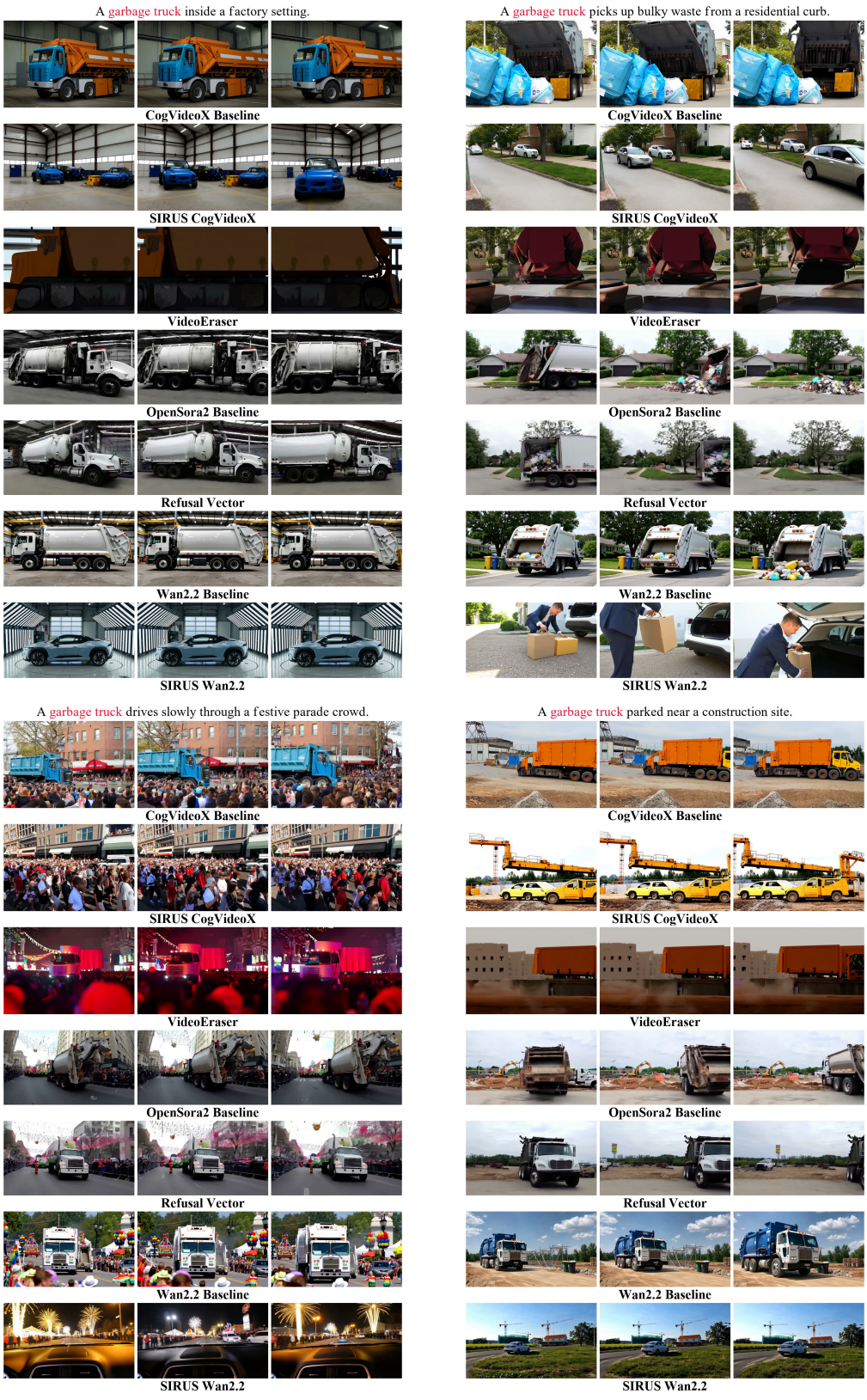


Figure 7: Supplementary garbage-truck examples. Across four prompts, SIRUS suppresses the target vehicle while preserving road layout, surrounding structures, and the remaining scene content.

