CARVE: Counterfactual Video Editing for Auditing and Hardening Video Detectors

Wei Zhou * **Yiming Chen** **Jinwei Xu** **Yang Zhou** **Quan Gan** **Li Yang**
School of Automation, Nanjing University of Science and Technology, Nanjing, China
{weizhou,yimingchen,jinweixu,yangzhou624,quangan,yangli945}@njust.edu.cn

Abstract

Video detectors are usually evaluated on observational splits, where target events co-occur with environmental and capture factors. A high AUC under this setup can reflect either the event itself or its surrounding conditions, and standard reporting cannot tell them apart. We address this with CARVE, a counterfactual video editing protocol. For each source clip, CARVE generates a matched quartet (V^0, V^E, V^A, V^{AE}) that holds camera pose, layout, and surrounding traffic fixed while independently intervening on environment and event. A two-phase generator first produces a layout-conditioned reference image, then performs reference-guided video editing; a three-layer reference / objective / VLM-panel protocol filters the output. Each quartet supports thresholded and continuous diagnostics for false-positive purity, event faithfulness, environment-induced drift, and held-out factor compositions. Thresholded scores are paired with continuous ones so that a conservative detector cannot suppress scores everywhere and look robust. The same audit drives training: CGAA-IC samples weak factors weighted by their measured brittleness and adds quartet-level purity, consistency, and faithfulness losses; the inference graph is unchanged. On CCTV accident detection, the audit shows that aggregate clean AUC hides where detectors fail: a VideoMAE detector with 91.4% AUC on TAD scores only 72.3% CPS on the night slice of CARVE-Q. CGAA-IC raises purity from 84.2% to 89.9% and cross-dataset average AUC from 80.3% to 84.5%.

1 Introduction

Video detectors are usually evaluated on observational splits, where the target event co-occurs with the environmental and capture factors that surround it. Roadside CCTV accident detection [26, 20, 28, 4, 30, 27] is a sharp instance: night, rain, wet pavement, and compression artifacts cluster on accident clips, so a detector that simply reacts to dark, reflective, low-quality video can score well on standard splits without responding to the collision itself. What the model responds to—rather than its AUC—is therefore the right robustness question, and observational data cannot answer it because the relevant factors are entangled with the label by construction. Photometric augmentation does not resolve this either, because it perturbs pixels while leaving the underlying scene factors entangled, and ego-view robustness work [5, 29] addresses a different sensing regime in which camera motion is intrinsic.

We address this gap by intervening on those factors with controllable generators. For each real source clip we construct a matched quartet (V^0, V^E, V^A, V^{AE}) that holds camera pose, road layout, and unrelated trajectories fixed while editing environment, event, or both. Generation runs in two phases: Nano Banana 2 [10] produces a reference image for the requested environment under the source layout, and Seedance 2.0 [21] performs reference-guided video editing under a structured event script.

*Corresponding author: weizhou@njust.edu.cn.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

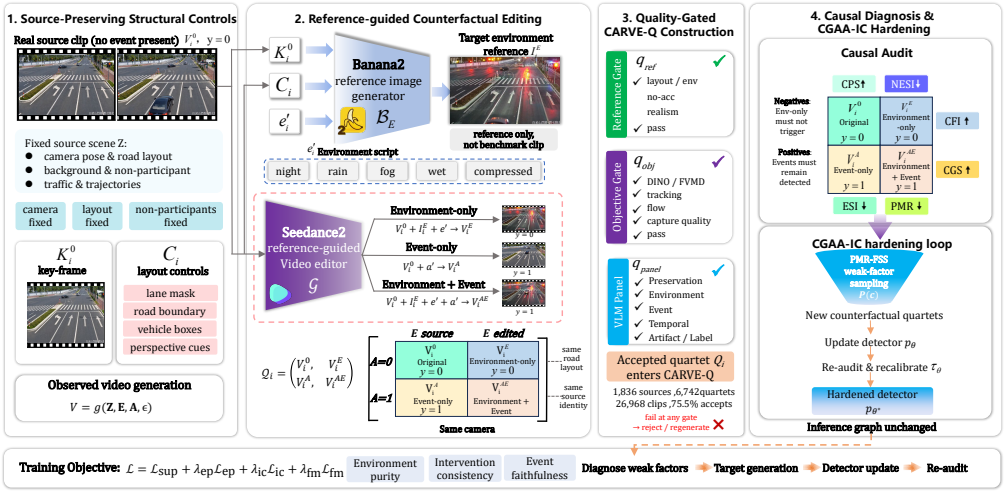

Figure 1: Overview of CARVE. From a real CCTV source clip, Nano Banana 2 produces a reference image for the requested environment under the original layout, and Seedance 2.0 generates the environment edit, the event edit, and their combination while keeping camera geometry fixed. The matched quartets serve two roles: a causal audit (CARVE-Q) and a closed-loop hardening procedure (CGAA-IC).

Three gates—a reference-image check, objective preservation tests, and a panel of vision–language judges [9, 14, 18]—filter the output. The quartets play two roles: as a *causal audit* they expose where a detector breaks, and as a *training signal* they let us repair it via weak-factor resampling and quartet-level losses (Fig. 1). We refer to the protocol as CARVE and to the CCTV audit suite it constructs as CARVE-Q; CGAA-IC is the closed-loop hardening procedure (weak-factor sampler plus quartet losses) that consumes the audit and is not part of benchmark construction.

The audit reveals brittleness that clean AUC hides: a VideoMAE backbone with 91.4% AUC on TAD drops to 72.3% purity (CPS) on night edits and 76.8% on fog, and the detector with the highest purity attains it largely through a weak response to inserted events rather than through event-specific evidence. CGAA-IC, which uses the same audit to drive a closed-loop training procedure, raises CPS to 89.9%, in-domain average AUC from 89.7% to 92.0%, and cross-dataset average from 80.3% to 84.5%, all without changing the inference graph. Our contributions are: **(i)** a counterfactual editing protocol that approximates do-style interventions on environment and event in real video through audited edits while preserving source-scene geometry; **(ii)** a five-metric causal audit (CPS, NESI, ESI, CFI, CGS) that pairs thresholded with continuous summaries to expose score-suppression artifacts; **(iii)** CGAA-IC, a closed-loop hardening procedure with weak-factor resampling and quartet-level losses, leaving the deployed inference graph untouched; and **(iv)** CARVE-Q, a CCTV instantiation with $1{,}836$ source clips and $6{,}742$ accepted quartets that exposes brittleness invisible on standard splits. The code for the full protocol (generation orchestration, three-layer gates, panel prompts and rubric, metric suite, and CGAA-IC training loop) is publicly released at `https://github.com/WeiZhou96/CARVE`.

2 Related work

CCTV traffic accident datasets. Public surveillance benchmarks now span clip-level (CADP [20], TAD [28]), gaze-supervised (TADS [3]), and spatio-temporally annotated (SO-TAD [4], FAD [30], STTAD [27]) corpora; very recent benchmarks such as ACCIDENT [17] extend evaluation to IID/OOD and zero-shot settings, and TU-DAT [11] expands road-anomaly coverage. These resources improve scale, annotation richness, and task diversity, but the environments they contain are still those that happened to co-occur with the recorded events. CARVE-Q supplies matched edits that hold source-scene geometry fixed while isolating nuisance sensitivity. Ego-view benchmarks [5, 29] are valuable but study a different sensing regime; we focus narrowly on fixed roadside infrastructure.

Video accident detection methods. Standard baselines include reconstruction anomaly models (MemAE [8]), 3D CNNs [23, 7, 6], video transformers [2, 22, 25], and traffic-specific models such

as TTHF [12]. FAD/STTAD methods focus on event-level localization with richer supervision; our protocol uses binary clip labels (accident present or not) under controlled nuisance shifts, i.e., edits of label-irrelevant environment factors (Section 3.2). Across these works, accuracy is reported on observational splits, so how scores change when a single factor is edited is never measured. CARVE metrics complement existing traffic-monitoring evaluation such as the FAD/STTAD localization protocols, which are computed on observational footage, whereas CPS, PMR, NESI, and ESI measure how a clip-level decision changes when one factor is edited.

Causal robustness and generative editing for diagnosis. Our formulation draws on invariant prediction [16, 1] and causal representation learning [19], but turns these abstract notions into pixel-level interventions on real video via the CARVE protocol. Recent controllable image and video generators [10, 21] support text-, image-, and video-conditioned edits; using them as a robustness-audit instrument requires careful preservation of source-scene geometry and a quality protocol that audits both visual fidelity and edit semantics, which CARVE supplies.

