CAUSAL-ROLE ERASURE FOR TEXT-TO-VIDEO DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video object removal is an important capability with broad applications in controllable generation and content editing. However, erasing an object does not always erase its physical causal effects: a droplet may vanish while its splash remains. Existing text-to-video concept-erasure methods target named concepts rather than their causal roles, leaving this conditional removal without an explicit formulation or training strategy. We formulate *causal-role erasure* as suppressing a source and its effects when it causes an event while preserving the same object's noncausal uses, and introduce Source-slot Randomized Counterfactual Distillation (**SRCD**) to address it. Specifically, we construct independently generated and screened source-free target videos that retain the receiver and scene without the source or its effects, then pair each target with event prompts that vary only the source identity. Across seven physical mechanisms, SRCD obtains a higher average graded erasure score than an otherwise matched control, but the corrected difference is not statistically significant; paired-role tests show no clear additional benefit, and human review does not confirm the automated comparison.

1 INTRODUCTION

Removing unwanted objects is a common requirement in controllable video generation and editing. Yet suppressing an object alone can leave behind physical changes it caused (Figure 1). Depending on the interaction, these can appear as splashes, fractures, deformations, or surface traces. We call these dependent changes the object's *causal footprint*. Recent effect-aware video-removal methods address residual effects in an observed clip (Motamed et al., 2026; Ekin et al., 2026). We instead edit a text-to-video (T2V) generator: given a prompt describing a targeted physical interaction, it should suppress both the event source and its footprint. This setting requires the model to learn the intervention across events, since it has no observed clip from which to locate the source or its effects at inference.

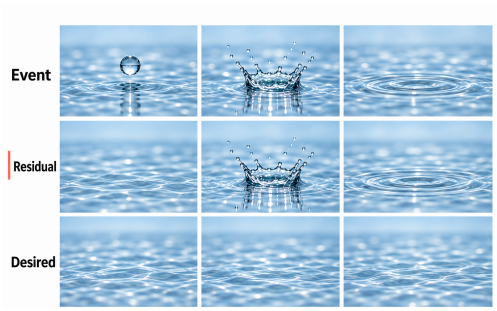

Figure 1: **A source can vanish while its footprint remains.** Conceptual illustration.

Crucially, the desired intervention depends on the object's role in the described event. An object may cause a physical response in one prompt but appear as a bystander in another; suppressing its name everywhere would remove valid content. Existing concept-erasure methods typically target a named object or semantic category (Gandikota et al., 2023; Xu et al., 2025), without specifying this role distinction or the source's dependent footprint. We therefore formulate *causal-role erasure*: suppress the source and its footprint when it causes the targeted event, while preserving noncausal uses of the same object. This formulation requires separate evaluation of source removal, footprint removal, and role-specific preservation.

Training for causal-role erasure requires target videos that show what should remain after the source and its effects are removed. A mask around the source alone cannot specify the desired state of the affected scene, and a factual event video does not contain its own source-free outcome. We therefore

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.

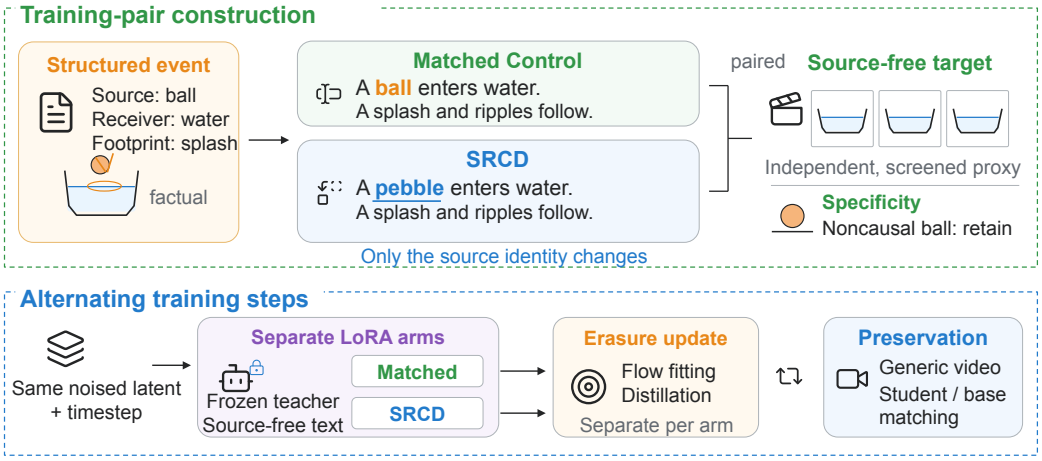

Figure 2: **Training pairs and updates.** Top: SRCD replaces the source object in the prompt while keeping the source-free target video fixed. Matched uses the original source identity. The target is generated independently, not edited from the original video. Bottom: both methods alternate erasure updates with teacher matching on generic preservation videos. The noncausal ball illustrates a preservation requirement tested during evaluation, not a training example.

generate source-free targets independently from prompts that retain the affected scene element (the receiver) and scene context while omitting the source and its footprint. Before training, we screen candidate videos for source and footprint absence, receiver preservation, and quality. These targets specify desired scene states rather than frame-aligned edits of observed videos.

For a fixed interaction, the desired source-free receiver state can be shared across different source identities. Source-slot Randomized Counterfactual Distillation (SRCD) uses this property by replacing only the source identity in each event prompt while keeping the interaction, receiver, expected footprint, scene description, and target video fixed (Figure 2). A student conditioned on the event prompt is trained on the target video and matched to a frozen teacher conditioned on the source-free prompt. This design aims to reduce dependence on the original source nouns, while generic preservation updates limit changes on unrelated scenes. To isolate the effect of source replacement, we train separate adapters for each mechanism and compare SRCD with Matched, which retains the original source identities under the same targets, objectives, and training settings.

Across seven physical mechanisms, our evaluation separates disappearance of the source, removal of its footprint, and preservation of noncausal uses. SRCD obtains a higher average graded erasure score than Matched under fixed targets and training settings, with gains in several mechanisms. The corrected overall comparison is not statistically significant; role-paired tests show no clear additional benefit, and human review does not confirm the automated comparison.

Our contributions are summarized as follows:

1. We formulate causal-role erasure for T2V models, distinguishing source-and-footprint suppression in causal prompts from preservation under noncausal uses of the same object.
2. We introduce Source-slot Randomized Counterfactual Distillation, combining independently generated and screened source-free targets with source-slot variation under a fixed desired outcome.
3. We provide a seven-mechanism evaluation that measures source and footprint removal, role preservation, and the agreement between automated and human judgments.

2 Related Work

Concept erasure in diffusion models. Concept erasure suppresses selected semantic content while preserving untargeted capabilities. Image methods use negative-guidance fine-tuning or closed-form

parameter edits (Gandikota et al., 2023; 2024). T2V approaches filter prompt embeddings and latent features, steer denoising noise, train adapters against negatively guided velocity targets, learn preferences over concept-free trajectories, or localize sparse features (Yoon et al., 2025; Xu et al., 2025; Ye et al., 2025; Wang et al., 2026a;b). Their target remains a specified semantic concept or category, evaluated through target presence, generation quality, and broad preservation. Causal-role erasure instead registers a source, its dependent footprint, and specificity: the same noun must remain valid in registered noncausal prompts.

Video object and effect removal. Video inpainting reconstructs masked regions across frames (Zhou et al., 2023). Effect-aware removal includes associated shadows, reflections, and appearance changes (Miao et al., 2025; Kushwaha et al., 2026); EffectErase adds deformation, and VOID extends removal to downstream interactions (Fu et al., 2026; Motamed et al., 2026). BeyondMasks separately evaluates object disappearance and residual after-effects against paired clean references (Ekin et al., 2026). These are direct precedents for separating object disappearance from residual effects, but the removal methods operate on an observed video using masks, reference pairs, or affected-region localization. We instead edit T2V generators: at inference, each mechanism-specific edited generator uses only the rendered prompt—no input video, object/effect mask, user-specified trajectory, or reference frame—and is evaluated on registered source-role and noncausal uses of the same noun.

Counterfactual supervision and event roles. Related systems construct alternative outcomes through paired present/absent supervision or assumed causal structure. EffectErase learns from paired videos, while VOID constructs paired counterfactual removals (Fu et al., 2026; Motamed et al., 2026). Causally Steered Diffusion optimizes video-editing prompts under a supplied graph (Spyrou et al., 2025). Separately, event-centric causal-chain prompting structures physical generation as temporally ordered events (Wang et al., 2026c). Video semantic-role labeling represents events through verbs, role-bearing entities, and event relations, but predicts structure rather than editing a generator (Sadhu et al., 2021). Our source-free target is independently generated rather than a pixel-aligned individual counterfactual, and event roles are supplied rather than discovered. We test role-conditioned behavior, not an internal causal representation.

3 Causal-Role Erasure

We modify a text-to-video generator to suppress an object's causal role and its effects while preserving its noncausal uses. At inference, the model receives a text prompt, not a video to edit.

3.1 Events and the Intervention Unit

A droplet falling into a bowl of water illustrates the event roles. The droplet is the *source*, the water is the *receiver*, and entry into the water is their *interaction*. The resulting splash and ripples are the *footprint*; unrelated motion or pre-existing ripples do not count.

For prompt construction, $e_i = (m_i, s_i, a_i, r_i, f_i, h_i, c_i)$ denotes the i-th structured event description. Here m_i is the mechanism (water impact); s_i is the source identity (droplet); a_i is the interaction (entry into water); r_i is the receiver (water in the bowl); and f_i is the expected footprint (splash and ripples). The field h_i specifies the desired source-free receiver state (here, undisturbed water), and c_i gives the lighting, viewpoint, and background. We specify these fields when constructing the data, so we can vary s_i while keeping the others fixed (§4.2). The tuple is used to construct prompts, not passed as a separate inference input.

3.2 Desired Outputs

For a prompt describing a droplet entering water, the modified generator should suppress the droplet and its splash or ripples. The water, bowl, and background should remain. In a noncausal prompt, a droplet suspended without touching the water should stay visible. If that prompt also specifies a mechanical paddle stirring the water, the paddle-driven motion should remain as well.

Several videos can meet these requirements. For a causal prompt, h_i describes the desired receiver state, not a video to copy frame by frame. Unrelated motion, such as shifting reflections, may remain.

3.3 Measuring Erasure and Preservation

Let u denote an evaluation case and k a method using backbone b. We rate each whole video for source visibility S, footprint visibility F, receiver preservation R, and quality Q. Single ratings use 0, 1, or 2. For visibility, these mean absent, partial or ambiguous, and clear; higher R, Q indicate better preservation. The formulas use ratings averaged per video and field, before applying gates or computing scores.