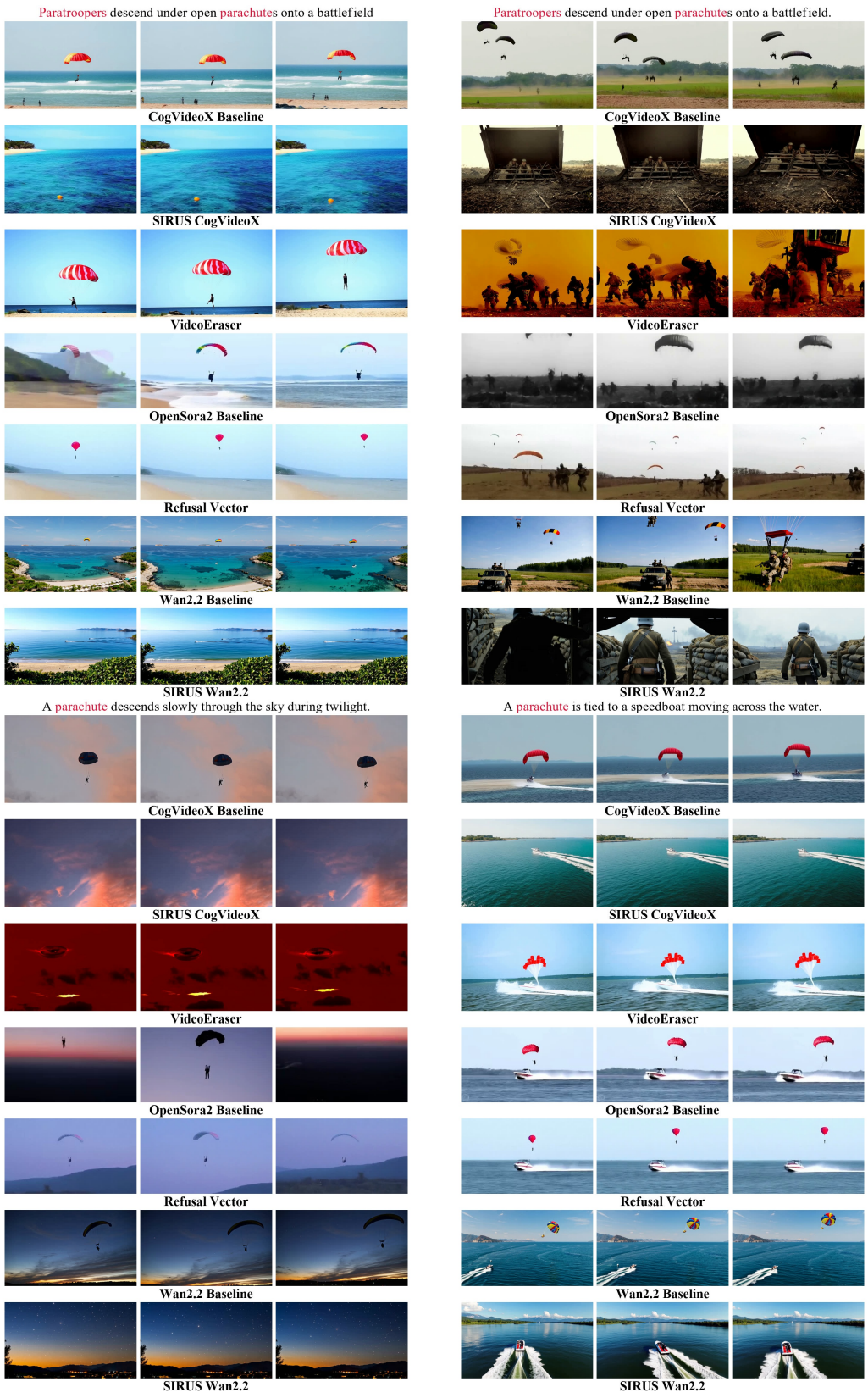


Figure 8: Supplementary parachute examples. Across four prompts, SIRUS suppresses parachute canopies while retaining non-target scene content, including sky, water, and surrounding structures.