3 Method

3.1 Setup and counterfactual quartets

We use a structural view of a video clip,

$$V = g(\mathbf{Z}, \mathbf{E}, \mathbf{A}, \epsilon), \tag{1}$$

where $\mathbf{Z}$ collects fixed source-scene variables (camera pose $\mathbf{Z}_{\mathrm{cam}}$, layout $\mathbf{Z}_{\mathrm{road}}$, identities and trajectories of agents outside the scripted event $\mathbf{O}_{\mathrm{nonpart}}, \mathbf{T}_{\mathrm{nonpart}}$), $\mathbf{E}$ collects environment variables (illumination, weather, road surface, capture quality), $\mathbf{A} = (\mathbf{I}_{\mathrm{part}}, \Delta\mathbf{T}_{\mathrm{part}}, \kappa_{\mathrm{type}}, \rho_{\mathrm{impact}}, t_0, \sigma_{\mathrm{sev}})$ encodes event participants, the localized trajectory perturbation that creates the event, type, region, onset, and severity, and ϵ captures residual nuisance variability. A detector $p_\theta : \mathcal{V} \to [0, 1]$ assigns each video clip an event confidence, with parameters θ trained against the binary label y. Fig. 2 shows the assumed dependence structure: only $\mathbf{A}$ is label-relevant, but in observational data the recording process produces statistical correlations between $\mathbf{E}$ and the label that an unconstrained detector can exploit. $\mathbf{A}$ parameterizes one scripted event per clip, and the detector outputs a single clip-level probability; clips containing several events fall outside the current label definition. The notation specifies the intended invariance target; the editor is a stochastic tool for approximate factor changes, and generated samples are retained only after explicit quality checks.

For each source clip we construct a matched quartet $\mathcal{Q}_i = (V_i^0, V_i^E, V_i^A, V_i^{AE})$ with target labels $y(V_i^0) = y(V_i^E) = 0$ and $y(V_i^A) = y(V_i^{AE}) = 1$. An ideal detector satisfies $p_\theta(V_i^0) \approx p_\theta(V_i^E) \approx 0$ and $p_\theta(V_i^A) \approx p_\theta(V_i^{AE}) \approx 1$. The quartet provides strictly more information than a clean/event pair: it supports auditing of false positives ($V^0 \to V^E$), stability checks on positive cases ($V^A \to V^{AE}$), and compositional generalization analysis under matched source scenes. The protocol assumes a static camera; extending it to ego-view data [5, 29] would require explicit treatment of camera motion and a different objective video gate, which we leave as future work.

3.2 Editing protocol and quality control

Two-phase generation. From V_i^0 we extract source key frames K_i^0 and a small set of layout cues C_i: lane masks and boxes around vehicles outside the scripted collision from off-the-shelf segmentation and detection models, road-boundary overlays from an edge-based estimator, and coarse perspective cues from a vanishing-point estimate, all with manual spot checks. Nano Banana 2 generates a reference image $I_i^E = \mathcal{B}_E(K_i^0, C_i, e_i'; \zeta_i^E)$ for the target environment script e_i'. The reference image acts as a visual prior under source-layout constraints and is excluded from benchmark clip counts. Seedance 2.0 then performs reference-guided video editing: $V_i^E = \mathcal{G}_E(V_i^0, I_i^E, e_i'; \xi_i^E)$, $V_i^A = \mathcal{G}_A(V_i^0, a_i'; \xi_i^A)$, and $V_i^{AE} = \mathcal{G}_{AE}(V_i^0, I_i^E, e_i', a_i'; \xi_i^{AE})$, where a_i' is the structured event script. V^A preserves the source environment and omits the environment reference; V^{AE} shares I_i^E with V^E to keep the requested environment appearance consistent inside the quartet.

Editable factor space. Environment factors cover illumination (day/dusk/night), weather (clear/rain/fog), road surface (dry/wet), and capture quality (clean/compressed/motion blur). Event

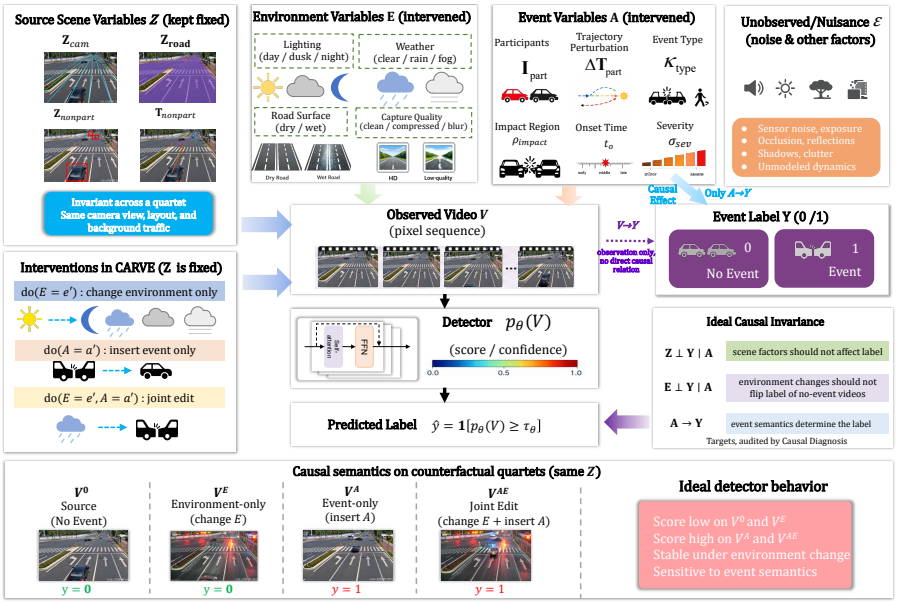

Figure 2: Causal view. Source-scene variables $\mathbf{Z}$, environment variables $\mathbf{E}$, and event variables $\mathbf{A}$ shape the observed clip, but only $\mathbf{A}$ is label-relevant. In observational datasets, spurious statistical correlations between environment and labels can still be exploited by detectors.

factors cover type (rear-end, side-impact, intersection t-bone, pedestrian strike, motorcycle collision, rollover), severity (minor/moderate/severe), and onset (early/middle/late). These 11 label-irrelevant environment levels (the nuisances) were fixed before construction, following adverse conditions recurrent in CCTV accident corpora and restricted to factors compatible with source-scene preservation. Traffic density, viewpoint, and distance are embedded in $\mathbf{Z}$ and remain fixed within each quartet. A feasibility mask $M_{\text{valid}}(e, a) \in \{0, 1\}$ removes physically implausible compositions: clear–wet scenes are retained only when the script explicitly denotes a plausible cause of wet pavement; rain–dry scenes require a short transitional onset of light rain; rollover scripts require sufficient participant scale and visible road space. Held-out compositions for compositional generalization are drawn from this feasible set. For 64-frame clips at 16 fps, event onset levels map to visible impact windows: early = frames 24–28, middle = 29–34, late = 35–40. Restricting onsets to this band leaves at least 1.5 s of pre- and post-impact evidence (onsets near the clip boundary would make the label ambiguous); real positives follow the same constraint.

Three-layer quality control. A candidate enters the suite only after passing three layers. *(i) Reference-image gate* $q_{\text{ref}} = q_{\text{layout}} \wedge q_{\text{env}} \wedge q_{\text{no-acc}} \wedge q_{\text{real}}$ checks that I_i^E respects source road topology and camera perspective, realizes the requested nuisance, contains no spurious crash evidence, and is plausible as a CCTV frame; failing references are discarded before video generation. This gate is necessary because the reference-image backend, while improving controllability, can subtly alter road topology before video generation, and downstream gates cannot always recover from layout drift introduced upstream. *(ii) Objective video gates* require masked DINO similarity [15] ≥ 0.92, FVMD [13] $\leq \tau_m = 85$, displacement outside the target region ≤ 4 pixels, flow-jump rate $\leq \tau_f = 0.12$ (the fraction of adjacent frame transitions whose mean optical-flow magnitude outside the edit changes by more than a validation-calibrated threshold γ_f), and a match between requested and detected capture quality. Apart from the validation-calibrated γ_f and the trained capture-quality probe, thresholds were fixed before construction on a small pilot batch against human accept/reject judgments, without a per-threshold sweep. *(iii) Panel gate.* A panel $\mathcal{M} = \{$Gemini 3.1 Pro, GPT-5.4 Pro, Qwen3-VL$\}$ [9, 14, 18] receives a uniform evidence packet (edit script, target labels, key-frame grids, matched source/generated thumbnails, I_i^E for V^E/V^{AE}, masks, and motion evidence) and returns seven-dimension scores $\mathbf{r}_m \in \{1, \ldots, 5\}^7$ covering preservation, environment correctness, event correctness, temporal continuity, artifact control, CCTV realism, and label reliability. The clip is accepted iff $\min_d \text{median}_m\, r_{m,d} \geq 4$ and $\max_d \text{IQR}_m\, r_{m,d} \leq 1$. We also report a continuous semantic-fidelity score $S_{\text{panel}} = \frac{100}{7} \sum_d (\text{median}_m\, r_{m,d} - 1)/4$ for diagnostic summaries; the