For causal cases, the binary gate $E_b(u)$ is 1 when the unmodified model's output (Original) is judged to exhibit the event. The usability gate $U_k(u)$ is 1 when $R_k(u), Q_k(u) \geq 1$. Each gate is 0 otherwise. The Causal Erasure Score is

$$\mathrm{CES}_k(u) = E_b(u)U_k(u)\frac{(2 - S_k(u)) + (2 - F_k(u))}{4}. \tag{1}$$

CES ranges from 0 to 1. If both gates pass, a video with no visible source but a clear footprint earns 0.5. Cases failing either gate remain in the average with score zero. CES measures graded removal, not a success rate; method comparisons use paired score differences.

On noncausal cases, $P_k(u)$ rates the protected object's visibility and $N_k(u)$ whether it stays noncausal. Both use higher-is-better 0–2 ratings. Specificity utility is $\mathrm{SU}_k(u) = U_k^{\mathrm{sp}}(u)[P_k(u) + N_k(u)]/4$. Here $U_k^{\mathrm{sp}}(u)$ applies the same usability rule to noncausal videos; no Original-event gate is required. Receiver preservation and quality are reported separately. Full rubrics and thresholds are in Appendix D.

4 Source-slot Randomized Counterfactual Distillation

SRCD changes the source identity in each training prompt while keeping its source-free target fixed (Figure 2). We train a separate LoRA adapter for each mechanism.

4.1 Constructing Source-Free Training Targets

We pair a prompt describing the original event with a target video intended to omit its source and footprint. For water impact, the prompt describes a droplet entering water and producing a splash. The target should show the water without the droplet, splash, or impact-induced ripples, while allowing unrelated motion such as shifting reflections.

To generate the target, we write a source-free prompt p_i^{cf} that keeps the receiver and scene description but requests the receiver state h_i. A frozen base model generates a video x_i^{cf} from this prompt. We screen candidates for source and footprint absence, receiver preservation, and video quality before training. These videos are generated independently, not by editing a factual video; the shared scene description does not imply frame-by-frame alignment. Generation settings and screening criteria are in Appendix B.

Our control, Matched, uses the original source identities with otherwise matched training conditions. It shares SRCD's selected targets. We construct the factual prompt as $p_i^{\mathrm{f}} = T_{m_i}^{\mathrm{f}}(s_i, a_i, r_i, f_i, c_i; w_i)$, where $T_{m_i}^{\mathrm{f}}$ is the prompt template for mechanism m_i and w_i selects its wording style. Matched uses this prompt unchanged.

4.2 Randomizing the Causal-Source Slot

With fixed source identities, an adapter may learn to suppress the named objects without extending that behavior to other sources of the same event. We therefore vary the source identity in the training prompt while keeping its source-free target fixed.

The interaction, receiver, expected footprint, and scene description remain unchanged. The target depicts the receiver without the source or its effects, so it can supervise the altered prompt as well. For water impact, changing the source leaves the desired undisturbed water state unchanged.

We replace source s_i with $q_i \neq s_i$ and rebuild the prompt:

$$p_i^{\mathrm{aug}} = T_{m_i}^{\mathrm{f}}(q_i, a_i, r_i, f_i, c_i; w_i). \tag{2}$$

Replacements are assigned before training from a 64-identity bank, with balanced usage; they are not redrawn at each update.

Matched trains on $(p_i^{\mathrm{f}}, x_i^{\mathrm{cf}})$, whereas SRCD trains on $(p_i^{\mathrm{aug}}, x_i^{\mathrm{cf}})$. They share targets, preservation data, and training settings (Appendix C). This comparison tests whether source replacement improves erasure beyond training on the original source names.

4.3 Counterfactual Distillation and Preservation

The student learns a source-free output while receiving a prompt that still describes the causal event. The frozen base model, with LoRA disabled, serves as the teacher and receives the source-free prompt instead. Both process the same noised target video at the same timestep.

We supervise the student with both the target video's flow-matching objective and the teacher's prediction. Let z_i^{cf} be the target video latent, ϵ_i Gaussian noise, and σ_i a uniformly sampled noise level. Their shared input is $z_{i,\sigma} = (1-\sigma_i)z_i^{\mathrm{cf}} + \sigma_i\epsilon_i$. Write $v_\phi^i(\rho)$ and $v_\theta^i(p_i^{\mathrm{cf}})$ for the student and teacher flow predictions. The student prompt is $\rho = p_i^{\mathrm{aug}}$ for SRCD and $\rho = p_i^{\mathrm{f}}$ for Matched. The erasure loss is

$$\mathcal{L}_{\mathrm{erase}}^i = \mathrm{MSE}(v_\phi^i(\rho), \epsilon_i - z_i^{\mathrm{cf}}) + \lambda_T \,\mathrm{MSE}(v_\phi^i(\rho), \mathrm{sg}[v_\theta^i(p_i^{\mathrm{cf}})]). \tag{3}$$

The first term fits the flow target $\epsilon_i - z_i^{\mathrm{cf}}$ obtained from the target video and sampled noise. The second matches the teacher's source-free prediction despite the student's event prompt. Here λ_T weights teacher matching and sg stops gradients through the teacher.

To limit changes on other scenes, we interleave preservation updates using generic non-target videos. For preservation sample j, both receive the same prompt p_j^{pr}, noised video latent, and timestep. We match their predictions with weight λ_P:

$$\mathcal{L}_{\mathrm{preserve}}^j = \lambda_P \,\mathrm{MSE}(v_\phi^j(p_j^{\mathrm{pr}}), \mathrm{sg}[v_\theta^j(p_j^{\mathrm{pr}})]). \tag{4}$$

These examples are separate from the noncausal-role evaluation cases. Erasure and preservation updates strictly alternate; the two losses are not summed on one minibatch. Only LoRA parameters are optimized. Matched and SRCD share both objectives and the update schedule. Weights, update counts, and other training settings are in Appendix C.2. At inference, the generator uses the mechanism-specific LoRA with the event prompt; neither the teacher nor a target video is needed.

5 Experiments

5.1 Experimental Setup

Cases and models. Each of seven equally weighted mechanisms contributes 24 causal and 18 specificity cases, for 294 semantic cases. The causal set varies source identity, receiver, wording, and whether the footprint is named explicitly. Specificity cases test noncausal use, role changes, and acceptable alternative causes. Wan 2.1 T2V 1.3B supplies Original, Matched, and SRCD. The separate CogVideoX-2B block supplies Original, Negative Prompt, VideoEraser, T2VUnlearning-adapted, and SAFREE. All eight streams use matched prompts and seeds within compatible settings. The main study contains 2,352 videos; Generic Paraphrase and Bystander Token on Water and Fracture add 96. Full model and baseline settings are in Appendix C.

Evaluation. Every judgment covers all 49 frames. Two independently ordered passes of the same VLM judge provide the primary scores. We average the two passes field by field before computing eligibility, CES, and specificity utility. Two human raters independently score 224 videos to assess evaluator reliability on a stratified subset; these labels do not replace the full automatic evaluation. CES and specificity utility are defined in §3.3; receiver preservation, quality, and usability are reported separately. Cases are averaged within mechanisms and the seven means receive equal weight. Paired, mechanism-stratified bootstrap intervals use 10,000 replicates; headline tests use Holm correction and preservation uses fixed non-inferiority margins. The controlled comparison is SRCD versus Matched. External Original-normalized contrasts additionally require shared Original capability. The complete protocol is in Appendix D.

Table 1: CES by mechanism with equal-weight macro averages. Bold marks the highest CES in each column within a backbone; cross-backbone rows are descriptive.

Method	Water	Rigid	Brittle	Powder	Elastic	Release	Trace	Macro
Wan 2.1 T2V 1.3B								
Original	0.0312	0.0365	0.0104	0.0573	0.0521	0.0260	0.0781	0.0417
Matched	0.2135	0.1562	0.3229	0.2708	0.1094	**0.1979**	**0.3646**	0.2336
SRCD	**0.3281**	**0.1719**	**0.3333**	**0.3229**	**0.2031**	0.1875	0.3021	**0.2641**
CogVideoX-2B								
Original	0.0208	0.0156	0.0104	0.0521	0.0521	0.0156	0.0417	0.0298
Negative Prompt	0.0417	**0.0521**	0.0208	0.0938	0.0729	0.0521	0.0677	0.0573
VideoEraser	0.0260	0.0365	**0.0625**	0.0938	0.1250	0.0885	0.0573	0.0699
T2VUnlearning-adapted	0.0208	0.0365	0.0156	**0.1667**	0.0833	0.0781	0.0573	0.0655
SAFREE	**0.0885**	0.0312	0.0260	0.1302	**0.1979**	**0.1510**	**0.0990**	**0.1034**

Table 2: SRCD minus lexical controls on Water and Fracture. Bold marks the larger point estimate (higher favors SRCD). Full identification requires the specificity CI above -0.10.

Control	ΔCES	95% CI	ΔImplicit	ΔSU	95% CI
Generic Paraphrase	+0.0990	[0.0104, 0.1927]	+0.1562	-0.0469	[-0.1354, 0.0052]
Bystander Token	**+0.1562**	[0.0573, 0.2708]	**+0.1771**	**-0.0052**	[-0.0156, 0.0000]

5.2 Main Results

Effect of source-identity variation. With source-free targets and training settings held fixed, the SRCD–Matched comparison tests the added value of varying source identities (Table 1). SRCD has a higher macro CES, but the Holm-adjusted test does not establish overall superiority ($p = 0.1300$). We next examine how this average difference is distributed across mechanisms.

Where the gain occurs. Water, Elastic, and Powder have the largest mean CES gains (Figure 4). For Powder, the mean footprint contribution is unchanged: its gain comes entirely from the source term. In Elastic, the footprint term accounts for most of the mean gain. Material Release has a higher footprint contribution, yet the larger decrease in its source contribution lowers its total CES. Component estimates and intervals are in Appendix E.4.

Base capability and external comparisons. SRCD and Matched share the Wan backbone; the external baselines use CogVideoX. Cross-block score differences mix editing effects with base-model differences. Each Original is an unedited reference for checking whether the base model can generate the target event, not an erasure method. Only 55 of 168 causal cases are eligible under both Originals, with too few shared cases in most mechanisms for the specified normalized macro comparison. The external rows in Table 1 are therefore descriptive results on CogVideoX. Eligibility counts and normalization requirements are in Appendix E.3.

5.3 Source Replacement and Lexical Controls

We test two alternative explanations for SRCD's gains on Water Impact and Brittle Fracture: changed wording and exposure to more object names. Generic Paraphrase rephrases the prompt without changing its event source. Bystander Token uses SRCD's replacement names at the same frequencies, but adds them as motionless background objects instead of replacing the original source. All three arms share source-free targets, preservation data, and training settings. We require a causal CES gain, a positive CES difference on implicit-footprint prompts, and specificity non-inferiority (Table 2).

