EVENTWORTH: RISK-CONTROLLED EVENT–FRAME PREDICTION WITH DELAYED VISUAL ANCHORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Conventional cameras often struggle under fast motion or challenging illumination, whereas event cameras capture fine-grained temporal changes that can complement frame observations for temporally causal video prediction. However, event quality varies across sensing conditions, and corrupted or unreliable events can degrade prediction rather than improve it. Existing event–frame methods mainly focus on exploiting events, but rarely determine whether they should be used for each prediction relative to a deployable frame-only fallback. This leaves negative-transfer risk largely uncontrolled. To address this problem, we propose EventWorth, a selector-agnostic framework for risk-controlled event use. EventWorth learns baseline-relative gain and harm scores from paired supervision provided by delayed future anchors. Validation data propose and freeze a selective policy, which is independently certified before deployment. Event assistance is enabled only after certification; otherwise, the system falls back to the frame-only branch. Under the stated round-level i.i.d. certificate assumption, EventWorth provides finite-sample control over the probability of deploying a selector whose selected harm risk exceeds a prescribed target. Experiments show that the analytic EventWorth policy is certified in all five seeds while retaining 50.44% development-test coverage, 1.30% harm, and 0.0220 system gain. These results show that selective event use can retain useful event information while controlling deployment risk.

1 INTRODUCTION

Event cameras report asynchronous brightness changes instead of sampling dense images at a fixed rate. Their high temporal resolution and dynamic range make them useful when conventional frames suffer from fast motion, sparse sampling, or difficult illumination (Gallego et al., 2022). Recent event-based video methods exploit this complementarity through recurrent reconstruction, event–frame synthesis and warping, nonlinear flow, cross-modal motion estimation, and event-reference synthesis (Zou et al., 2021; Weng et al., 2021; Tulyakov et al., 2021; 2022; He et al., 2022; Zhang & Yu, 2022; Kim et al., 2023; Sun et al., 2023; Liu et al., 2024b). Pretrained video diffusion priors further expand the available visual prior (Chen et al., 2025). These methods recover motion information that is missing from sparsely sampled frames.

The benefit of events is nevertheless condition dependent. Event streams can be uninformative in static regions and unreliable under background activity, hot pixels, threshold mismatch, event dropout, low-light latency, or temporal misalignment. Real-world denoising data and reconstruction benchmarks expose strong failures under such conditions (Duan, 2024; Ercan et al., 2023; Gao et al., 2025). Existing methods correct specific defects, adapt event models at test time, or weight regional event features according to reliability (Yang et al., 2024a; Kim et al., 2022; Cho et al., 2024; Liang et al., 2024). These advances improve average robustness, but they do not determine whether the current event stream improves a particular prediction relative to a deployable frame-only fallback.

This question is related to dynamic modality selection, but temporally causal event–frame prediction introduces a distinct supervision structure. General methods quantify modality-subset utility during training or choose among available modalities at inference time (He et al., 2024; Du et al., 2026). In the event–frame setting considered here, the value of events is unavailable when the deployment decision must be made. The next visual anchor arrives only after the prediction suffix has

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.

been generated, but it can retrospectively evaluate both the frame-only and event-assisted branches. Rejecting every event stream is trivially conservative, whereas enabling events unconditionally may violate a prescribed negative-transfer budget. This requires a selective event-use policy whose risk can be certified without using the certification data for tuning.

We propose EventWorth, a framework for independently certified selective event use. Two frozen temporally causal predictors provide a frame-only fallback and an event-assisted candidate. Future anchors yield paired proxy targets for training baseline-relative gain and harm scoring models. The core object certified by EventWorth is a *frozen selector*, rather than a particular gating architecture. A sequence-disjoint validation split proposes one candidate for each predeclared selector. Our default utility-aware instantiation jointly thresholds gain and harm scores, while a matched risk-only selector removes the gain gate and serves as a controlled comparison. Each proposed selector is frozen before the certificate split is opened for its formal policy-risk test. A single predeclared deployment selector may use the full certification failure budget. When multiple predeclared selectors are tested simultaneously, we apply family-wise correction. If certification fails, EventWorth disables event use and falls back to the frame-only branch. The certification result is therefore selector agnostic: joint gain–risk gating is a utility-aware instantiation, not a necessary condition for risk control.

Our contributions are summarized as follows:

- We propose EventWorth, a selector-agnostic framework for risk-controlled event use in temporally causal video prediction. Delayed visual anchors provide paired baseline-relative gain and harm supervision for deciding when event assistance should be used relative to a frame-only fallback.
- We develop a four-way sequence-disjoint protocol that separates score training, selector proposal, independent certification, and evaluation. The resulting policy is frozen before certification and falls back to the frame-only branch when certification fails.
- We establish exact finite-sample control of selected harm risk for validation-frozen selectors, together with a baseline-relative excess-loss guarantee, zero-harm sample complexity, and a family-wise extension for multiple predeclared selectors.
- Extensive experiments verify the effectiveness of EventWorth in retaining useful event assistance while controlling harmful deployment. The analytic policy passes all five marginal certificates with 50.44% development-test coverage, 1.30% harm, and 0.0220 system gain, supporting risk-aware event–frame prediction under unreliable event streams.