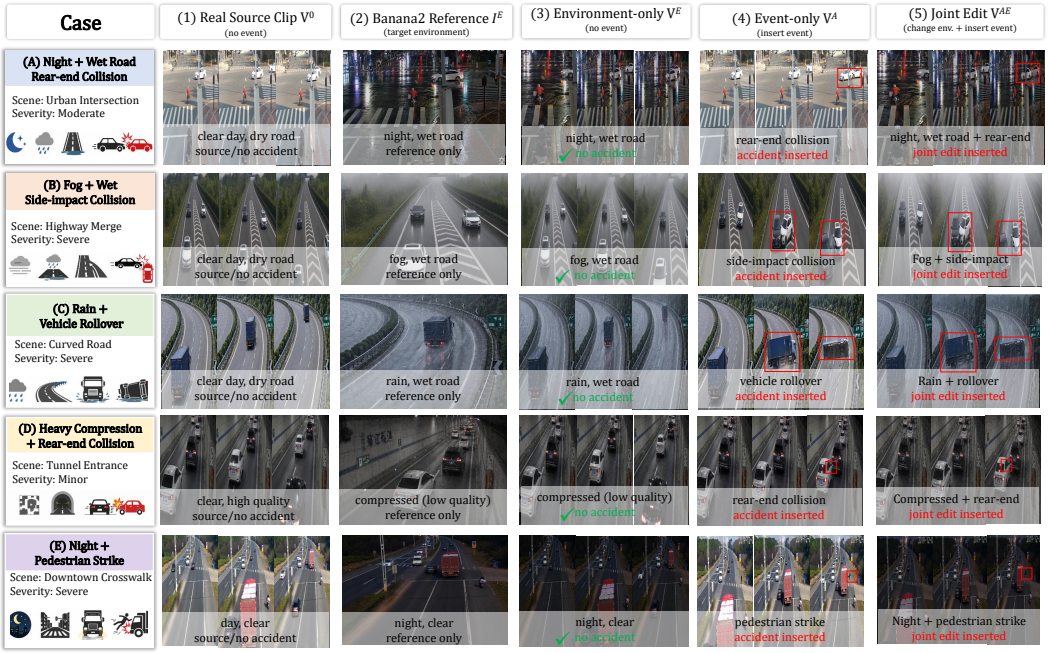

Figure 3: Illustrative counterfactual edits from the generation backend. Nano Banana 2 provides the reference image I^E for the requested environment, and Seedance 2.0 generates the final video edits. The examples show the environment branch, localized accident insertion, and the joint edit while preserving camera geometry, road layout, and surrounding traffic.

binary gate determines acceptance. A quartet is retained only if all three of V^E, V^A, V^{AE} pass both objective and panel gates and I_i^E passes the reference gate. Fig. 3 shows representative outputs that pass these gates.

Freezing and re-calibrating the panel. Judges answer under a fixed rubric and JSON template, and prompts, model identifiers, access dates, raw and parsed scores, and decisions are stored for every candidate, so later API changes cannot silently alter CARVE-Q. For future audits, the open-weights Qwen3-VL serves as a frozen anchor joined by any two contemporary VLMs; a replacement panel is admissible if its agreement with the same stratified 200-item blinded human validity audit (Section 4) is comparable to κ=0.78, with FAR and FRR reported. The human audit, not any particular API, is the reference standard.

3.3 Causal diagnosis metrics

The matched quartets enable five complementary metrics that probe different failure modes. All thresholded quantities use an operating point τ_θ chosen on the source-dataset validation split to maximize F1; threshold selection excludes target test clips and CARVE-Q test quartets. We pair thresholded with continuous summaries so that conservative detectors cannot trade purity for response strength without being detected.

Counterfactual Faithfulness (CFI) and event-side stability (ESI). On the positive side, CFI measures how strongly inserted events move the score, and ESI measures how stable event scores remain when the environment changes. For example, a detector scoring 0.10 on V^0, 0.12 on V^E, and 0.15 on both V^A and V^{AE} has ESI 0 but CFI 0.04, exposing that it barely detects the inserted event:

$$\begin{aligned} \text{CFI} &= \tfrac{1}{2N}\textstyle\sum_{i=1}^{N}\left[p_\theta(V_i^A) - p_\theta(V_i^0) + p_\theta(V_i^{AE}) - p_\theta(V_i^E)\right], \\ \text{ESI} &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\left|p_\theta(V_i^{AE}) - p_\theta(V_i^A)\right|. \end{aligned} \tag{2}$$

ESI must be read jointly with CFI: a weakly responsive detector can trivially achieve low ESI by reacting weakly to both V^A and V^{AE}, which a low CFI exposes.

Causal Purity Score (CPS) and continuous drift. On the negative side, CPS is one minus the false-positive rate on event-free environment edits, and NESI is the mean absolute score change they cause. The detector above has CPS 100% and NESI 0.02; conversely, a detector whose threshold lies above 0.62 and that scores $p_\theta(V^0)=0.10$ but $p_\theta(V^E)=0.62$ keeps CPS at 100% yet has NESI 0.52:

$$\mathrm{CPS} = 1 - \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbb{1}\big[p_\theta(V_i^E) \geq \tau_\theta\big]\,, \qquad \mathrm{NESI} = \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \big|p_\theta(V_i^E) - p_\theta(V_i^0)\big|. \tag{3}$$

NESI is the continuous companion to CPS: a detector that suppresses false alarms but still drifts in score under environment edits has high CPS yet large NESI, and is flagged as fragile on the negative side.

Compositional Generalization Score (CGS). To measure robustness on unseen feasible compositions, we hold out a subset of factor combinations Ω_{unseen} and report the two quantities below (e.g., AUC 0.90 on seen and 0.81 on unseen compositions gives CGS 90 and a gap of 0.09):

$$\mathrm{CGS}_{\mathrm{ratio}} = \frac{\mathrm{AUC}_{\Omega_{\mathrm{unseen}}}}{\mathrm{AUC}_{\Omega_{\mathrm{seen}}}}, \qquad \mathrm{CGS}_{\mathrm{gap}} = \mathrm{AUC}_{\Omega_{\mathrm{seen}}} - \mathrm{AUC}_{\Omega_{\mathrm{unseen}}}. \tag{4}$$

A ratio close to one and a small gap indicate that the detector retains performance on compositions absent from training. In main tables CGS denotes $100 \times \mathrm{CGS}_{\mathrm{ratio}}$; the gap is used internally for weak-factor scoring.

Positive Miss Rate (PMR) and the per-factor brittleness score. For positive event factors we additionally measure $\mathrm{PMR}_c = \frac{1}{2N_c}\sum_{i\in c}\big[\mathbb{1}(p_\theta(V_i^A) < \tau_\theta) + \mathbb{1}(p_\theta(V_i^{AE}) < \tau_\theta)\big]$, which prevents weak-factor sampling from focusing solely on false positives. The five per-factor signals combine into a brittleness score

$$\mathrm{FSS}(c) = \eta_1\widetilde{(1-\mathrm{CPS}_c)} + \eta_2\widetilde{\mathrm{NESI}_c} + \eta_3\widetilde{\mathrm{ESI}_c} + \eta_4\widetilde{\mathrm{PMR}_c} + \eta_5\widetilde{\mathrm{CGS}_{\mathrm{gap},c}}, \tag{5}$$

where $\widetilde{\cdot}$ denotes min–max normalization across the factor levels for which a term is defined, with the weights renormalized over the available terms when a term is undefined for a level, and $(\eta_1,\ldots,\eta_5)=(0.25, 0.15, 0.20, 0.25, 0.15)$ weight false alarms, negative drift, positive drift, misses, and composition transfer; the two thresholded failure modes carry equal and highest weight. The weights were fixed a priori by this rule, never tuned on test data, and not swept. Per-level estimates are stabilized by empirical-Bayes shrinkage before sampling. We refer to the resulting weak-factor sampler as PMR-FSS.

3.4 CGAA-IC: closed-loop hardening

Algorithm 1 summarizes the hardening loop. After each round we recalibrate τ_θ, audit the detector on validation quartets to obtain per-factor scores, and sample new edits with probability $P(c) \propto \mathrm{FSS}(c)^\alpha$, $\alpha = 1.5$. The training objective combines supervised binary cross-entropy on $\mathcal{D}_r \cup \mathcal{D}_s$ with three quartet-level terms that target the failure modes measured above:

$$\mathcal{L}_{\mathrm{ep}} = \tfrac{1}{N}\textstyle\sum_i\big[\mathrm{BCE}(p_\theta(V_i^0), 0) + \mathrm{BCE}(p_\theta(V_i^E), 0)\big], \tag{6}$$

$$\mathcal{L}_{\mathrm{ic}} = \tfrac{1}{N}\textstyle\sum_i\big|p_\theta(V_i^A) - p_\theta(V_i^{AE})\big|, \tag{7}$$

$$\mathcal{L}_{\mathrm{fm}} = \tfrac{1}{N}\textstyle\sum_i\big[\max(0, \delta - p_\theta(V_i^A) + p_\theta(V_i^0)) + \max(0, \delta - p_\theta(V_i^{AE}) + p_\theta(V_i^E))\big], \tag{8}$$

with $\delta = 0.4$ and $\mathcal{L} = \mathcal{L}_{\mathrm{sup}} + \lambda_{\mathrm{ep}}\mathcal{L}_{\mathrm{ep}} + \lambda_{\mathrm{ic}}\mathcal{L}_{\mathrm{ic}} + \lambda_{\mathrm{fm}}\mathcal{L}_{\mathrm{fm}}$, $(\lambda_{\mathrm{ep}}, \lambda_{\mathrm{ic}}, \lambda_{\mathrm{fm}}) = (0.5, 1.0, 0.25)$. $\mathcal{L}_{\mathrm{ep}}$ targets the CPS failure mode, $\mathcal{L}_{\mathrm{ic}}$ targets ESI drift on positives, and $\mathcal{L}_{\mathrm{fm}}$ keeps a margin between event and clean counterparts. $\mathcal{L}_{\mathrm{ic}}$ penalizes only the gap between the two positives; their level is set by the supervised terms. CGAA-IC operates only at training time; the deployed detector keeps the base inference graph.