Averaged over the two mechanisms, SRCD has higher causal CES and a positive implicit-footprint CES difference against Bystander Token, while meeting the specificity non-inferiority criterion. Adding the same names to the background does not reproduce SRCD's erasure scores. This comparison also changes whether the original source is retained, so it does not isolate semantic role placement alone. Against Generic Paraphrase, the erasure differences are again positive, but the

Table 3: SRCD minus Matched on role-related subsets. Intervals are unadjusted supporting intervals; the subsets overlap on 56 cases.

Subset	Cases	ΔCES	95% CI
Implicit footprint	84	+0.0402	[0.0060, 0.0789]
Held-out source	112	+0.0458	[0.0112, 0.0815]

Table 4: Preservation contrasts for SRCD minus Matched. Non-inferiority requires the interval lower bound to exceed the prespecified margin.

Metric	Difference	95% CI	Margin
Specificity utility	-0.0040	[-0.0377, 0.0268]	-0.1000
Causal receiver	-0.0015	[-0.0238, 0.0208]	-0.1000
Causal quality	+0.0000	[-0.0208, 0.0208]	-0.1000
Causal usable	+0.0000	[0.0000, 0.0000]	-0.0500

specificity interval includes both no loss and losses beyond the allowed margin, leaving preservation within tolerance unresolved.

5.4 Generalization beyond Training Names and Explicit Effects

We compare SRCD with Matched within two subsets of the causal evaluation. Implicit-footprint prompts retain the source interaction but omit explicit effect descriptions. For example, a prompt can describe an object entering water without mentioning a splash or ripples. Held-out-source prompts use identities excluded from both the original training-source inventory and the augmentation bank. This split refers to erasure training, not the base model's pretraining. Both tests cover the same seven event mechanisms.

SRCD has higher mean CES than Matched under both conditions (Table 3). The observed gain thus extends to test objects outside the erasure-training source inventory and prompts without explicit effect words. As in the main results, Water, Elastic, and Powder have the largest mean gains, while Material Release has negative differences in both comparisons. Since CES combines source and footprint scores, the implicit-prompt result concerns overall erasure rather than footprint removal alone. The subsets share cases and use unadjusted intervals, so they provide overlapping evidence, not independent replications.

5.5 Erasure Behavior and Preservation

Table 5: Additive source and footprint contributions to CES over 168 causal cases. Intervals are exploratory, unadjusted paired 95% intervals for SRCD minus Matched.

Contribution	Matched	SRCD	Difference	95% CI
Source	0.0737	0.0878	+0.0141	[-0.0037, 0.0327]
Footprint	0.1600	0.1763	+0.0164	[-0.0037, 0.0372]
Total CES	0.2336	0.2641	+0.0305	[0.0000, 0.0625]

What the CES gain measures. Table 5 shows positive mean differences in both source and footprint contributions, although both intervals include zero. The two components need not improve in the same videos. Usability decisions agree across arms on eligible cases (Appendix E.4). Figure 4 shows increased source absence, unchanged footprint absence, and slightly lower strict success. CES credits partial reduction, whereas absence endpoints require zero visibility. Single-field absence rates include all causal outputs without Original gating. Strict success requires an eligible Original, both source and footprint absent, and top receiver and quality scores.

Preservation and receiver failures. All four macro contrasts in Table 4 meet their non-inferiority margins. They measure the added preservation cost of source replacement relative to Matched, since

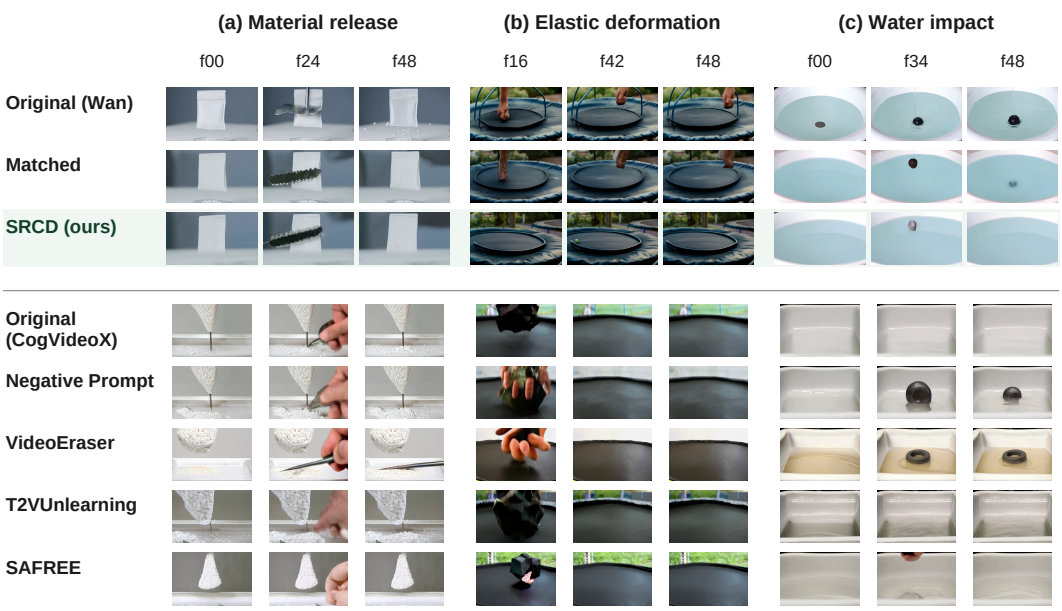

Figure 3: **Selected actual-output comparisons.** Each panel uses matching frame indices across methods. The tool in (a), small object at frame 42 in (b), and transient source at frame 34 in (c) remain visible. Examples are outcome-aware selections; cross-backbone contrasts are descriptive.

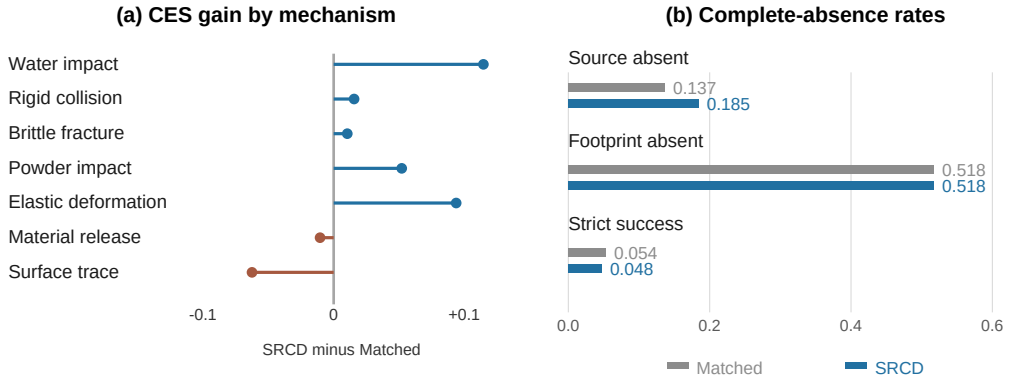

Figure 4: **Mechanism and endpoint decomposition.** Mechanism-level CES differences and complete-absence endpoints. Panel (a) shows point estimates; component intervals are reported in Appendix E.4. Binary absence and strict success do not measure the same outcome as graded CES.

both arms already undergo erasure training. The averages conceal a receiver-preservation failure in Brittle Fracture. In the existing VLM scores, two of its 18 noncausal SRCD outputs receive zero receiver preservation despite high visual-quality scores, giving them zero SU. The other 16 outputs have full SU, but the two failures remain in the reported average. This SU loss directly reflects failed receiver preservation; it does not by itself show that the protected object was incorrectly deleted. Full preservation results are in Appendix E.1.

Same-object role pairs. The 42 semantic role pairs match object identity, receiver, and wording style across causal and noncausal prompts. Both arms have high noncausal SU. SRCD has slightly lower mean causal CES and slightly higher SU than Matched, with both difference intervals including zero (Appendix E.5). Retention is high, but the paired comparison does not establish an additional role-selective erasure benefit.

Visual comparison. The Wan comparisons in Figure 3 show several forms of partial change. In Material Release (a), released beads are less conspicuous under SRCD, but the tool remains and foreground blur obscures the collection region. In Elastic Deformation (b), the source is less conspicuous under SRCD than in Original and Matched, with a small object visible at frame 42. In Water Impact (c), Matched retains a source after entry; SRCD shows a brief source around frame 34 and calmer water later. Receiver framing and geometry also vary across outputs. The displayed frames retain transient residues found during full-sequence inspection; they do not imply continuous source absence. Further selected examples appear in Appendix H.

5.6 Human Evaluation and Judge Reliability

Two computer science PhD students independently scored the same 224 formal videos, formed by crossing 28 semantic cases with the eight evaluated streams. Method labels and VLM scores were hidden. Human raters viewed the original-resolution videos with replay; the VLM viewed low-resolution panels covering all 49 frames. We retained the independent human labels without adjudication. This comparison includes differences in input presentation as well as judgment.

Table 6: Exact agreement (%) on 224 videos. H–H compares the two humans; H–V averages all four human–VLM-pass comparisons. Fields have 112 videos each, except receiver and quality (224).

Field	H–H	H–V
Source visibility	61.6	67.2
Footprint visibility	44.6	63.6
Protected-object visibility	69.6	75.9
Noncausal role	88.4	87.7
Receiver preservation	86.2	85.9
Video quality	91.5	93.1
All fields	77.5	81.6

Agreement is higher for receiver preservation and video quality than for source and footprint visibility (Table 6). Receiver and quality labels account for half the scored items and are concentrated at the top of the scale. Pooling these fields with source and footprint visibility masks weaker consistency in erasure judgments. The humans also disagree substantially on footprints, so the issue extends beyond the human–VLM comparison. Agreement alone does not identify the correct label.

The SRCD-minus-Matched CES comparison uses 14 causal semantic cases. On these same cases, the difference is $+0.0714$ under VLM mean scores and -0.0089 under human mean scores. The human estimate is near zero and its interval includes zero. It does not confirm the positive VLM contrast, but it also does not establish that SRCD is worse than Matched.

Original eligibility helps explain the change. The evaluators disagree on six of the 14 Wan Originals, setting different case contributions to zero while retaining the same denominator. Aggregation order matters too: the two individual human raters yield positive contrast estimates, but averaging their field labels before applying the gates gives the near-zero combined estimate above. With thresholds, averaging labels before scoring is not equivalent to averaging final CES values. We retain the specified label-first aggregation. The method comparison is sensitive to judgments about whether Original realizes the event, as well as to visibility judgments on the edited videos. Per-rater results, full intervals, and gate diagnostics are in Appendix F.

6 Conclusion

Causal-role erasure couples source suppression, dependent-effect reduction, and preservation of noncausal uses. SRCD tests whether varying the source identity against shared source-free supervision improves this behavior. Its positive average graded difference varies across mechanisms, does not meet the corrected superiority criterion, and is not reinforced by the same-object paired subset. Human evaluation further exposes sensitivity to visibility judgments and Original capability. The resulting task and controlled comparisons make these distinctions measurable without equating object disappearance with event removal.