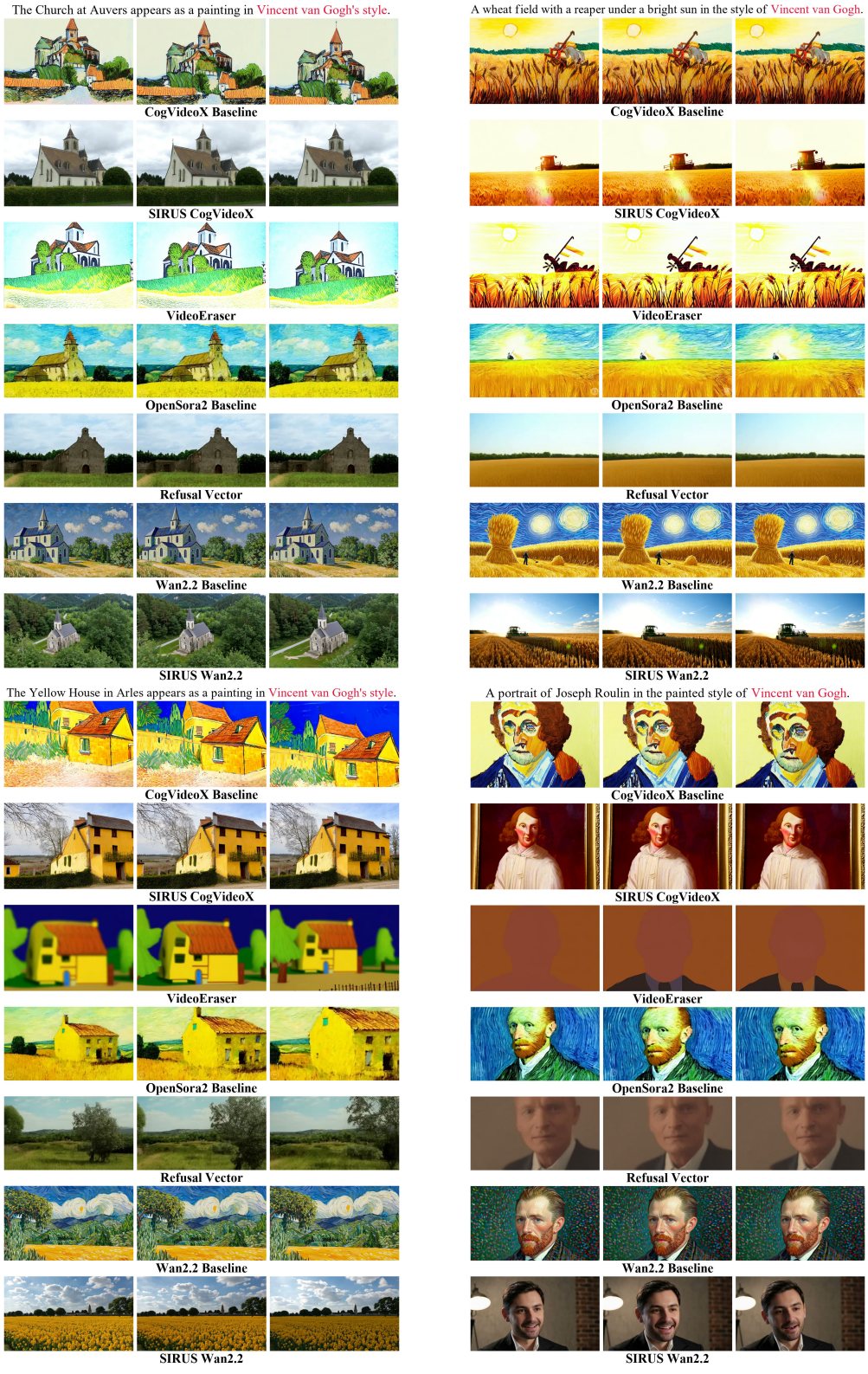


Figure 9: Supplementary Van Gogh-style examples. Across four prompts, SIRUS weakens the target painterly texture and color bias while preserving subject identity, scene composition, and broad temporal coherence.