Why a closed loop is necessary. A static augmentation pool, even one drawn from CARVE-Q, ignores that a detector's brittleness profile shifts across rounds (once night-rain false alarms are suppressed, weaknesses such as heavy compression become binding) and that the initial τ_θ stops maximizing validation F1 as quartet losses reshape the scores. We therefore recalibrate τ_θ and recompute $\{\mathrm{FSS}(c)\}$ every round; empirically, the iterative rediagnosis row in Table 6 contributes the final 0.7 points of cross-dataset AUC and the lowest ESI.

Algorithm 1 CGAA-IC: weak-factor sampling with quartet losses

Require: real training set $\mathcal{D}_r$, source clips $\mathcal{S}$, detector p_θ, rounds R

Pretrain p_θ on $\mathcal{D}_r$
for $r = 1$ to R **do**
    Calibrate τ_θ on the source-dataset validation split
    Audit p_θ on validation quartets; compute $\{\mathrm{FSS}(c)\}$ via Eq. (5)
    Sample target factors with $P(c) \propto \mathrm{FSS}(c)^\alpha$
    Generate references with $\mathcal{B}_E$, gate them; generate quartets with Seedance 2.0; gate them
    Form $\mathcal{D}_s^{(r)}$ and update p_θ on $\mathcal{D}_r \cup \bigcup_{k \le r} \mathcal{D}_s^{(k)}$ with $\mathcal{L}$
end for
return hardened detector p_θ

4 CARVE-Q: instantiation on CCTV accident detection

We build CARVE-Q from CCTV source clips drawn from TAD (127), eligible CADP events with sufficient verified pre-impact context (364), and SurveillanceCrash (1,345), totaling $1{,}836$ source clips and $8{,}931$ candidate tuples. After the three-layer gates, $6{,}742$ quartets remain (acceptance rate 75.5%; per-source rates $76.1/73.7/75.9\%$). Each retained source contributes 1–6 quartets (median 3, mean 3.67), giving $26{,}968$ clip entries and $20{,}226$ generated edits in addition to $1{,}836$ unique real V^0 clips. Every V^0 label is negative; for CADP, only events whose pre-impact window can be verified as accident-free serve as V^0 sources. All partitions are disjoint by source unit ($89/13/25$ for TAD, $255/36/73$ for CADP, $941/135/269$ for SurveillanceCrash, totaling $1{,}285/184/367$), with $4{,}701/657/1{,}384$ accepted quartets; the test split additionally reserves 341 quartets from 18 held-out feasible compositions for CGS.

Generation quality and audit caveats. Filtered V^E edits achieve masked DINO similarity 0.953, FVD [24] 86.2, FVMD 60.9, and $S_{\text{panel}} = 94.6$; V^A and V^{AE} are progressively harder ($S_{\text{panel}} = 92.4$ and 91.7), but all filtered rows are well above an ungated diagnostic batch of V^{AE} outputs ($S_{\text{panel}} = 83.1$, DINO 0.892; Table 2), which shows what the gates remove: the batch averages below the 0.92 DINO gate because failed candidates are typically catastrophic rather than marginal. Of the $2{,}189$ rejected candidates, 287 fail the reference gate, $1{,}346$ the objective gates, and 556 the panel (Appendix A). In a blinded realism study, 18 raters who did not know each clip's provenance judged each of 240 clips (120 real, 120 synthetic) as real or synthetic, with three judgments per clip and 40 clips per rater; 408 of 720 judgments were correct (56.7%; 95% CI 53.2–60.1%, clip-clustered bootstrap with $10{,}000$ resamples), only modestly above chance despite statistical detectability. Accident physical plausibility on 96 collision clips passes all four dimensions for 79.2% of clips ($76/96$), with remaining failures concentrated around small distant impacts, wheel/debris artifacts, and ambiguous post-impact evidence. Two caveats follow. First, even after all three gates, generated clips remain audited counterfactual approximations rather than physical simulations of crashes, so we report a CARVE-Q clean AUC inside the suite separately from real-dataset AUCs in Tables 3 and 4. Second, on a blinded audit sample of 120 panel-accepted and 80 panel-rejected items, the panel agrees with humans on 89.5% (κ=0.78 on this sample); 5.8% of accepted items are judged invalid (FAR) and 17.5% of rejected items valid (FRR), both conditioned on the panel decision. These rates place a soft ceiling on synthetic-distribution cleanliness and bias the accepted set toward clearly correct edits.

5 Experiments

5.1 Setup

We use three CCTV-only datasets under a unified binary protocol at clip level, decoded at 16 fps with 64-frame clips and positive onset constrained to frames 24–40: **TAD** [28] (277 accident, 127 non-accident videos), **CADP-Binary** (constructed from $1{,}120$ eligible CADP events [20], each providing one positive clip centered on impact and one manually verified negative clip ending before impact), and **SurveillanceCrash**, the processed CCTV crash corpus from prior work [31] ($3{,}556$ accident, $4{,}102$ non-accident videos). Cross-dataset experiments train only on the source training split, calibrate the threshold and select models on the source validation split, and evaluate without

Table 1: Causal audit of representative detectors on CARVE-Q. Clean AUC, CPS, CGS in %; ESI, CFI unitless. Bold = best per column. NESI and PMR enter the per-factor brittleness score (Eq. 5); per-level values are in Appendix B.

Method	Clean AUC↑	CPS↑	ESI↓	CGS↑	CFI↑
Reconstruction-based					
MemAE [8]	77.9	**91.8**	**0.121**	64.5	0.43
3D convolutional					
C3D [23]	80.8	77.6	0.236	70.1	0.49
SlowFast [7]	85.8	80.9	0.203	75.2	0.56
X3D-M [6]	84.6	82.2	0.211	74.8	0.54
Video Transformer					
TimeSformer [2]	88.1	83.4	0.183	79.8	0.61
TTHF [12]	89.2	82.6	0.177	80.7	**0.67**
VideoMAE [22]	89.9	84.2	0.171	82.4	0.65
VideoMAE-V2 [25]	**90.8**	85.0	0.164	**83.5**	0.66

Table 2: Generation quality on CARVE-Q (filtered V^E, V^A, V^{AE} vs. an ungated V^{AE} batch). Lower FVD/FVMD, higher DINO/Panel are better; bold = best per column; gray row = ungated diagnostic reference, not a method.

Clip	FVD↓	FVMD↓	DINO↑	Panel↑
Accepted by 3-layer gates				
V^E	**86.2**	**60.9**	**0.953**	**94.6**
V^A	98.7	68.1	0.938	92.4
V^{AE}	103.4	76.2	0.931	91.7
Negative reference (no gating)				
Unfilt. V^{AE}	124.8	95.6	0.892	83.1

VLM gate vs. blinded human gate on an enriched 120/80 accepted/rejected sample: agreement $179/200{=}89.5\%$, $\kappa{=}0.78$; FAR $7/120{=}5.8\%$, FRR $14/80{=}17.5\%$.

target training on the full target test split. We audit MemAE [8], C3D [23], SlowFast [7], X3D-M [6], TimeSformer [2], TTHF [12], VideoMAE [22], and VideoMAE-V2 [25]; for hardening experiments we use VideoMAE-B as the default backbone. Training uses AdamW (lr $3{\times}10^{-4}$, weight decay 0.05, batch 32, cosine decay, 30 base epochs), and CGAA-IC runs for $R{=}3$ rounds; each round adds generated edits equal to roughly 25% of the original real training clips. The random quartet baseline is an internal ablation, not an external method: it uses the same generation backend, the same gates, and the same accepted per-round edit count, but samples factors uniformly from the feasible set instead of by PMR-FSS. Full split sizes, per-round edit counts, and inference throughput appear in the appendix.