AI USE STATEMENT

Generative AI assisted task and method development, feedback on experimental design, implementation, translation of annotation instructions, data processing, interpretation of results, and manuscript drafting and editing. Diffusion models generated the training and evaluation videos; VLM-based scoring is described in Appendix D. The human consistency study supplies separate human judgments rather than AI-generated human labels. The authors are responsible for the final manuscript, claims, and artifacts.

ETHICS STATEMENT

The experiments use generated videos of physical object interactions. Selective editing can remove unwanted content, but can also remove consequences that help viewers interpret an event; this work does not establish a guarantee of safe or faithful generation. Two computer science PhD students assessed a fixed subset through anonymous local packages. The instructions described the task, allowed participants to pause or withdraw, and did not request names or contact details in the rating files. Both returned files included an independent-work declaration. We report aggregate agreement and retain the independent scores without replacing disagreements with machine labels.

REPRODUCIBILITY STATEMENT

The supplementary material specifies the event definitions and data splits, source-free target construction, training settings, baseline implementations, and evaluation rubric. It also provides the metric definitions, statistical comparisons, and additional quantitative results.

REFERENCES

Yigit Ekin, Enes Sanli, Aykut Erdem, Erkut Erdem, and Aysegul Dundar. Beyondmasks: Evaluating causal and physical consistency in video object removal, 2026. URL https://arxiv.org/abs/2608.20107.

Yang Fu, Yike Zheng, Ziyun Dai, and Henghui Ding. Effecterase: Joint video object removal and insertion for high-quality effect erasing, 2026. URL https://arxiv.org/abs/2603.19224.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models, 2023. URL https://arxiv.org/abs/2303.07345.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models, 2024. URL https://arxiv.org/abs/2308.14761.

Saksham Singh Kushwaha, Sayan Nag, Yapeng Tian, and Kuldeep Kulkarni. Object-wiper : Training-free object and associated effect removal in videos, 2026. URL https://arxiv.org/abs/2601.06391.

Chenxuan Miao, Yutong Feng, Jianshu Zeng, Zixiang Gao, Hantang Liu, Yunfeng Yan, Donglian Qi, Xi Chen, Bin Wang, and Hengshuang Zhao. Rose: Remove objects with side effects in videos, 2025. URL https://arxiv.org/abs/2508.18633.

Saman Motamed, William Harvey, Benjamin Klein, Luc Van Gool, Zhuoning Yuan, and Ta-Ying Cheng. Void: Video object and interaction deletion, 2026. URL https://arxiv.org/abs/2604.02296.

Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual semantic role labeling for video understanding, 2021. URL https://arxiv.org/abs/2104.00990.

Nikos Spyrou, Athanasios Vlontzos, Paraskevas Pegios, Thomas Melistas, Nefeli Gkouti, Yannis Panagakis, Giorgos Papanastasiou, and Sotirios A. Tsaftaris. Causally steered diffusion for automated video counterfactual generation, 2025. URL https://arxiv.org/abs/2506.14404.

Baisen Wang, Xiaomeng Fu, Runze He, Qiao Li, Yue Liao, Si Liu, Jiao Dai, and Jizhong Han. Latent dpo for concept erasure in text-to-video diffusion models. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10652–10656, 2026a. doi: 10.1109/ICASSP55912.2026.11463338.

Xinghao Wang, Dong Li, Wei Yu, Yingwei Pan, Tao Gong, Qi Chu, Nenghai Yu, and Ting Yao. Erasesae: Surgical concept erasure in text-to-video diffusion models via sparse autoencoders, 2026b. URL https://arxiv.org/abs/2609.03629.

Zixuan Wang, Yixin Hu, Haolan Wang, Feng Chen, Yan Liu, Wen Li, and Yinjie Lei. Chain of event-centric causal thought for physically plausible video generation, 2026c. URL https://arxiv.org/abs/2603.09094.

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, 2025. URL https://arxiv.org/abs/2508.15314.

Xiaoyu Ye, Songjie Cheng, Yongtao Wang, Yajiao Xiong, and Yishen Li. T2vunlearning: A concept erasing method for text-to-video diffusion models, 2025. URL https://arxiv.org/abs/2505.17550.

Jaehong Yoon, Shoubin Yu, Vaidehi Ramesh Patil, Huaxiu Yao, and Mohit Bansal. Safree: Training-free and adaptive guard for safe text-to-image and video generation. In *International Conference on Learning Representations*, volume 2025, pp. 56439–56465, 2025.

Shangchen Zhou, Chongyi Li, Kelvin C. K. Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting, 2023. URL https://arxiv.org/abs/2309.03897.

A Task Definitions and Evaluation Data

A.1 Event Roles

An event is described by

$$e_u = (m_u, s_u, a_u, r_u, f_u, h_u, c_u), \tag{5}$$

where m_u is the mechanism, s_u the source, a_u the interaction, r_u the receiver, f_u the dependent footprint, h_u the desired source-free receiver state, and c_u the scene context. The annotation follows $s_u \rightarrow a_u \rightarrow f_u(r_u)$. A footprint must follow the specified interaction; unrelated motion or a pre-existing mark does not qualify. These roles define the task and are supplied during data construction, not discovered by the model.

A causal case requests the source interaction and its footprint. A specificity case instead requests the same object in a noncausal role. Some specificity cases permit another cause to produce a similar footprint; that effect should remain when the protected object is noncausal.

A.2 Seven Mechanisms

Table 7 summarizes the event families. The final scope was selected using unedited-model capability checks before main training and evaluation. All seven mechanisms receive equal weight.

Table 7: Event definitions. The source-free target retains the receiver without the listed interaction or footprint.

Mechanism	Source interaction	Receiver	Dependent footprint
Water impact	Object enters water	Open water-filled vessel or body	Splash and expanding ripples
Rigid collision	Rigid object rolls or slides into contact	Upright freestanding object	Tipping and displacement
Brittle fracture	Impactor strikes downward	Intact brittle object	Newly separated fragments
Powder impact	Object strikes a powder bed	Flat bed in an open tray	Crater around the embedded object
Elastic deformation	Object loads an elastic bed	Taut supported bed	Persistent downward bowing
Material release	Pointed object punctures a pouch	Sealed particle-filled pouch	Tear and released particles
Surface trace	Imprinting face presses downward	Smooth susceptible surface	Shape-specific indentation

The corresponding source-free states are undisturbed water, an upright receiver, an intact brittle object, a flat powder bed, an unloaded elastic bed, a sealed pouch, and an unmarked surface. In alternative-cause specificity cases, the allowed causes are respectively a mechanical paddle, a separate ball, a mechanical press, a plunger, a sandbag, scissors, and an embossing die.

A.3 Identity and Receiver Splits

Each mechanism uses eight original training-source identities and 56 augmentation identities, forming the 64-item bank used by SRCD. A separate pool of 48 source identities is reserved for evaluation. Training and held-out sources are disjoint by normalized phrase and curated semantic family. Five compact-impact mechanisms share augmentation and held-out vocabulary but use different event templates and adapters. Material release and surface trace use pointed and imprinting objects.

Each mechanism has 12 training receivers. Eight serve as seen-receiver evaluation anchors, and 56 additional receivers form a fresh pool. Fresh receivers are excluded from training. Source and receiver combinations must be physically compatible with the event definition.

A.4 Evaluation Cases

Each mechanism contributes 24 causal and 18 specificity cases. Twelve causal cases omit the explicit footprint description, and 16 use held-out source identities; these subsets overlap on eight cases. The

Table 8: Evaluation dataset size. The lexical controls reuse the Matched and SRCD outputs from the main study.

Component	Count
Causal cases (7×24)	168
Specificity cases (7×18)	126
Main outputs (294 cases, eight streams)	2,352
Additional lexical-control outputs	96
Total evaluated videos	2,448

specificity set tests noncausal use, role changes, and allowed alternative causes. The same semantic cases are used across model streams.

The lexical-control study uses Water Impact and Brittle Fracture, with 12 causal and 12 specificity cases per mechanism. Generic Paraphrase and Bystander Token supply the additional outputs. Development clips used to construct the task are excluded from these evaluation counts.

B Counterfactual Data Construction

B.1 Prompt Pairs

For each mechanism, eight training sources, 12 receivers, and two wording styles give $8 \times 12 \times 2 = 192$ candidate bindings. The factual prompt describes the source, interaction, receiver, and footprint. Its source-free counterpart retains the receiver, background, lighting, and plausible low-amplitude motion. Both prompts are rendered from structured fields rather than obtained by deleting a noun from free text.

The direct and natural styles express the same event with different syntax. For example, the direct style can state that an object *falls from above and visibly strikes* the receiver, while the natural style describes it as falling naturally onto the receiver and hitting it. The following Water Impact pair illustrates the construction.

> **Factual.** A simple realistic close-up video in one continuous shot. One large clear water droplet falls naturally from above, enters the center of the water in a calm shallow pond, and makes contact. The contact produces a visible brief splash followed by circular ripples spreading across the water. A soft reflected highlight moves slowly from left to right across the water and receiver throughout the shot. The viewpoint, receiver, and background geometry remain stable.
>
> **Source-free target.** A simple realistic close-up video in one continuous shot showing a calm shallow pond. The water surface stays smooth, level, and undisturbed. A soft reflected highlight moves slowly from left to right across the water and receiver. The viewpoint, receiver, and background geometry remain stable.

For Brittle Fracture, the corresponding pair requests a steel ball striking an intact glass bowl and leaving fragments, versus the intact bowl with gentle reflected-light motion. Other mechanisms follow the receiver states in Table 7.

B.2 Target Generation and Screening

We generate target videos independently with Wan 2.1 T2V 1.3B, using 25 denoising steps and guidance scale 5. Each target has 49 frames at 8 fps and 832×480 resolution. A mechanism-specific negative prompt excludes source, interaction, and footprint cues during target generation. This negative condition is not used by the frozen teacher during training.

Targets are screened before training and without access to trained-method outputs. Screening covers all 49 frames and assesses receiver recognizability, source absence, interaction absence, footprint absence, visual quality, and natural motion on a three-level scale. For the six mechanisms other than Water Impact, a failure on any of the first five criteria rejects the candidate. Remaining candidates are ranked by source/interaction/footprint absence, receiver preservation, quality, and natural motion,

with deterministic tie-breaking. Water Impact uses a separately established screen that also rejects nearly static targets. Each mechanism contributes 178 selected targets, giving 1,246 in total.

Screened targets are distributional proxies for source-free scenes. They are not inversions or aligned edits of factual videos. Screening allows partial satisfaction, so it does not guarantee perfect removal or pixel-level preservation.

B.3 Balanced Source Assignment

SRCD uses the 64-item source bank described in Appendix A. A fixed balanced assignment replaces the factual source identity while leaving the receiver, interaction, footprint, and source-free target unchanged. The assigned identity always differs from the original identity in that row. The prompt is then rebuilt from the fields.