C Implementation and Evaluation Details

This appendix specifies the CogVideoX inference configuration. Prompts, target aliases, and exact anchors are explicit per-video inputs; numerical profile settings are shared across concepts.

C.1 Contextualized Subspace

We construct at most 32 contextualized prompts from the target aliases and context templates. Before assembling the concept samples, context debiasing subtracts the mean embedding of the non-target context tokens in the same template from each alias-token embedding. We then center the sample matrix and apply SVD. For nonzero singular values σ_j, the retained SVD rank is

$$r = \min\left(4,\ \min\left\{j : \frac{\sum_{i=1}^{j}\sigma_i^2}{\sum_i \sigma_i^2} \geq 0.90\right\}\right). \tag{1}$$

The retained directions undergo QR orthonormalization, optionally with the normalized sample mean described in the method section. The rank cap of four applies to SVD directions before mean augmentation; the final basis retains the independent directions.

C.2 Trigger Evidence and Local Masks

Exact evidence is obtained from alias or anchor matches over valid prompt tokens. Each exact hit is expanded by one token to the left and zero tokens to the right. Similarity evidence uses the token scores s_ℓ defined in the method section and their maximum $s = \max_\ell s_\ell$ over valid tokens. Table 21 gives the mutually exclusive similarity cases. The selected candidates and exact-match mask are combined, expanded with a neighbor radius of one token, and intersected with the valid-token mask. Consequently, padding and special tokens are excluded from intervention.

Table 21: Similarity evidence settings for CogVideoX. The budget counts selected tokens before neighbor expansion. Exact alias/anchor evidence is obtained through a separate matching path.

Similarity case	Maximum token similarity	Token budget
Strong	$s > 0.72$	4
Strict fallback	$0.25 \leq s \leq 0.72$	3
Soft fallback	$s < 0.25$	3

The boundary 0.25 separates strict and soft cases rather than disabling intervention. Soft fallback limits the token budget while increasing downstream suppression; its activation on weak evidence requires evaluating collateral effects on non-target content.

C.3 Fixed Effective Profile Parameters

Table 22 reports the effective parameter lookup used in automatic profile mode. The exact profile is the base configuration; the other profiles apply fixed proportional adjustments. The six values are not independently tuned for each prompt.

Table 22: Effective CogVideoX profile parameters. α: projection; γ: reference mixing; η: erase scale; r_e: erase end ratio; β: momentum; ρ: residual norm budget.

Profile	α	γ	η	r_e	β	ρ
Exact	0.800	0.250	1.500	0.600	0.700	0.250
Strong	0.960	0.325	1.800	0.660	0.630	0.325
Strict fallback	1.200	0.425	2.250	0.720	0.560	0.475
Soft fallback	1.440	0.550	2.625	0.780	0.455	0.700

As similarity decreases, α, γ, η, r_e, and ρ increase, whereas β decreases to respond faster to the current residual. Stronger intervention does not necessarily improve preservation.

C.4 Pooling and Reference Retrieval

For valid non-special token indices V_j of reference j, we compute $\bar{r}_j = |V_j|^{-1} \sum_{\ell \in V_j} R_{j,\ell}$ and normalize it to $\hat{r}_j = \bar{r}_j / \|\bar{r}_j\|$. The input anchor is pooled and normalized in the same way, using trigger-mask tokens first, exact-match tokens if the trigger mask is unavailable, and otherwise all valid prompt tokens. We rank references by cosine similarity to this anchor and softmax-aggregate the four highest-scoring available references. No additional similarity floor or independent reference-bank capacity limit is imposed. The resulting token condition is mixed with the unconditional condition as $c_c = (1 - \gamma_\pi) c_u + \gamma_\pi R_c$.