5.2 Detectors are brittle under controlled edits, and synthetic edits are well-controlled

Tables 1 and 2 ask whether detectors break under matched edits and whether the edits are trustworthy. Table 1 shows that no single detector dominates. Among the seven supervised detectors, clean AUC tracks the ranking closely (CGS follows it exactly and ESI falls monotonically), yet even VideoMAE-V2, with the highest clean AUC (90.8) and CGS (83.5), reaches only 85.0% CPS, i.e., false alarms on roughly one in seven event-free edits. MemAE breaks the pattern: lowest clean AUC (77.9) but highest CPS (91.8%) and lowest ESI (0.121), with a CFI of just 0.43 showing that much of its purity comes from a weak response to inserted events. TTHF has the strongest CFI (0.67) but lower purity than the transformer backbones. Clean AUC thus orders supervised detectors but hides how many false alarms nuisance edits trigger and where they concentrate. For the base VideoMAE detector, night edits drop CPS to 72.3%, fog to 76.8%, and heavy compression to 79.5%, all well below the overall 84.2%. Table 2 makes poor edits an unlikely main driver of the audit signal (all filtered rows: DINO > 0.93, panel ≥ 91.7), though isolated artifacts remain.

5.3 Hardening with CGAA-IC: in-domain and cross-dataset results

Table 3 compares augmentation strategies on the three real CCTV datasets with VideoMAE-B. RandAugment and MixUp/CutMix change the average by at most 0.2 points, suggesting that unstructured pixel perturbations do not address the CCTV nuisance space. Hard-sample oversampling and IRM [1]-style training give modest gains (90.2, 90.6); random quartet synthetic augmentation is stronger (91.2) because it changes semantic factors while preserving the source layout; full CGAA-IC gives the best average AUC ($89.7 \rightarrow 92.0\%$, $+2.3$ points). Cross-dataset evaluation (Table 4) measures whether the cues transfer beyond the source dataset: CGAA-IC improves every direction and the average rises from 80.3% to 84.5%. Gains are largest on transfers from TAD (T$\rightarrow$S $+8.5$, T$\rightarrow$C $+5.8$), which has the smallest training split among the three sources and therefore benefits most from diagnosed weak-factor edits; gains between CADP-Binary and SurveillanceCrash are smaller ($+3.7$,

Table 3: In-domain AUC at clip level (%) with VideoMAE-B. Bold = best per column; blue row = our final method.

Training strategy	TAD↑	CADP-Binary↑	SurveillanceCrash↑	Avg.↑
No / standard augmentation				
No augmentation	91.4	89.2	88.4	89.7
RandAugment	91.6	89.5	88.1	89.7
MixUp / CutMix	91.5	89.7	88.6	89.9
Sample- and invariance-based				
Hard-sample oversampling	92.0	89.8	88.7	90.2
IRM-style training	92.2	90.4	89.2	90.6
Counterfactual quartet augmentation (ours)				
Random quartet synthetic	93.0	90.8	89.8	91.2
CGAA (no consistency)	**93.9**	91.4	90.1	91.8
CGAA-IC	93.7	**91.8**	**90.6**	**92.0**

Table 4: Cross-dataset AUC (%) for CCTV-only zero-shot transfer with VideoMAE-B; T = TAD, C = CADP-Binary, S = SurveillanceCrash. Target sets are never used for tuning. Bold = best per column; blue row = our final method.

	From TAD		From CADP		From Surv		
Method	T→C	T→S	C→T	C→S	S→T	S→C	Avg.↑
No augmentation	74.3	71.1	85.8	82.7	84.4	83.2	80.3
IRM-style	75.9	73.4	86.3	83.1	84.6	83.7	81.2
Random quartet synthetic	78.8	77.4	87.5	84.6	85.5	84.5	83.1
CGAA-IC	**80.1**	**79.6**	**89.0**	**86.4**	**86.3**	**85.5**	**84.5**

$+2.3$). Quartet-level constraints thus appear to help most when the source training split is narrow. Because every target test set is real footage that never passed through a generator, these gains cannot stem from generator fingerprints alone.

5.4 Ablation, hard slices, and qualitative analysis

Table 6 decomposes components against the same random-quartet baseline. Quartet synthetic data alone improves CPS from 84.2 to 86.5 and the cross-dataset average from 80.3 to 83.1; PMR-FSS focuses the budget on diagnosed weak factors, raising CPS to 87.7 without a cross-dataset gain ($83.1 \rightarrow 82.7$); $\mathcal{L}_{\text{ep}}$ gives the largest CPS gain among the quartet losses ($+1.7$; Xfer 83.6); $\mathcal{L}_{\text{ic}}$ gives the largest single ESI step ($0.138 \rightarrow 0.117$, about a third of the total drop) at a small CPS cost and lifts Xfer to 84.2; $\mathcal{L}_{\text{fm}}$ adds margin separation; iterative rediagnosis gives the best final balance (84.5). Hard slices (Table 5) show where gains concentrate: $+7.8$ on night+rain and $+3.7$ on heavy compression, matching the weaknesses flagged in Section 5.2, and $+6.1$ on low severity; these slices were chosen as the real-data counterparts of the weakest audit-flagged factors. AP and F1 lifts on SurveillanceCrash ($+2.6$, $+2.5$) confirm that gains are visible at the operating threshold, not only in ranking; Appendix A translates the audit metrics into deployment quantities such as per-condition false-alarm budgets.

Fig. 4 illustrates the pattern on one accepted quartet: the baseline drifts on V^E (Grad-CAM following wet-road reflections and headlight streaks) and weakens on V^{AE}, while the hardened detector concentrates attention on the colliding vehicles and impact region across both positives.

Remaining failure modes. The factor space is conservative: camera viewpoint, distance, and traffic density are held fixed in $\mathbf{Z}$ because relaxed-geometry pilots violated source-layout constraints (Appendix A); converting them into editable factors with gates redefined against an explicit geometry model is future work. Small distant impacts and occluded collisions remain hard, and a residual fraction of accepted clips contains local artifacts, matching the 5.8% panel false-acceptance rate and motivating retention of the human audit.

Table 5: Hard-slice AUC (%) on real TAD and SurveillanceCrash test clips with VideoMAE-B; slice definitions in Appendix A. Bold = better of the two.

Slice (Pos/Neg)	Base	CGAA-IC	Δ
Hard nuisance slices			
Night+rain (126/160)	71.6	**79.4**	**+7.8**
Low severity (107/107)	69.1	**75.2**	**+6.1**
Heavy compr. (147/184)	74.3	**78.0**	**+3.7**
SurveillanceCrash overall			
AUC	88.4	**90.6**	+2.2
AP	82.8	**85.4**	+2.6
F1	81.3	**83.8**	+2.5

Table 6: Component ablation with VideoMAE-B. CPS, TAD, Xfer in %; ESI unitless. Bold = best per column; blue row = our final method.

	CARVE-Q audit		Real-data AUC	
Variant	CPS↑	ESI↓	TAD↑	Xfer↑
Base VideoMAE	84.2	0.171	91.4	80.3
+ quartet synth.	86.5	0.154	93.0	83.1
+ PMR-FSS sampler	87.7	0.142	93.1	82.7
+ $\mathcal{L}_{\text{ep}}$	89.4	0.138	92.9	83.6
+ $\mathcal{L}_{\text{ic}}$	89.1	0.117	93.4	84.2
+ $\mathcal{L}_{\text{fm}}$	89.6	0.112	**93.9**	83.8
+ iterative rediagnosis	**89.9**	**0.105**	93.7	**84.5**

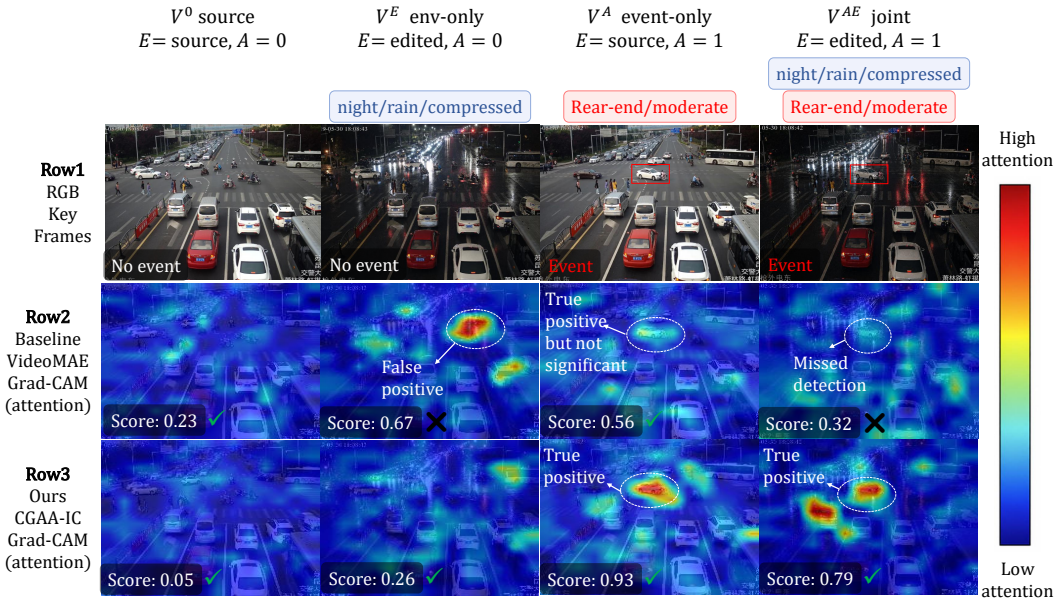

Figure 4: Qualitative Grad-CAM on a matched quartet (`night/rain/compressed`, `rear-end/moderate`). Top: RGB; middle: baseline VideoMAE; bottom: CGAA-IC-hardened. Scores are accident probabilities under τ_θ.