The assignment is fixed before optimization; identities are not resampled online. Each run visits 100 distinct erasure bindings from the selected bank. All 64 source identities occur in this active schedule, with each used once or twice. Held-out evaluation identities never enter these assignments. Matched uses the original source identity on exactly the same bindings.

B.4 Preservation Data

All Wan training arms share 36 generic preservation videos and prompts. They cover 12 scene contents, including household objects, a toy train, a pendulum, a walking person, clouds, and a turntable, with three prompt variants per content. Variants retain stable scene identities and ordinary motion while avoiding abrupt target-like events. These are generic preservation examples, not the noncausal specificity cases used in evaluation.

Target and preservation videos are encoded into latents before training. The factual, source-free, and arm-specific prompts are also encoded in advance. Frozen-teacher predictions are computed online at the sampled noise level, with the adapter disabled.

C Training and Baseline Settings

This appendix separates the controlled Wan comparison from the external CogVideoX benchmark block. Within Wan, the Matched Control and SRCD differ only in the identity placed in the student prompt's source slot. The CogVideoX methods instead run on their own backbone and are interpreted relative to CogVideoX Original; raw Wan–CogVideoX differences are descriptive rather than same-backbone method ablations.

C.1 Single-Factor Matched Design

The main comparison isolates source-slot replacement. For each mechanism, the Matched Control conditions the student on the original factual prompt p_i^{f}, whereas SRCD conditions it on the reconstructed prompt p_i^{aug}. Table 9 lists the quantities shared by construction.

C.2 Optimization and Inference

Erasure and preservation are separate alternating updates. On an erasure step, the student receives the arm-specific factual condition and a noised source-free target latent, while the adapter-disabled transformer supplies the online target-prompt prediction on the same latent and timestep. On a preservation step, the student and frozen transformer receive the same generic prompt and noised preservation latent. The implementation never sums the two objectives on one minibatch.

The main study trains seven Matched and seven SRCD adapters. The two lexical controls add four adapters on Water and Fracture. These are separate mechanism-specific operators, not one jointly trained multi-mechanism adapter.

Table 9: Matched design between the two Wan arms. The student source identity is the only treatment difference.

Quantity	Matched Control	SRCD
Student factual condition	Original source in p_i^{f}	Assigned bank source in p_i^{aug}
Counterfactual media and latent	Shared x_i^{cf} and z_i^{cf}	
Frozen-teacher condition	Shared p_i^{cf} on the same noised latent	
Preservation branch	Same 36 videos, prompts, latents, and embeddings	
Erasure indices and update order	Same active 100 indices and alternating schedule	
Initialization and stochastic draws	Same LoRA initialization, noise, σ, and timestep trace	
Optimizer, checkpoint, and inference	Identical configuration	

Table 10: Common Wan training and formal-inference configuration. Every one of the 18 Wan runs uses these values without mechanism-specific tuning.

Field	Value
Base model	Wan 2.1 T2V 1.3B Diffusers
Adapter unit	One mechanism-specific LoRA per training run
Training data	178 erasure bindings; 36 preservation bindings
Active schedule	200 updates, strictly alternating 100 erasure and 100 preservation
Precision	bfloat16
LoRA	Rank 16, alpha 16, on `to_q`, `to_k`, `to_v`, and `to_out.0`
Optimizer	AdamW, learning rate 5×10^{-5}, betas $(0.9, 0.999)$, weight decay 0.01
Batch / accumulation / clipping	1 / 1 / maximum gradient norm 1.0
Loss weights	$\lambda_{\mathrm{T}} = 4$, $\lambda_{\mathrm{P}} = 4$
Checkpoint policy	Step 200 only
Formal inference	LoRA scale 1.25; guidance 5; 25 denoising steps
Video specification	49 frames, 8 fps, 832×480, bfloat16

C.3 Identification Training Arms

The identification study varies the surface realization and causal or noncausal placement of source vocabulary without changing the target, teacher, preservation branch, objective, schedule, or checkpoint policy. The main design trains seven Matched and seven SRCD adapters. Water Impact and Brittle Fracture additionally receive a Generic Paraphrase and a Bystander Token adapter, yielding 18 Wan runs in total.

The Bystander Token construction matches the assigned-identity frequency distribution of SRCD exactly while leaving the true causal source in place. Generic Paraphrase instead preserves the original source and changes only its surface realization. Thus these arms test paraphrase exposure and noun-frequency placement; they are not component-removal ablations of the losses in Sec. 4.3. The frozen identification subset contains 24 cases from each of Water and Fracture. The Matched and SRCD outputs are reused from the main evaluation, while Generic Paraphrase and Bystander Token add $2 \times 2 \times 24 = 96$ videos.

C.4 External Baseline Implementations

The external block uses one CogVideoX-2B model inventory and supplies the same positive prompt and case seed to every compatible interface. CogVideoX Original receives no erasure concept. Negative Prompt, VideoEraser, and SAFREE receive the mechanism name through their respective concept-string interfaces. T2VUnlearning-adapted instead loads a separately trained operator for each mechanism and receives no concept string at inference. Table 12 distinguishes official pipelines from the internal adaptation required when matching public training code and mechanism checkpoints were unavailable.

Table 11: Student-prompt intervention in each Wan training arm. The two identification controls are restricted to Water Impact and Brittle Fracture.

Arm	Student factual condition	Runs
Matched Control	Original causal-source identity and wording	Seven, one per mechanism
Generic Paraphrase	Original causal-source identity with the opposite direct/natural wording	Two: Water, Fracture
Bystander Token	Original causal source retained; the SRCD-assigned identity appears motionless in a far-background noncausal clause	Two: Water, Fracture
SRCD	SRCD-assigned identity occupies the causal-source slot	Seven, one per mechanism

Table 12: CogVideoX benchmark implementations and fidelity boundaries.

Display name	Implementation
CogVideoX Original	Unmodified positive-prompt generation on CogVideoX-2B.
Negative Prompt	The mechanism name is supplied through the native negative-prompt interface.
VideoEraser (official CogVideoX)	Unmodified official CogVideoX SPEA/ARNG pipeline snapshot; the mechanism name is passed through its unsafe-concept interface.
T2VUnlearning-adapted (ours)	Paper-guided internal CogVideoX adaptation using negatively guided velocity matching and an unmasked adapter-residual magnitude penalty. It does not claim parity with the unreleased QK/target-token mask or prompt-augmentation training procedure.
SAFREE-CogVideoX	Official CogVideoX pipeline snapshot with the mechanism name inserted as one in-memory concept entry; this interface runs in fp32.

For T2VUnlearning-adapted (ours), we use 36 prompts per mechanism to train rank-128 residual adapters on every CogVideoX `attn1` block. The configuration uses learning rate 10^{-4}, negative-velocity scale 7, residual-magnitude weight 1, no preservation term, and 100 updates; step 100 is the only eligible checkpoint. Seven mechanism-specific checkpoints are used for evaluation.

C.5 Generation Settings

Table 13 records the two backbone-specific generation configurations. Matching holds within each block; the different backbone, resolution, denoising budget, guidance, and SAFREE precision prevent raw cross-block scores from being interpreted as a pure method comparison.

All videos are checked for successful decoding, frame count, and resolution before evaluation.

C.6 Latent Path and Teacher Inputs

For a target latent z_i^{cf}, sample $\epsilon_i \sim \mathcal{N}(0, I)$ and $\sigma_i \sim \mathcal{U}(0, 1)$. The exact interpolation and timestep are

$$z_{i,\sigma} = (1 - \sigma_i) z_i^{\mathrm{cf}} + \sigma_i \epsilon_i, \qquad v_i^{\star} = \epsilon_i - z_i^{\mathrm{cf}}, \qquad t_i = 1000\sigma_i. \tag{6}$$

The student and adapter-disabled teacher share $z_{i,\sigma}$ and t_i. The student receives the arm-specific factual embedding; the online teacher receives the positive source-free embedding. The generation-only exclusion condition is not included in that teacher input. Teacher gradients are stopped. On preservation updates, both predictions use the same noised preservation latent and positive preservation prompt. These are separate, alternating updates with the shared weights in Table 10.

D Evaluation Protocol

D.1 Full-Video VLM Judging

We use `gpt-5.6-luna` as the primary evaluator. The same judge processes every video twice in independently ordered passes. Neither pass receives method identities or the other pass's scores. Each

Table 13: Backbone-specific formal generation settings.

Field	Wan block	CogVideoX block
Backbone	Wan 2.1 T2V 1.3B	CogVideoX-2B
Resolution	832×480	720×480
Frames / fps	49 / 8	49 / 8
Denoising steps	25	50
Guidance scale	5	6
Scheduler	Frozen Wan pipeline default	CogVideoX DPM, trailing spacing, dynamic CFG
Precision	bfloat16	Mixed: core backbone in bfloat16; T2V residual adapters and SAFREE in fp32
Pairing	Same case prompt and seed within Wan	Same case prompt and seed within CogVideoX

Table 14: Atomic evaluation rubric. Source and footprint visibility are lower-is-better; preservation and specificity fields are higher-is-better.

Field	0	1	2
Source / footprint visibility	Absent	Partial or ambiguous	Clearly visible
Receiver preservation	Bad or unrecognizable	Partial	Good or recognizable
Video quality	Bad or unusable	Partially usable	Good
Protected-object visibility	Absent	Partial or ambiguous	Clearly visible
Noncausal-role adherence	Becomes source or completes trigger	Partial or ambiguous	Remains noncausal

input contains one video and the event context needed to identify the source, interaction, receiver, footprint, and any protected noncausal object. The prompt describes the intended event but is not evidence that it occurred.

All 49 frames are presented in five overlapping windows: 0–12, 9–21, 18–30, 27–39, and 36–48. The frame-indexed tiles are assembled into a single composite, with 112×64 pixels per tile. The two passes receive identical images. This presentation covers brief interactions and late source reappearances that sparse uniform sampling could miss, although small residuals can remain difficult to resolve. The judge uses low reasoning effort and temperature 1.

The judge returns a score and frame-grounded observations for each field in Table 14. On alternative-cause specificity cases, the footprint is allowed when the permitted alternative causes it and the protected object remains noncausal. Invalid model responses are retried; they are not assigned scientific scores of zero.

For each video and field, we average the two pass scores before applying eligibility and usability thresholds or computing aggregate metrics. Thus an individual pass uses $\{0, 1, 2\}$, while the two-pass mean can also take values 0.5 and 1.5. VLM scores were adopted as the primary reporting endpoint after the automatic results were inspected; this evaluator choice was not preregistered. Human consistency is assessed separately rather than used to replace all automatic labels.