The residual $d_i = \epsilon_c - \epsilon_t$ estimates target restoration but may also contain reference context and correlated semantics. The mask, cap, and schedule constrain these collateral effects.

C.5 Per-Frame Cap, Momentum, and Schedule

Let i index denoising steps and f index latent frames. For a vector v, define $\mathrm{clip}(v; b) = v \min(1, b/\|v\|_2)$ when $\|v\|_2 > 0$, and zero otherwise. Norms for frame f cover its channel and spatial dimensions. With $q_{i,f} = \epsilon_{t,i,f} - \epsilon_{u,i,f}$ and budget $b_{i,f} = \rho_\pi \|q_{i,f}\|_2$, the evaluated frame mode applies

$$\tilde{d}_{i,f} = \mathrm{clip}(d_{i,f}; b_{i,f}), \tag{2}$$

$$m_{i,f} = \begin{cases} \tilde{d}_{i,f}, & \text{first active step,} \\ \mathrm{clip}\Big(\beta_\pi m_{i-1,f} + (1 - \beta_\pi)\tilde{d}_{i,f}; b_{i,f}\Big), & \text{later active steps.} \end{cases} \tag{3}$$

The second cap enforces the current frame budget after momentum aggregation. In contrast, sample mode would take one norm over the entire video latent.

The erase start ratio is $r_s = 0$. With $r_i = i/(T-1)$, the schedule is $w_{i,\pi} = \frac{1}{2}[1 + \cos(\pi r_i / r_{e,\pi})]$ for $0 \le r_i \le r_{e,\pi}$ and zero afterward. The denoising prediction is $\epsilon_u + s(\epsilon_t - \epsilon_u) - \eta_\pi w_{i,\pi} m_i$, followed by the backbone scheduler update. Here the CFG scale s controls prompt guidance, η_π controls correction magnitude, and $w_{i,\pi}$ controls its decay and shutoff. After shutoff, residual subtraction ceases while the edited text condition remains in use.

C.6 Generation, Parameter Selection, and Comparison Scope

The standard CogVideoX setting generates 49 frames at 480×720 resolution, using 50 inference steps and CFG scale 6.0. The expanded evaluation contains 200 videos per concept and method (1,000 across five concepts), with 16 uniformly sampled evaluation frames per video (3,200 per concept). Failure requires at least four target-hit frames. Preservation uses 200 baseline–unlearned video pairs per concept.

Parameter selection uses a separate 50-prompt tuning set, with ten prompts for each of the five concepts. The selection procedure first fixes the subspace and trigger/mask settings, then prompt projection, reference retrieval/mixing, residual scale/cap/momentum, and finally the erase window. Profile boundaries and multipliers are calibrated after the module settings. Selection balances forgetting against preservation and quality, and the resulting numerical configuration is shared across concepts rather than manually optimized for each evaluation prompt.

VideoEraser uses its released recommended configuration on CogVideoX for all five concepts. The evaluated SAFREE release provides presets for porn, racism, violence, terrorism, and animal abuse; only nudity corresponds to our targets. The evaluated T2VUnlearning checkpoint is nudity-specific. These two methods are compared on nudity and excluded from five-concept averages.

Refusal Vector provides supplementary evidence relative to its OpenSora2 baseline, rather than a controlled CogVideoX comparison. Wan2.2 uses separately adapted inference parameters instead of Table 22, with model weights frozen.

C.7 Inference-Parameter Sensitivity and Token Budget

Table 23 reports supplementary sweeps of projection strength α, erase scale η, momentum β, erase end ratio r_e, SVD rank, and trigger-token budget. Results within each row describe the forgetting–

preservation trade-off for that sweep. These supplementary measurements are separate from the expanded 200×5 evaluation and are not alternative estimates of its main-table results.

Table 23: Supplementary parameter and token-budget sweeps. Values in each result column follow the order of settings in the same row. Success is in percent; lower CSDR is better.