6 Conclusion

We presented CARVE, a counterfactual editing protocol that turns environment-vs.-event entanglement in video detectors into a measurable property and a training signal. Matched quartets drive both the audit on CARVE-Q and the closed-loop hardening CGAA-IC, raising CPS from 84.2% to 89.9% and cross-dataset AUC from 80.3% to 84.5% without changing the inference graph. The audit shows that aggregate clean AUC hides large slice-level purity drops and that purity must be read jointly with event response. The protocol transfers to another video task if (i) the label is caused by a localized event, so an edit leaving the event untouched cannot change it (icy-road detection fails this); (ii) non-event scene content can be held fixed (ego-view without ego-motion modeling fails this); and (iii) preservation and edit semantics are checkable. Relaxing the static-camera constraint and substituting open generation backends are natural next steps toward a general-purpose audit primitive.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Natural Science Foundation of China under Grants 52502402 and 92370203, in part by the Basic Research Program of Jiangsu Province under Grant BK20251462, and in part by the Fundamental Research Funds for the Central Universities under Grant 30926010702.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2020.

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 813–824, 2021.

[3] Yachuang Chai, Jianwu Fang, Haoquan Liang, and Wushouer Silamu. TADS: a novel dataset for road traffic accident detection from a surveillance perspective. *The Journal of Supercomputing*, 80(18):26226–26249, 2024. doi: 10.1007/s11227-024-06429-7.

[4] Xingyuan Chen, Huahu Xu, Mingyang Ruan, Minjie Bian, Qishen Chen, and Yuzhe Huang. SO-TAD: A surveillance-oriented benchmark for traffic accident detection. *Neurocomputing*, 618:129061, 2025. doi: 10.1016/j.neucom.2024.129061.

[5] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, He Wang, and Sen Li. DADA-2000: Can driving accident be predicted by driver attention? analyzed by a benchmark. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 4303–4309, 2019. doi: 10.1109/ITSC.2019.8917218.

[6] Christoph Feichtenhofer. X3D: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020.

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019.

[8] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1705–1714, 2019.

[9] Google Cloud. Gemini 3.1 pro. `https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/3-1-pro`, 2026. Accessed 2026-04-19.

[10] Google DeepMind. Nano banana 2: Combining pro capabilities with lightning-fast speed. `https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/`, 2026. Accessed 2026-04-21.

[11] Pavana Pradeep Kumar and Krishna Kant. TU-DAT: A computer vision dataset on road traffic anomalies. *Sensors*, 25(11):3259, 2025. doi: 10.3390/s25113259.

[12] Rongqin Liang, Yuanman Li, Jiantao Zhou, and Xia Li. Text-driven traffic anomaly detection with temporal high-frequency modeling in driving videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 34:8684–8697, 2024. doi: 10.1109/TCSVT.2024.3390173.

[13] 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. In *ICML Workshop on Controllable Video Generation*, 2024. URL `https://openreview.net/forum?id=tTZ2eAhK9D`.

[14] OpenAI. GPT-5.4 pro model documentation. `https://developers.openai.com/api/docs/models/gpt-5.4-pro`, 2026. Accessed 2026-04-19.

[15] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL `https://openreview.net/forum?id=a68SUt6zFt`.

[16] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

[17] Lukas Picek, Michal Čermák, Marek Hanzl, and Vojtěch Čermák. ACCIDENT: A benchmark dataset for vehicle accident detection from traffic surveillance videos. *arXiv preprint arXiv:2604.09819*, 2026. doi: 10.48550/arXiv.2604.09819.

[18] Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[19] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning. *arXiv preprint arXiv:2102.11107*, 2021.

[20] Ankit Parag Shah, Jean-Baptiste Lamare, Tuan Nguyen-Anh, and Alexander Hauptmann. CADP: A novel dataset for CCTV traffic camera based accident analysis. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–9, 2018. doi: 10.1109/AVSS.2018.8639160.

[21] Team Seedance. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026. Model page: `https://seed.bytedance.com/en/seedance2_0`, accessed 2026-04-19.

[22] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, volume 35, pages 10078–10093, 2022.

[23] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015.

[24] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenges. In *International Conference on Learning Representations Workshop*, 2019.

[25] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE V2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023.

[26] World Health Organization. Road traffic injuries. Fact sheet, 2023. URL `https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries`. Accessed 2026-04-16.

[27] Sixuan Xu, Tianran Zhang, Lei Zhao, Wei Zhou, Stefan Nastic, and Chen Wang. Surveillance-based spatial temporal traffic accident detection: A novel dataset and tailored algorithm. In *2025 IEEE 28th International Conference on Intelligent Transportation Systems (ITSC)*, 2025. doi: 10.1109/ITSC60802.2025.11423029.

[28] Yajun Xu, Huan Hu, Chuwen Huang, Yibing Nan, Yuyao Liu, Kai Wang, Zhaoxiang Liu, and Shiguo Lian. TAD: A large-scale benchmark for traffic accidents detection from video surveillance. *IEEE Access*, 13: 2018–2033, 2025.

[29] Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Yuchen Wang, Ella M. Atkins, and David J. Crandall. DoTA: Unsupervised detection of traffic anomaly in driving videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):444–459, 2023. doi: 10.1109/TPAMI.2022.3150763.

[30] Hongyang Yu, Xinfeng Zhang, Yaowei Wang, Qingming Huang, and Baocai Yin. Fine-grained accident detection: Database and algorithm. *IEEE Transactions on Image Processing*, 33:1059–1069, 2024. doi: 10.1109/TIP.2024.3355812.

[31] Wei Zhou, Longhui Wen, Yunfei Zhan, and Chen Wang. An appearance-motion network for vision-based crash detection: Improving the accuracy in congested traffic. *IEEE Transactions on Intelligent Transportation Systems*, 24(12):13742–13755, 2023. doi: 10.1109/TITS.2023.3297589.

A Additional details

Editing prompt template. We use a strict minimal-change rule: "Preserve camera pose, road geometry, lane count, vehicle identities, and all object trajectories. Treat the provided reference image only as a target environment appearance prior. Keep accident state unchanged. Change only the requested environment factors. Do not add or remove vehicles. Do not alter scene layout or camera motion."

Flow-jump rate. $\mathrm{FJR}(V) = \frac{1}{T-2}\sum_{t=1}^{T-2} \mathbb{1}[\frac{1}{|\Omega_{\mathrm{non}}|}\sum_{x\in\Omega_{\mathrm{non}}} \big|\|F_{t+1}(x)\|_2 - \|F_t(x)\|_2\big| > \gamma_f]$, where F_t denotes optical flow from frame t to $t+1$, Ω_{non} is the region outside the edit, and γ_f is a percentile threshold calibrated on the validation split. The capture-quality probe is a lightweight classifier with three classes (clean, compressed, motion blur) trained on manually labeled CCTV clips and used only for filtering.

Source composition. TAD contributes $127/612/466$ source/candidate/accepted (76.1%); CADP $364/1{,}769/1{,}303$ (73.7%); SurveillanceCrash $1{,}345/6{,}550/4{,}973$ (75.9%). Of the $2{,}189$ rejections among the $8{,}931$ candidates, 287 fail the reference gate, $1{,}346$ the objective gates, and 556 the panel, so the panel accounts for about a quarter of all rejections and 92.4% of the candidates that reach it survive. Applying the 17.5% FRR to the 556 panel rejections gives an estimated 97 wrongly discarded candidates, about 1.1% of all candidates and 1.4% of the $6{,}742$ accepted quartets; the dominant yield cost lies in the objective gates, which are fully reproducible.

Relaxed-geometry pilots. The pilots let the editor move viewpoint and camera distance. Three failure modes dominated: lane topology drifted (lane counts changing mid-clip), non-participant vehicles appeared or disappeared between frames, and the edited background lost perspective consistency with the source geometry, so DINO-similarity and layout-gate pass rates collapsed to an impractical yield. Once viewpoint, distance, or density moves, the preservation set is no longer well defined and the objective gates lose their reference, because masked DINO similarity, displacement, and flow checks are meaningful only relative to a region that ought to stay fixed.

Hard-slice definitions. The slices in Table 5 are subsets of the real TAD and SurveillanceCrash test clips: night+rain ($126/160$ pos/neg), low severity ($107/107$), and heavy compression ($147/184$). Labels come from dataset metadata or from two annotators applying the CARVE-Q factor definitions, with adjudication; a clip may belong to several slices. Because severity is an event attribute, the low-severity AUC uses matched accident-free clips from the same split and a similar capture-quality band as negatives. The slices were chosen as the real-data counterparts of the weakest audit-flagged factors.