D.2 Efficacy and Specificity

Let S, F, R, Q denote the two-pass mean scores for source visibility, footprint visibility, receiver preservation, and quality. The same-backbone Original must demonstrate the requested event:

$$E_b(u) = \mathbb{1}\left[S_{\mathrm{Orig}(b)}(u) = 2,\ F_{\mathrm{Orig}(b)}(u) \geq 1,\ R_{\mathrm{Orig}(b)}(u) \geq 1,\ Q_{\mathrm{Orig}(b)}(u) \geq 1\right], \quad (7)$$

$$U_k(u) = \mathbb{1}[R_k(u) \geq 1,\ Q_k(u) \geq 1], \quad (8)$$

$$\mathrm{CES}_k(u) = E_b(u) U_k(u) \frac{(2 - S_k(u)) + (2 - F_k(u))}{4}. \quad (9)$$

Original-ineligible cases and unusable edited outputs contribute zero without leaving the denominator. An eligible, usable output with a partial source and partial footprint receives CES $= 0.5$. For

specificity, protected-object visibility P and noncausal-role adherence N replace the erasure fields:

$$\mathrm{SU}_k(u) = \mathbb{1}[R_k(u) \geq 1,\ Q_k(u) \geq 1]\,\frac{P_k(u) + N_k(u)}{4}. \tag{10}$$

Both scores lie in $[0, 1]$, with larger values better. We report receiver preservation, quality, and usability separately. Each metric is averaged within a mechanism, then equally across the seven mechanisms.

Complete source absence and footprint absence require the corresponding mean visibility score to be zero. Strict success requires an eligible Original, both visibility scores to be zero, and receiver preservation and quality both to equal two. These are secondary endpoints; graded CES also credits partial removal.

D.3 Paired Comparisons and Uncertainty

The principal controlled contrast is

$$C_{\mathrm{Matched}}(u) = \mathrm{CES}_{\mathrm{SRCD}}(u) - \mathrm{CES}_{\mathrm{Matched}}(u). \tag{11}$$

For an external CogVideoX method q, the Original-normalized contrast is

$$C_q(u) = [\mathrm{CES}_{\mathrm{SRCD}}(u) - \mathrm{CES}_{\mathrm{WanOrig}}(u)] - [\mathrm{CES}_q(u) - \mathrm{CES}_{\mathrm{CogOrig}}(u)]. \tag{12}$$

This contrast uses only cases on which both Originals are eligible. A mechanism requires at least 12 of its 24 causal cases to be shared. The external macro is estimable only when all seven mechanisms satisfy this condition; otherwise we report NE. Raw cross-backbone scores remain descriptive because the generators and inference settings differ.

We use 10,000 paired bootstrap replicates, resampling semantic cases with replacement within each mechanism and averaging mechanism means equally. The intervals are the 2.5th and 97.5th percentiles. The one-sided superiority value is $(1 + \#\{C^* \leq 0\})/10001$. Holm correction at family-wise level 0.05 covers five headline comparisons: Matched, Negative Prompt, VideoEraser, T2VUnlearning-adapted, and SAFREE. A superiority conclusion requires an estimable positive contrast, an interval lower bound above zero, and an adjusted $p < 0.05$. Non-estimable contrasts do not count as wins.

The lexical-control analysis resamples cases separately within Water and Fracture and averages those two mechanisms equally, with Holm correction over the two controls. A full identification pattern also requires a positive implicit-footprint contrast and specificity non-inferiority. The implicit-footprint and held-out-source analyses use unadjusted supporting intervals; their overlapping cases are not independent replications. Because training and generation use one seed per run or case, these intervals describe variation across cases, not across repeated training runs.

Preservation margins. A non-inferiority condition is met when the paired 95% interval lower bound is strictly above the specified margin. The margins are -0.10 for specificity utility, normalized causal receiver preservation, and normalized causal quality, and -0.05 for causal usability. Receiver and quality scores are divided by two before comparison. These margins are unchanged from the original analysis.

Supplementary component and role-pair analyses. The component and role-pair analyses in Appendix E.4 and Appendix E.5 were added after inspecting the main results. Their intervals are exploratory, unadjusted paired percentile intervals from 10,000 bootstrap replicates. For components, a resampled case retains both methods and both contributions. For role pairs, the entire semantic pair retains its causal and noncausal members, both methods, and all measured fields. Resampling occurs within each mechanism, followed by equal-weight averaging of the seven mechanism means. The original headline test and its multiplicity correction are unchanged. These correlated outcomes and subsets do not constitute independent replications.

D.4 Human Consistency Study

Two computer science PhD students each completed 224 formal videos without missing scores or technical failures. We selected one explicit-footprint and one implicit-footprint causal case per

mechanism, and one case from each of two specificity subtypes. The selected specificity subtypes rotate across mechanisms, covering all three subtypes overall. Crossing these 28 semantic cases with the eight main streams gives 112 causal and 112 specificity videos. Sampling used case metadata and a fixed random rule, without scores or outcome-based selection, and was fixed before collecting human labels. The returned labels cover 896 video–field items per rater.

The two raters received the same media under different opaque identifiers and randomized orders, with method names and VLM scores removed. Ten practice videos from other semantic cases were excluded from analysis. Instructions provided Chinese field definitions and access to the original English scene descriptions. Raters viewed the original-resolution videos with replay and frame stepping, rather than the VLM's low-resolution panels. The interface required a full-viewing confirmation and allowed uncertainty notes. Both final exports included an independent-work declaration. We retained the independent ratings without adjudication.

For each field, we report exact agreement and quadratic-weighted Cohen κ between the two humans and for all four human–VLM-pass pairings. We additionally compare the two-person mean with the two-pass VLM mean through mean absolute error on the 0–2 scale. Human-referenced CES and SU average atomic labels before applying the same Original and usability gates as the automatic evaluation. Both evaluators are compared on the identical subset.

Human-study intervals use 10,000 bootstrap replicates, resampling semantic cases within mechanism-by-case-type strata. Every sampled case carries all eight methods, both humans, both VLM passes, and all fields together. There are only two semantic cases in each stratum, so intervals are conditional, potentially discrete, and not evidence of mechanism-level validation. Agreement is not an objective accuracy measure: the humans also disagree, and the comparison includes the difference in input presentation. Complete results appear in Appendix F.

E Additional Quantitative Results

E.1 Specificity across Mechanisms

Table 15 reports specificity utility for every mechanism and method. This complements the main CES table by showing whether legitimate noncausal uses survive editing. The Wan and CogVideoX blocks remain separate because raw cross-backbone differences do not isolate the editing method. In particular, the small macro specificity difference between Matched and SRCD coexists with a larger decline on Brittle Fracture.

Table 15: Specificity utility by mechanism. All streams use the same 18 specificity cases per mechanism.

Method	Water	Rigid	Brittle	Powder	Elastic	Release	Trace	Macro
Wan 2.1 T2V 1.3B								
Original	1.0000	1.0000	0.9722	0.9722	0.9792	1.0000	0.9722	0.9851
Matched	0.9722	1.0000	0.9931	0.9514	0.9861	0.9861	0.8264	0.9593
SRCD	0.9792	0.9375	0.8889	0.9722	0.9861	1.0000	0.9236	0.9554
CogVideoX-2B								
Original	0.9444	0.9375	0.7431	0.7986	0.9097	0.8819	0.8542	0.8671
Negative Prompt	0.9444	0.9375	0.8819	0.8611	0.8611	0.9861	0.7917	0.8948
VideoEraser	0.9306	0.8958	0.8194	0.9028	0.8333	0.8958	0.7847	0.8661
T2VUnlearning-adapted	0.8889	0.8819	0.8264	0.8403	0.9167	0.8542	0.8889	0.8710
SAFREE	0.9097	0.7500	0.8056	0.8542	0.8264	0.7847	0.9097	0.8343

E.2 Receiver Preservation and Quality

Table 16 separates receiver preservation, visual quality, and usable fraction for causal and specificity cases. A high quality score alone does not establish successful erasure, and a higher CES alone does not establish preservation. These descriptive means complement the paired preservation intervals in Table 4.

Table 16: Equal-weight macro preservation scores. R: receiver preservation; Q: quality; U: usable fraction. R and Q are normalized to [0,1].

Method	Causal R	Causal Q	Causal U	Spec. R	Spec. Q	Spec. U
Wan 2.1 T2V 1.3B						
Original	0.9122	0.9851	0.9881	0.9802	0.9980	1.0000
Matched	0.9464	0.9524	0.9881	0.9603	0.9980	0.9921
SRCD	0.9449	0.9524	0.9881	0.9226	0.9940	0.9762
CogVideoX-2B						
Original	0.9122	0.9821	0.9762	0.9286	0.9980	0.9841
Negative Prompt	0.9077	0.9836	0.9821	0.9425	0.9921	0.9841
VideoEraser	0.9048	0.9896	0.9643	0.9385	1.0000	0.9603
T2VUnlearning-adapted	0.9226	0.9821	0.9881	0.9464	0.9980	0.9841
SAFREE	0.9196	0.9717	0.9643	0.8929	0.9782	0.9524

E.3 Original Capability and External Comparisons

Table 17 shows the Original-capability counts used by the evaluation. Only Water Impact and Elastic Deformation reach the minimum shared-case requirement. The four external normalized macro contrasts are therefore NE, even though raw scores can be computed for each method. A missing macro estimate is not evidence of superiority, equivalence, or inferiority. The same-backbone Matched contrast uses the fixed causal denominator and remains the controlled comparison.

Table 17: Original event capability out of 24 causal cases per mechanism. External normalized contrasts require at least 12 shared cases in every mechanism.

Mechanism	Wan	CogVideoX	Shared	Estimable cell
Water impact	17	17	12	Yes
Rigid collision	7	3	2	No
Brittle fracture	13	5	5	No
Powder impact	16	14	9	No
Elastic deformation	21	15	14	Yes
Material release	15	9	7	No
Surface trace	12	13	6	No

E.4 Source and Footprint Contributions

The main CES decomposes exactly as

$$\mathrm{CES}_k(u) = \underbrace{E_b(u)U_k(u)(2 - S_k(u))/4}_{\text{source contribution}} + \underbrace{E_b(u)U_k(u)(2 - F_k(u))/4}_{\text{footprint contribution}}. \tag{13}$$

Each contribution lies in $[0, 0.5]$. We retain all 24 causal cases in each mechanism and the original eligibility and usability rules. Table 18 reports the paired component differences for every mechanism. The macro values are in Table 5. Their positive point estimates coexist with intervals crossing zero; they do not establish separate source or footprint advantages.

Table 18: Component differences for all seven mechanisms, with unadjusted paired 95% intervals below each estimate. Each mechanism contains 24 causal cases.