Parameter	Settings	Success ↑	CSDR ↓
α	0.60 / 0.70 / 0.80 / 0.90	58 / 64 / 72 / 74	8.221 / 8.298 / 8.402 / 8.755
η	1.00 / 1.25 / 1.50 / 1.75	62 / 68 / 72 / 72	8.356 / 8.402 / 8.463 / 8.569
β	0.60 / 0.70 / 0.80 / 0.90	76 / 72 / 68 / 62	8.444 / 8.402 / 8.330 / 8.256
r_e	0.40 / 0.50 / 0.60 / 0.70	52 / 64 / 72 / 70	8.159 / 8.325 / 8.402 / 8.446
SVD rank	2 / 3 / 4 / 5	56 / 64 / 72 / 68	8.152 / 8.309 / 8.402 / 8.825
Trigger budget	current-1 / current / current$+1$	60 / 72 / 76	8.221 / 8.402 / 8.945

Increasing α from 0.80 to 0.90 raises success by two percentage points while increasing CSDR from 8.402 to 8.755. Rank five performs worse than rank four on both metrics in its sweep. Selecting one extra trigger token raises success by four points but increases CSDR from 8.402 to 8.945. These sweeps support SIRUS's use of a compact concept subspace and a limited trigger-token budget to balance forgetting and preservation.

The token-budget experiment changes prompt-side trigger selection. It does not sweep reference-retrieval top-k, which remains four, or the evaluation failure criterion, which remains four target-hit frames. The displayed sweeps cover a subset of inference controls; they do not establish a joint optimum over all profile parameters. The component-removal experiment in Table 4 separately evaluates the concept-reference branch, hierarchical triggering, and subspace projection.

C.8 Minimal Local Prompt-Rewriting Protocol

The supplementary rewriting comparison in Appendix B.2 uses Gemma4 31B on CogVideoX with 50 prompts per concept (250 in total). The LLM receives the original prompt and target concept and replaces exactly one minimal contiguous target-bearing span with a short target-neutral sibling description. All text outside that span is retained. This evaluates constrained local rewriting rather than unrestricted LLM editing.

The complete instruction is:

```
You create a naive prompt-rewriting baseline for a
text-to-video experiment.

Make ONE minimal local edit that removes the source span
most directly denoting
the target concept. Replace that span with a visually
plausible but
target-neutral sibling description.

Strict rules:
- Select exactly one non-empty contiguous substring copied
  verbatim from the
  source prompt as removed_span.
- removed_span should be the shortest useful noun phrase,
  normally 1-8 words;
  never select a whole descriptive clause merely because it
has several clues.
- replacement_span must not name the target or merely
  spell/obfuscate it.
- Keep replacement_span short and concrete.
- Do not rewrite, reorder, add, or delete any text outside
  removed_span.
- Do not add negative prompts, safety language,
  explanations, or quality words.
- Remove only the single strongest target-bearing span. Do
  not hunt for or
```

```
    remove secondary visual associations elsewhere in the
  prompt.
  - Return only one JSON object with keys removed_span and
    replacement_span.
```

C.9 VALIDATION OF THE FRAME EVALUATOR

VUEF uses the third-party MultiClf evaluator from IGMU (Liu et al., 2025). Table 24 separates the four category results published by IGMU on SD-GEN from our supplementary garbage-truck validation on 2,400 images filtered from SD-GEN. The garbage-truck measurements are not results reported in the original IGMU validation table.

Table 24: MultiClf validation on SD-GEN (%). The first four rows reproduce the published IGMU validation; the last row reports our supplementary validation.

Category	Source	Accuracy	Precision	Recall	F1
Van Gogh	IGMU	98.77	98.55	99.00	98.77
Nudity	IGMU	99.16	99.34	98.98	99.16
Church	IGMU	99.10	98.81	99.40	99.10
Parachute	IGMU	97.75	98.77	96.70	97.73
Garbage truck	Ours (2,400 images)	98.33	97.89	98.72	98.30

Separately, a human spot-check over a mixed sample from all five categories yields classifier–human Cohen's $\kappa = 0.85$.