CGS held-out compositions. The 18 held-out (environment, event-type, severity) compositions used for CGS are drawn by stratified sampling from the feasibility mask M_{valid}: for each of the three severity levels we hold out two combinations from the easy block (clear/dry day) and four from the hard block (rain, fog, night, or wet road), yielding $6{+}12 = 18$ compositions matched on per-block prevalence. None of the held-out compositions appear in the training or validation quartets used for FSS computation. For detectors trained only on real data (Table 1), no composition is seen during training; CGS then compares the held-out compositions with the rest of the suite, and because 12 of the 18 come from the hard block, it also reflects composition difficulty rather than generalization alone.

Real-dataset splits. TAD uses $194/89$, $28/13$, $55/25$ accident/non-accident clips for train/val/test; CADP-Binary uses $784/112/224$ events with one positive and one negative clip each; SurveillanceCrash uses $2{,}489/2{,}871$, $356/410$, $711/821$ accident/non-accident clips with hash-level deduplication.

Per-round synthetic budget and inference cost. Each round of CGAA-IC adds $72/393/1{,}341$ generated edits for TAD/CADP-Binary/SurveillanceCrash respectively (roughly 25% of the original real training clips). On a Jetson AGX Orin 64GB, VideoMAE-B runs at 8.3 clips/s before hardening and 8.2 clips/s after, well within measurement noise.

Deployment considerations. Per-condition CPS is a per-condition false-alarm budget at the deployed operating point: on a stream of $1{,}000$ event-free candidate clips per day, the base detector's night CPS of 72.3% against 84.2% overall means roughly 277 versus 158 nuisance alarms per day. PMR covers missed detections, and NESI and ESI govern alarm-threshold hysteresis and how far an operator can trust scores near the threshold. Hardening leaves the inference graph and edge throughput unchanged (previous paragraph), so the per-factor audit serves as a condition-specific pre-deployment check for a given camera.

Training compute. A single complete training run (the 30-epoch base detector together with $R=3$ rounds of CGAA-IC) takes approximately 12 hours on $8\times$ NVIDIA RTX 5880 Ada workstation GPUs (48 GB VRAM each, 384 GB aggregate). The Jetson AGX Orin throughput numbers above refer to inference-time deployment hardware, not the training cluster.

Human study composition. The blinded validation study uses 18 raters who participated voluntarily as internal research collaborators: six traffic or vehicle-safety researchers/practitioners, six computer-vision researchers, and six general annotators drawn from the authors' research network. Each clip receives three independent real/synthetic judgments. Raters were not informed of the editing protocol or of any model's decisions, and the 200-sample validity audit (next paragraph) used the same blinding rule. No external crowdsourcing platform was used and no monetary compensation was provided; participation followed standard internal evaluation protocols within the authors' lab and partner groups, and could be withdrawn at any time.

Independent VLM-vs-human gate audit. The 5.8% false-acceptance and 17.5% false-rejection rates reported in Section 4 come from an independent 200-sample blinded re-audit, not from VLM self-assessment. We sample 120 candidates accepted by the VLM panel and 80 candidates that passed the objective gates but were rejected by the VLM panel; human raters then judge each item as semantically valid or invalid for its intended target labels *without seeing the VLM panel's decisions or scores*. The two evaluations agree on $179/200 = 89.5\%$ of items (Cohen's $\kappa = 0.78$), with 7 of 120 accepted items judged invalid (FAR $= 5.8\%$) and 14 of 80 rejected items judged valid (FRR $= 17.5\%$). This audit and the 18-rater realism study in the previous paragraph measure different properties: the realism study asks whether a clip looks real (a binary real/synthetic discrimination, only weakly above chance at 56.7% with 95% CI $[53.2\%, 60.1\%]$ that excludes 50%), while the validity audit asks whether a clip correctly implements its target environment and event labels. The two are not directly comparable.

Released code and withheld media. The public release (`https://github.com/WeiZhou96/CARVE`) contains (i) the CGAA-IC training loop with all quartet losses and the PMR-FSS sampler; (ii) the metric suite (CPS, NESI, ESI, CFI, CGS, PMR, FSS) with audit scripts; (iii) the objective-gate implementations with every configured threshold (masked DINO 0.92, motion gate, 4-pixel displacement, flow-jump gate with calibrated γ_f, capture-quality probe); (iv) the panel protocol: prompts, the seven-dimension rubric, JSON parsing and validation, and the median/IQR acceptance rule; (v) generation-orchestration code with the structured factor scripts, the full two-phase prompt templates, and the feasibility mask; and (vi) the evaluation harness, split-construction tooling, all configuration files, and a synthetic-score test suite that exercises the metric and sampling arithmetic without media. The pipeline is generator-agnostic at the interface level, so a group with licensed access to the source datasets and generator endpoints can re-run construction, audit, and hardening end to end.

Factor configurations, structured editing instructions, masks, gate thresholds, VLM prompts and rubrics, raw and parsed judge scores, and acceptance decisions are recorded for every candidate quartet. The media and the raw traces, which embed source-derived frames and thumbnails, are not redistributed. First, V^0 clips are the source corpora themselves and remain obtainable under those corpora's own terms, whose licenses restrict redistribution of derived footage. Second, the generators re-render rather than copy pixels: in the accepted edits we inspected, faces and other small deformable structures are not reconstructed faithfully at CCTV resolution, but rigid high-contrast structures survive, and because the protocol preserves vehicle appearance and non-participant trajectories, license plates often remain legible. We therefore treat every branch derived from a real source as identifying media: it is held under team-only access, not redistributed, and never used for re-identification. Blurring does not remove this constraint, because each V^E stays scene-matched

to its publicly available source clip, so a viewer who links the pair recovers what the blur removed. Third, synthetic crash footage derived from real scenes carries dual-use risk (next paragraph).

Dual-use risk. We see two concrete misuse channels. The first is fabricated evidence: a V^A or V^{AE} clip is real CCTV footage of a real location with a plausible scripted collision inserted, which lends itself to insurance fraud and staged-evidence uses. The second is misinformation: location-anchored surveillance footage reads as an official recording, and the realism study shows limited separability from real clips (56.7% rater accuracy). The asymmetry in what we release is deliberate: code, prompts, and thresholds are open because their value lies in auditability, and turning them into new forgeries still takes generator access and effort, whereas a ready-made crash clip tied to a real scene is usable as is. Crash media derived from real sources therefore stay unreleased.

Reproducibility and statistical reporting. Detector results in Tables 1–6 are single-seed point estimates from one complete training run per configuration; the realism-study confidence interval in Section 4 uses clip-clustered bootstrap with $10{,}000$ resamples.

Additional limitations. Two further caveats inform interpretation. (i) The VLM panel composition (Gemini 3.1 Pro / GPT-5.4 Pro / Qwen3-VL) reflects models available at construction time; replicability under future API revisions or model deprecation is not guaranteed; the forward recalibration standard in Section 3.2 specifies how a replacement panel is validated against the human audit. (ii) $\mathcal{L}_{\text{ic}}$ enforces score stability across environment edits and can therefore suppress legitimate visibility-driven changes (e.g., a fog-occluded collision being intrinsically harder to detect). In the ablation, adding $\mathcal{L}_{\text{ic}}$ gives the largest single ESI reduction while TAD AUC rises from 92.9 to 93.4 and cross-dataset AUC also improves, which argues against a collapse into uniform overconfidence; the per-edit visibility scores recorded in the audit traces help diagnose this when it occurs.

B Per-factor diagnostics

Table 7 lists diagnostics of the base VideoMAE detector for the most-sampled factor levels on the 657-quartet validation pool (after empirical-Bayes shrinkage) together with the sampling share that PMR-FSS ($P(c) \propto \mathrm{FSS}(c)^{1.5}$) assigns to each level from these diagnostics. These validation-split values differ from the test-split CPS reported in Section 5.2 (72.3 night, 76.8 fog, 79.5 heavy compression). The overall validation NESI is 0.126 and the overall PMR is 0.172. The hard-slice gains in Table 5 fall on the most-sampled factors.

Table 7: Validation diagnostics of base VideoMAE and PMR-FSS sampling shares for the most-sampled levels (%). CPS and NESI are defined for environment levels, PMR for event levels.

Level	CPS	NESI	PMR	Share
Night	73.4	0.201	–	14.7
Fog	77.5	0.176	–	9.8
Compressed	80.1	0.153	–	6.4
Rain	81.9	0.143	–	5.2
Minor severity	–	–	0.271	12.9
Pedestrian strike	–	–	0.235	8.4
Late onset	–	–	0.211	7.2
Motorcycle collision	–	–	0.194	5.4
Remaining 15 levels				30.0

C Notation

Table 8: Main notation.