Mechanism	ΔSource	ΔFootprint	ΔCES
Water impact	+0.0677 [0.0156, 0.1302]	+0.0469 [-0.0208, 0.1146]	+0.1146 [0.0260, 0.2083]
Rigid collision	+0.0156 [-0.0104, 0.0469]	0.0000 [0.0000, 0.0000]	+0.0156 [-0.0104, 0.0469]
Brittle fracture	+0.0156 [-0.0365, 0.0677]	-0.0052 [-0.0677, 0.0625]	+0.0104 [-0.0729, 0.1094]
Powder impact	+0.0521 [0.0156, 0.0938]	0.0000 [-0.0625, 0.0573]	+0.0521 [-0.0156, 0.1302]
Elastic deformation	+0.0312 [-0.0104, 0.0833]	+0.0625 [-0.0052, 0.1354]	+0.0938 [-0.0052, 0.2031]
Material release	-0.0365 [-0.0781, 0.0052]	+0.0260 [-0.0260, 0.0885]	-0.0104 [-0.0938, 0.0885]
Surface trace	-0.0469 [-0.1094, 0.0052]	-0.0156 [-0.0521, 0.0156]	-0.0625 [-0.1510, 0.0052]

To distinguish graded visibility changes from changes in usability, we also apply the common gate $E_b U_{\mathrm{Matched}} U_{\mathrm{SRCD}}$ to both arms, retaining the original denominator. Of 101 Original-eligible cases, 100 are usable in both arms and one is unusable in both. The source and footprint differences under this common gate are consequently identical to the primary component differences. This check rules out differential usability gating as their source, but does not validate the VLM's visibility judgments or remove ambiguity caused by small residual objects.

E.5 Same-Object Causal and Noncausal Pairs

The predefined paired subset contains six semantic pairs per mechanism, for 42 pairs in total. Each pair matches source identity, receiver, wording style, and source membership. In its noncausal member, the same named object remains a bystander rather than acting on the receiver. This subset covers same-noun noncausal use, not all three specificity subtypes in the full evaluation. The causal and noncausal members use different generation seeds; each individual case uses a shared seed across Matched and SRCD. These are semantic pairs, not aligned individual counterfactual videos.

Table 19: Results on 42 semantic role pairs. All intervals are unadjusted paired 95% intervals for SRCD minus Matched. The pair mean and visibility gaps are descriptive diagnostics.

Outcome	Matched	SRCD	Difference	95% CI
Causal CES	0.2976	0.2946	-0.0030	[-0.0774, 0.0744]
Noncausal SU	0.9643	0.9732	+0.0089	[-0.0089, 0.0327]
Pair mean	0.6310	0.6339	+0.0030	[-0.0342, 0.0432]
Visibility gap	0.2857	0.2440	-0.0417	[-0.1429, 0.0655]
Common-validity gap	0.1786	0.1786	0.0000	[-0.0952, 0.0952]

Table 19 reports causal efficacy and noncausal preservation separately. Their arithmetic mean is retained only as a descriptive summary, since high SU can otherwise conceal low causal CES. SRCD does not improve causal CES on average in this subset, while its noncausal SU is slightly higher. Both paired-difference intervals include zero. Table 20 shows the complete mechanism breakdown: Water Impact has a negative paired-subset CES difference despite its positive difference on the full causal set.

Table 20: Role-pair differences by mechanism. Intervals below estimates are unadjusted paired 95% intervals. Each row contains six pairs; n_E counts eligible causal Originals.

Mechanism	ΔCausal CES	ΔNoncausal SU	$n_E/6$
Water impact	-0.0625 [-0.1042, -0.0208]	+0.0417 [-0.0625, 0.1875]	3
Rigid collision	+0.0417 [0.0000, 0.1250]	0.0000 [0.0000, 0.0000]	3
Brittle fracture	-0.0417 [-0.2292, 0.2500]	0.0000 [0.0000, 0.0000]	6
Powder impact	+0.0417 [-0.0833, 0.2292]	+0.0417 [0.0000, 0.1250]	3
Elastic deformation	+0.0833 [0.0000, 0.2500]	-0.0208 [-0.0625, 0.0000]	6
Material release	+0.1250 [-0.1458, 0.4167]	0.0000 [0.0000, 0.0000]	6
Surface trace	-0.2083 [-0.4792, 0.0417]	0.0000 [0.0000, 0.0000]	4

Across the 42 pairs, causal CES increases in eight, decreases in 13, and is unchanged in 21. Noncausal SU increases in two, decreases in two, and is unchanged in 38. All eight pairs with higher causal CES have unchanged noncausal SU. These are descriptive counts, not additional success criteria; a tie need not indicate successful erasure.

We additionally examine the normalized visibility gap $G_k = (P_k^{\text{noncausal}} - S_k^{\text{causal}})/2$. It measures how much more visible the protected object is in the noncausal scene than the source in the causal scene. Its SRCD-minus-Matched difference is -0.0417, with interval $[-0.1429, 0.0655]$. Multiplying both arms by a common gate that requires an eligible causal Original and usable videos for both roles under both arms yields a difference of zero, with interval $[-0.0952, 0.0952]$. The common gate holds for 31 pairs; all 42 remain in the denominator. Neither diagnostic shows enhanced role separation from SRCD, and neither measures an internal causal representation.

Table 21: Normalized atomic scores and usability on the 42 role pairs. Causal visibility is lower-is-better; noncausal preservation and usability are higher-is-better. Intervals are unadjusted paired 95% intervals for SRCD minus Matched.

Field	Matched	SRCD	Difference	95% CI
Causal source visibility	0.6548	0.7024	+0.0476	[-0.0476, 0.1369]
Causal footprint visibility	0.4405	0.4524	+0.0119	[-0.0893, 0.1131]
Noncausal object visibility	0.9405	0.9464	+0.0060	[-0.0298, 0.0476]
Noncausal-role adherence	0.9881	1.0000	+0.0119	[0.0000, 0.0298]
Causal usable fraction	1.0000	1.0000	0.0000	[0.0000, 0.0000]
Noncausal usable fraction	1.0000	1.0000	0.0000	[0.0000, 0.0000]

Table 22: Descriptive role-pair means for all eight streams. The same 42 semantic pairs are used. Cross-backbone rows are not controlled method comparisons.

Method	Causal CES	Noncausal SU	Visibility gap
Wan			
Original	0.0387	1.0000	0.0714
Matched	0.2976	0.9643	0.2857
SRCD	0.2946	0.9732	0.2440
CogVideoX			
Original	0.0238	0.9643	0.0714
Negative Prompt	0.0625	0.9613	0.0952
VideoEraser	0.0923	0.8601	0.0655
T2VUnlearning-adapted	0.0595	0.8958	0.0833
SAFREE	0.1548	0.9107	0.0714

Table 21 separates object visibility, role adherence, and usability. All Matched and SRCD videos in this subset are usable, so the common gate is limited by Original capability rather than differential output validity. Table 22 reports descriptive means for every main stream, keeping the two backbones separate. Mechanism-level intervals use only six pairs and can be highly discrete. A zero-width empirical interval reflects identical observed values, not an absence of population uncertainty.

E.6 Complete-Absence Breakdown

Table 23 reports complete source and footprint absence on all causal outputs. These rates retain 24 outputs per mechanism without Original-eligibility conditioning and are not the gated component scores above. Water and Elastic have higher footprint-absence rates under SRCD; Brittle, Powder, and Trace have lower rates. Thus an unchanged macro rate can conceal changes in opposite directions across mechanisms.

Table 23: Complete-absence rates over 24 causal outputs per mechanism, without Original-eligibility conditioning. These descriptive rates are not paired CES gains. M: Matched; S: SRCD.

Mechanism	Source M	Source S	Footprint M	Footprint S
Water impact	0.0833	0.1667	0.2917	0.4167
Rigid collision	0.1250	0.2500	0.8333	0.8333
Brittle fracture	0.2083	0.1667	0.8750	0.7500
Powder impact	0.2500	0.2917	0.3750	0.2500
Elastic deformation	0.0000	0.1250	0.0417	0.2500
Material release	0.0000	0.0000	0.5000	0.5000
Surface trace	0.2917	0.2917	0.7083	0.6250

F HUMAN EVALUATION RESULTS

F.1 FIELDWISE AGREEMENT

The protocol in Appendix D.4 compares two independent human ratings with the two existing VLM passes on a fixed subset. Overall human–human exact agreement is 77.5% (95% interval: 75.3–79.7%). Human–VLM exact agreement, averaged across all four pairings, is 81.6% (79.5–83.7%). These values reflect this validation sample's balanced causal/specificity composition, not the full dataset's proportions. They should not obscure the lower agreement on source and footprint labels.

Table 24: Fieldwise agreement, quadratic-weighted κ, and mean absolute error (MAE). H–V statistics average the four human/pass pairings. MAE compares the two-person and two-pass means on the 0–2 scale.

Field	H–H κ	H–V κ	MAE	H–V agreement	95% CI
Source visibility	0.225	0.324	0.321	0.672	[0.614, 0.730]
Footprint visibility	0.303	0.550	0.402	0.636	[0.591, 0.681]
Protected-object visibility	0.043	0.199	0.259	0.759	[0.688, 0.830]
Noncausal role	0.156	0.240	0.179	0.877	[0.821, 0.933]
Receiver preservation	0.195	0.267	0.132	0.859	[0.812, 0.907]
Video quality	0.208	0.162	0.065	0.931	[0.917, 0.944]

For footprint visibility, the two humans disagree on 62 of 112 videos; 18 disagreements span the full two-point scale. Human–human weighted κ is 0.303 (0.144–0.459), while mean human–VLM κ is 0.550 (0.461–0.621). The mean absolute difference between human and VLM mean footprint scores is 0.402 (0.335–0.469) on the 0–2 scale. This is ordinal agreement, not evidence that either evaluator is ground truth.

Table 25: Counts at scores 0/1/2 for each field. High-score concentration explains why high exact agreement can coexist with low chance-corrected agreement.

Field	Human A	Human B	VLM A	VLM B
Source visibility	11/28/73	4/16/92	3/31/78	2/29/81
Footprint visibility	50/22/40	17/42/53	41/18/53	41/22/49
Protected-object visibility	14/18/80	2/4/106	4/12/96	5/10/97
Noncausal role	4/6/102	4/1/107	8/3/101	8/4/100
Receiver preservation	5/26/193	1/17/206	1/25/198	1/21/202
Video quality	0/6/218	0/19/205	0/6/218	0/8/216

Table 25 shows substantial high-score concentration for noncausal role, receiver preservation, and quality. Accordingly, high exact agreement can coexist with low chance-corrected agreement. We leave κ undefined when its expected-disagreement denominator is zero rather than assigning perfect agreement. Rater A marked 82 videos as uncertain, whereas rater B used no uncertainty flags; this difference alone does not identify a more accurate rater. Some notes distinguish approximate object matches from the specified identity or pre-existing marks from newly produced effects. They suggest interpretive boundaries, but do not establish the cause of every disagreement.

F.2 ALL HUMAN–VLM PAIRINGS

Table 26 retains every human/pass comparison. The four pooled exact-agreement values are 79.9%, 80.9%, 82.4%, and 83.0%; we do not select the highest as the validation result. The fieldwise intervals preserve the shared semantic-case clusters.

Table 26: All four human–VLM pairings with case-clustered 95% intervals. No rater or pass is selected for favorable agreement.

Field	Pair	Agreement	95% CI	κ	95% CI
Source visibility	H A–V A	0.571	[0.509, 0.634]	0.231	[0.127, 0.333]
Source visibility	H A–V B	0.589	[0.518, 0.661]	0.158	[0.072, 0.244]
Source visibility	H B–V A	0.768	[0.705, 0.830]	0.510	[0.293, 0.631]
Source visibility	H B–V B	0.759	[0.696, 0.821]	0.398	[0.250, 0.536]
Footprint visibility	H A–V A	0.661	[0.589, 0.732]	0.570	[0.406, 0.719]
Footprint visibility	H A–V B	0.652	[0.625, 0.679]	0.570	[0.501, 0.616]
Footprint visibility	H B–V A	0.589	[0.473, 0.705]	0.514	[0.343, 0.677]
Footprint visibility	H B–V B	0.643	[0.554, 0.732]	0.548	[0.367, 0.699]
Protected-object visibility	H A–V A	0.679	[0.571, 0.777]	0.299	[0.153, 0.447]
Protected-object visibility	H A–V B	0.705	[0.616, 0.795]	0.383	[0.137, 0.579]
Protected-object visibility	H B–V A	0.830	[0.759, 0.902]	0.085	[-0.048, 0.249]
Protected-object visibility	H B–V B	0.821	[0.741, 0.902]	0.029	[-0.077, 0.195]
Noncausal role	H A–V A	0.875	[0.812, 0.938]	0.330	[0.107, 0.581]
Noncausal role	H A–V B	0.893	[0.830, 0.955]	0.434	[0.139, 0.789]
Noncausal role	H B–V A	0.875	[0.812, 0.938]	0.101	[-0.071, 0.341]
Noncausal role	H B–V B	0.866	[0.795, 0.938]	0.096	[-0.070, 0.325]
Receiver preservation	H A–V A	0.848	[0.799, 0.897]	0.352	[0.222, 0.475]
Receiver preservation	H A–V B	0.871	[0.830, 0.911]	0.355	[0.292, 0.426]
Receiver preservation	H B–V A	0.848	[0.790, 0.911]	0.119	[-0.024, 0.282]
Receiver preservation	H B–V B	0.871	[0.812, 0.929]	0.240	[0.142, 0.383]
Video quality	H A–V A	0.955	[0.933, 0.978]	0.144	[-0.025, 0.287]
Video quality	H A–V B	0.946	[0.924, 0.969]	0.116	[-0.030, 0.242]
Video quality	H B–V A	0.915	[0.897, 0.933]	0.208	[-0.023, 0.379]
Video quality	H B–V B	0.906	[0.884, 0.929]	0.181	[-0.036, 0.335]

F.3 Evaluator Sensitivity of Method Contrasts

Table 27: SRCD minus Matched on the identical human-study subset. Each evaluator applies its own Original eligibility and usability gates. Mean rows average atomic scores before applying gates.

Evaluator	ΔCES	95% CI	ΔSU	95% CI
Human A	0.054	[-0.018, 0.125]	0.000	[-0.089, 0.089]
Human B	0.089	[-0.036, 0.214]	-0.018	[-0.089, 0.054]
Human mean	-0.009	[-0.071, 0.054]	-0.009	[-0.080, 0.062]
VLM A	0.125	[0.054, 0.196]	0.036	[-0.018, 0.089]
VLM B	0.054	[-0.036, 0.125]	0.000	[0.000, 0.000]
VLM mean	0.071	[0.045, 0.098]	0.018	[-0.009, 0.045]

Table 27 applies each evaluator's own Original gate to the same 14 causal and 14 specificity semantic cases. The human-mean CES contrast is -0.0089, compared with $+0.0714$ under VLM mean scores. Their difference is -0.0804, with interval $[-0.1429, -0.0179]$. For SU, the corresponding human-minus-VLM contrast difference is -0.0268, with interval $[-0.0893, 0.0268]$. The human result is imprecise and does not establish equivalence or a universal negative effect; it also does not reproduce the positive VLM contrast on this subset.

Capability judgments explain much of this sensitivity. Among 14 Wan causal cases, human mean scores make seven Originals eligible, VLM mean scores make nine eligible, and only five are eligible under both. For CogVideoX, both evaluators select seven eligible Originals, again with five in common. Thus equal eligible counts can still select different events.

Keeping the denominator fixed at 14, the descriptive CES contrast-gap decomposition for Wan is -0.0179 from different output scores on jointly eligible cases, $+0.0089$ from human-only eligible cases, and -0.0714 from VLM-only eligible cases. Their sum is the observed -0.0804 gap. This post-result diagnostic does not replace either primary evaluator.

Aggregation order also matters. Applying the gates separately to each human's labels gives positive CES contrasts of $+0.0536$ and $+0.0893$, both with intervals spanning zero. Their arithmetic mean is $+0.0714$. However, the specified procedure first averages the atomic human scores and then gates them, yielding -0.0089. Since eligibility contains a threshold, these operations do not commute. We retain the specified order rather than choosing the more favorable contrast after observing the results.

G Scope and Limitations

SRCD is evaluated as seven mechanism-specific adapters on Wan 2.1 T2V 1.3B. The CogVideoX block supplies external baselines, not a second-backbone replication of SRCD. Each training run uses one seed, and each semantic case uses one generation seed shared across compatible methods. Our results therefore do not estimate training-seed sensitivity or cross-backbone robustness. The task covers locked-camera events within a supplied ontology; held-out names test transfer within that scope.

The source-free targets are independently generated distributional proxies, not unique counterfactuals aligned to factual clips. Event roles are supplied during data construction. Consequently, the behavioral controls test dependence on source identity and placement, not whether the model learns an internal causal representation.

Evaluation depends on both base-model capability and judge reliability. Original gating limits erasure credit when the generator cannot realize an event, while shared-capability requirements restrict external comparisons. VLM judgments may miss small objects, ambiguous footprints, or temporal details in the frame composite. The completed human study finds substantial source and footprint disagreement and an evaluator-sensitive CES contrast. Only two causal and two specificity cases per mechanism enter that study, limiting both precision and subtype coverage. Its human raters view full videos while the VLM views frame panels, so disagreement includes the effect of input presentation. It does not establish interchangeable evaluators. Graded improvement should also be distinguished from complete removal. Selected qualitative examples illustrate individual outcomes rather than the distribution of performance.

H Additional Qualitative Examples

Figures 5 and 6 show six additional cases in the layout of Figure 3. These outcome-aware selections were checked over all 49 frames. Frame indices match across methods within each panel; the horizontal rule separates the Wan and CogVideoX blocks.

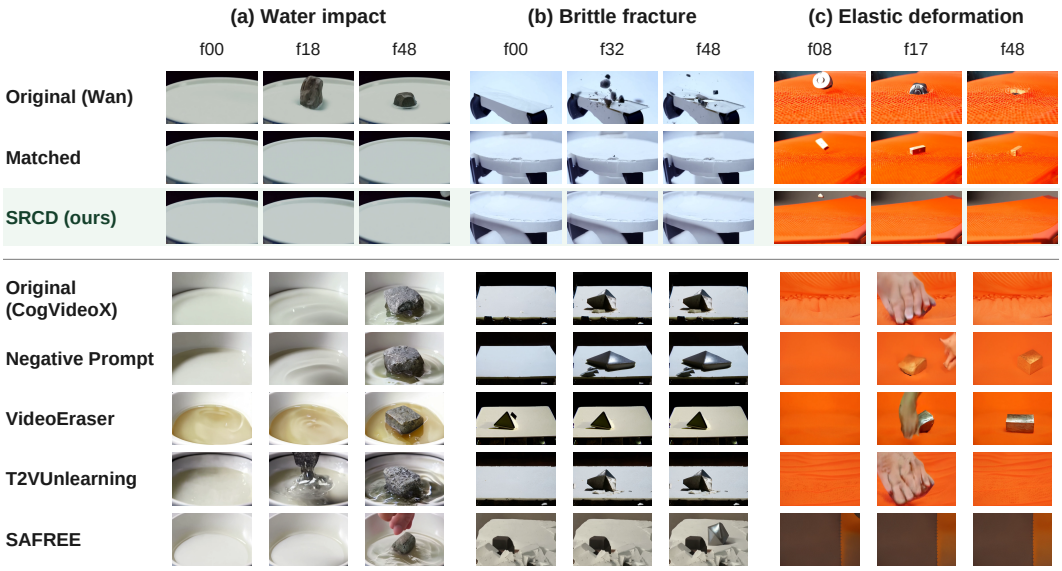

Figure 5: **Water impact, brittle fracture, and elastic deformation.** SRCD reduces the prominent events while retaining the receivers. Matched also suppresses much of (a) and (b). Small SRCD sources remain at frame 48 in (a) and frame 8 in (c); receiver geometry varies.

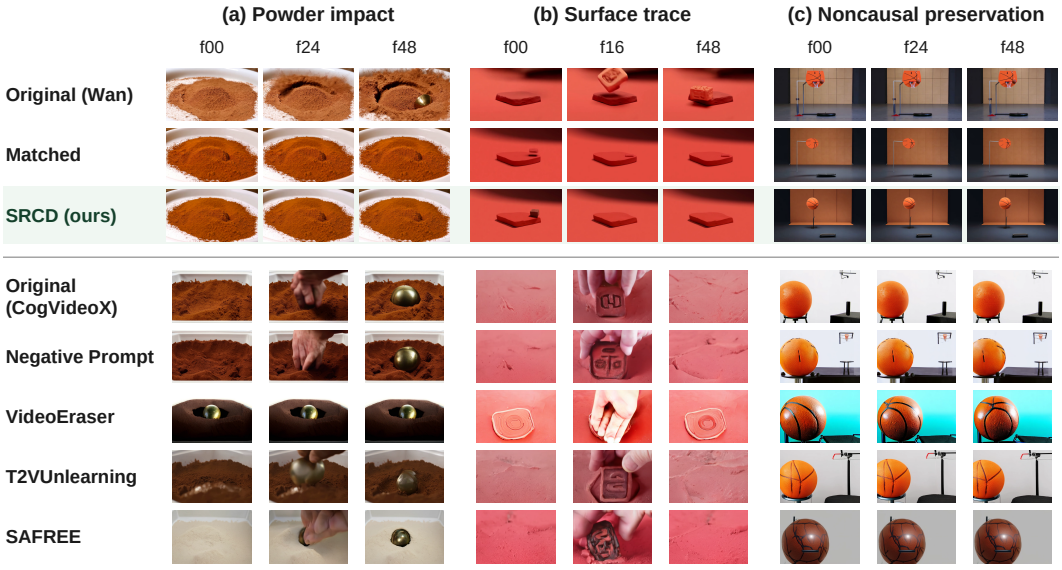

Figure 6: **Surface changes and noncausal preservation.** Both trained arms reduce the developing cavity in (a), while retaining initial surface structure. In (b), SRCD retains an early small stamp; the Original footprint is partly occluded. In (c), the basketball remains separate from the receiver, including under SRCD.