Symbol	Meaning
$\mathbf{Z}$	Fixed source-scene variables: camera pose, road layout, non-participant identities and trajectories
$\mathbf{E}$	Environment variables: illumination, weather, road surface, capture quality
$\mathbf{A}$	Event variables: participants, trajectory perturbation, type, impact region, onset, severity
V^0, V^E, V^A, V^{AE}	Source clip, environment edit, event edit, joint edit of a quartet
I^E	Reference image for the requested environment
p_θ, τ_θ	Detector and its operating threshold (max-F1 on the source validation split)
CPS, NESI	Purity on V^E and mean absolute score drift $\lvert p_\theta(V^E) - p_\theta(V^0)\rvert$
CFI, ESI	Event response and positive-side drift $\lvert p_\theta(V^{AE}) - p_\theta(V^A)\rvert$
CGS	$100\times$ ratio of AUC on held-out to seen compositions; $\mathrm{CGS}_{\mathrm{gap}}$ is their difference
PMR_c	Positive miss rate for event level c
$\mathrm{FSS}(c)$, η, α	Brittleness score, its weights, and the sampling exponent
$\mathcal{L}_{\mathrm{ep}}, \mathcal{L}_{\mathrm{ic}}, \mathcal{L}_{\mathrm{fm}}, \delta$	Purity, consistency, and faithfulness losses; margin of $\mathcal{L}_{\mathrm{fm}}$
S_{panel}	Continuous panel semantic-fidelity score

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state four contributions (the CARVE counterfactual editing protocol, a five-metric causal audit, the CGAA-IC closed-loop hardening procedure, and the CARVE-Q CCTV instantiation) and the headline numbers (CPS $84.2 \rightarrow 89.9\%$, in-domain average $89.7 \rightarrow 92.0\%$, cross-dataset average $80.3 \rightarrow 84.5\%$); each is established in Sections 3–5 and the corresponding tables.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The *Remaining failure modes* paragraph in Section 5.4 discusses the conservative editable factor space (camera viewpoint, distance, and traffic density held fixed in $\mathbf{Z}$), residual local artifacts in accepted clips, and the difficulty of small distant impacts and occluded collisions; the *Generation quality and audit caveats* paragraph in Section 4 further notes that generated clips are audited counterfactual approximations rather than physical simulations of crashes and quantifies the soft ceiling imposed by the VLM panel's 5.8% false-acceptance and 17.5% false-rejection rates.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present formal theorems or proofs; the structural-equation notation in Section 3.1 is descriptive and is used to specify the intended invariance target, not as a basis for theoretical claims.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: Sections 3.2–3.4 specify the editing protocol, the three quality gates and their numerical thresholds, the five metric definitions with full equations, the FSS weights, and the algorithm; Section 5.1 specifies datasets, splits, the unified clip protocol, optimizer, hyperparameters, the synthetic budget per round, and the audited backbones; Appendix A records the editing prompt template, the flow-jump-rate definition, per-source acceptance and rejection statistics, real-dataset splits, hard-slice definitions, and per-round edit counts. The released code (Item 5) implements every stage, and a group with licensed access to the source datasets and generator endpoints can re-run construction, audit, and hardening end to end.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code is publicly released at `https://github.com/WeiZhou96/CARVE`: the CGAA-IC training loop with all quartet losses and the PMR-FSS sampler, the metric suite, the three-layer gates with every configured threshold, the panel prompts, rubric, parsing, and acceptance rule, generation-orchestration code with the prompt templates and feasibility mask, split-construction tooling, all configuration files, and a synthetic-score test suite (Appendix A). Media derived from real sources and the raw audit traces that embed them are not redistributed, because source-dataset licenses restrict redistribution of derived footage, the derived clips remain identifying even after blurring, and synthetic crash footage of real scenes carries dual-use risk (Appendix A, *Released code and withheld media* and *Dual-use risk*).

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5.1 lists the optimizer (AdamW), learning rate (3×10^{-4}), weight decay (0.05), batch size (32), epochs (30 base + R=3 rounds), per-dataset synthetic-edit budget, and the threshold-calibration protocol; Section 3 lists the loss weights $(\lambda_{\text{ep}}, \lambda_{\text{ic}}, \lambda_{\text{fm}}) = (0.5, 1.0, 0.25)$, $\delta = 0.4$, $\alpha = 1.5$, and the FSS weights; Appendix A records the full real-dataset split sizes and the per-round edit counts for TAD/CADP/SurveillanceCrash.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: AUC, CPS, ESI, CGS, CFI, and AP/F1 in Tables 1–6 are reported as point estimates from a single backbone configuration, which is consistent with the convention in the comparison literature (MemAE, C3D, SlowFast, X3D-M, TimeSformer, TTHF, VideoMAE, VideoMAE-V2). The human validation study in Section 4 reports a 95% confidence interval via clip-clustered bootstrap with $10{,}000$ resamples, which is the place where statistical variability most directly affects a quantitative claim.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix A specifies that a single complete training run (30-epoch base detector together with R=3 rounds of CGAA-IC) takes approximately 12 hours on $8\times$ NVIDIA RTX 5880 Ada workstation GPUs (48 GB VRAM each, 384 GB aggregate), and reports inference throughput on a Jetson AGX Orin 64GB (8.3 vs. 8.2 clips/s before and after hardening). The per-round synthetic-edit budget and inference cost paragraph documents deployment-time numbers separately from the training cluster.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?

   Answer: [Yes]

   Justification: We use only existing publicly available CCTV accident corpora under their licenses, work with privacy-filtered subsets where available, and refrain from redistributing raw or generated media that could be re-identified or sensationalized. The paper follows the principles in the NeurIPS Code of Ethics; the residual ethical risks (dual-use of synthetic accident footage through fabricated evidence and misinformation, and identifying content such as legible license plates in derived clips) are discussed in Appendix A and motivate the non-release decisions documented in Items 5 and 11.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 1 motivates the work via road-safety reliability, and the *Deployment considerations* paragraph in Appendix A relates the audit metrics to false-alarm budgets and operator workflows. On the negative side, the *Dual-use risk* paragraph in Appendix A names two misuse channels: fabricated evidence (real CCTV of a real location with an inserted collision, usable for insurance fraud or staged evidence) and misinformation (location-anchored footage that reads as an official recording, with only 56.7% rater accuracy in separating synthetic from real clips). The *Generation quality and audit caveats* paragraph in Section 4 discusses the gap between audited counterfactual approximations and real crash physics.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse?

    Answer: [Yes]

    Justification: CARVE-Q is retained internally and is not released as a freely downloadable corpus of generated crash videos; SurveillanceCrash is used only in its privacy-filtered redistributable form; CADP is used at the event level via verified pre-impact windows under its original license; TAD is used under its original license. No pretrained model checkpoints, raw audit traces, or media derived from real sources are redistributed; derived clips are held under team-only access and never used for re-identification. Code, prompts, and thresholds are released because their value lies in auditability; no media are released. These choices are documented in Item 5 and Appendix A (*Released code and withheld media*, *Dual-use risk*).

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets (TAD, CADP, SurveillanceCrash) and pretrained backbones (MemAE, C3D, SlowFast, X3D, TimeSformer, VideoMAE, VideoMAE-V2, TTHF) are cited at first use; the generation models (Nano Banana 2, Seedance 2.0) and the VLM judge panel (Gemini 3.1 Pro, GPT-5.4 Pro, Qwen3-VL) are cited with provider documentation and access dates. We do not redistribute any of these assets and use them only within the terms of their original licenses.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

Justification: The new released asset is the code, documented in Appendix A and in the repository. CARVE-Q itself, the raw audit traces, and media derived from real sources are not released, for the reasons given in Item 5; their construction protocol is documented in Sections 3.2 and 4 and Appendix A.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: We did not use any crowdsourcing platform. The blinded validation study described in Section 4 involves 18 raters who are internal research collaborators (six traffic or vehicle-safety researchers/practitioners, six computer-vision researchers, and six general annotators drawn from the authors' research network); they participated voluntarily as part of standard internal evaluation protocols within the authors' lab and partner groups, with no monetary compensation, and were free to decline or withdraw at any time. Because no crowdsourcing was conducted, the NeurIPS minimum-wage condition for crowdsourced labor does not apply. The composition of the rater pool, the task structure (240 balanced clips, three judgments each), and the bootstrap protocol are documented in Section 4 and Appendix A.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: The validation study is non-interventional: raters view short, non-graphic CCTV clips (length-limited, with explicit pre-screening for graphic post-impact content) and provide real/synthetic and physical-plausibility judgments. No personally identifying information about raters was collected, and no behavioral or biological data were recorded. Formal IRB approval was not sought; raters were briefed about the task, participation was voluntary, and the data they generated consist only of categorical judgments on already-public CCTV footage. We acknowledge that practices around low-risk internal validation studies vary across institutions and that this is one reason we report Item 14 and the present item as [No] rather than overstate the procedural status.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

    Answer: [Yes]

    Justification: A panel of vision–language models (Gemini 3.1 Pro, GPT-5.4 Pro, Qwen3-VL) is used as a quality-control gate on candidate edits; their role, prompt structure, response template, acceptance rule, identifiers, and access dates are documented in Section 3.2 and Section 5.1, and the agreement against blinded human judgments (89.5%, κ=0.78) is reported in Section 4. The judges score candidate edits only; the accident detectors evaluated in this paper never receive judge scores during training or inference.