2 Related Work

2.1 Event-assisted video formation

Event-based video formation includes event-only reconstruction and event–frame fusion. Recurrent and Transformer models reconstruct dense videos from events (Zou et al., 2021; Weng et al., 2021), while event–frame systems use intensity anchors to recover absolute appearance. Time Lens and Time Lens++ combine synthesis with event-derived warping and nonlinear flow (Tulyakov et al., 2021; 2022); TimeReplayer reduces its dependence on high-frame-rate supervision through cycle consistency (He et al., 2022); and joint deblurring–interpolation methods connect event integration to latent sharp frames (Zhang & Yu, 2022; Sun et al., 2023). More recent architectures introduce cross-modal motion fields, event-reference synthesis, and pretrained video priors (Kim et al., 2023; Liu et al., 2024b; Chen et al., 2025). Causal reconstruction from past visual observations and subsequent events has also been studied for continuous video decompression (Wang et al., 2025). These methods primarily ask how to exploit events. EventWorth instead assumes that competent temporally causal predictors are available and asks when the event-assisted predictor should be deployed.

2.2 Reliability and selective event use

Real event streams vary with illumination, motion, sensor parameters, bandwidth, and synchronization (Gallego et al., 2022). EVREAL documents large performance variation across reconstruction settings (Ercan et al., 2023); LED isolates real background-activity noise (Duan, 2024); and recent

reconstruction work highlights the weak observability of static content (Gao et al., 2025). Task-specific methods correct latency, adapt representations under test-time shift, or use regional reliability for fusion (Yang et al., 2024a; Kim et al., 2022; Cho et al., 2024; Liang et al., 2024). EventWorth addresses a complementary deployment problem. It leaves both prediction branches fixed and learns paired baseline-relative gain and harm signals for selective event use. These signals support either joint gain–risk or risk-only selectors. In both cases, final risk control comes from the independent policy-risk certificate rather than from the score values themselves.

2.3 Modality selection and risk certification

Beyond event vision, modality-selection methods identify informative subsets, choose recovered modalities at inference time, or adapt fusion according to sample-dependent uncertainty (He et al., 2024; Du et al., 2026; Gao et al., 2024; Yang et al., 2024b; Zhang et al., 2023). Related multimodal studies analyze optimization imbalance, missing inputs, and dynamic inference paths (Wang et al., 2020; Peng et al., 2022; Ma et al., 2022; Han et al., 2022; Maheshwari et al., 2024; Xue & Marculescu, 2023). Their main objectives are task accuracy, robustness, or computational efficiency, rather than a certificate for baseline-relative harm on selected predictions.

Selective prediction trades coverage against error (Zaoui et al., 2020; Kalai & Kanade, 2021; Pugnana & Ruggieri, 2023). Holdout risk control and conformal risk control provide finite-sample guarantees for fixed predictors or monotone policy families (Bates et al., 2021; Angelopoulos et al., 2024), while exact binomial confidence limits provide nonasymptotic coverage for Bernoulli risks (Clopper & Pearson, 1934). Online risk control, confidence sequences, and delayed online learning address a different sequential setting in which policies or calibrators continue to update during deployment (Feldman et al., 2023; Howard et al., 2020; 2021; Karampatziakis et al., 2021; Howson et al., 2023). EventWorth instead freezes each candidate selector before certification. It combines delayed paired supervision, validation-only selector proposal, independent selected-risk certification, and fail-closed event deployment in a baseline-relative formulation. The resulting certificate applies to joint gain–risk, risk-only, or other predeclared frozen selectors, with multiplicity control when several selectors are tested simultaneously. The gain gate seeks utility rather than creating the risk-control guarantee.

3 Method

Figure 1 summarizes EventWorth. Future anchors provide paired baseline-relative supervision, validation freezes a selector, and an independent split certifies its selected harm risk. Joint gain–risk gating is the default utility-aware instance, while matched risk-only serves as a control; certification failure triggers the frame-only fallback.

3.1 Baseline-relative event utility

EventWorth controls two fixed temporally causal predictors. Here "causal" refers only to the information timeline and does not invoke causal inference. The branch f_0 provides the frame-only fallback, which serves as the deployment baseline, while f_1 is the event-assisted candidate. Round k is delimited by visual anchors at t_k and t_{k+1}. Before a prescribed decision time $s_k \in (t_k, t_{k+1})$, the system observes the available visual history and event prefix. We denote the causal information by $\mathcal{G}_k$ and compute a health representation:

$$X_k = \phi_k = H_\psi(\mathcal{G}_k) \in \mathbb{R}^d. \tag{1}$$

The representation uses only information available at the decision time, including event statistics, recent visual history, observability, and cross-modal consistency. It estimates the relative gain and harm of event assistance. The two prediction branches themselves remain fixed.

Exactly one branch is selected for the prediction suffix. For $u \in [0, 1]$, define:

$$\tau_k(u) = s_k + u(t_{k+1} - s_k), \qquad y_k(u) = y(\tau_k(u)). \tag{2}$$

A representative causal realization is:

$$\begin{aligned} \widehat{y}_{0,k}(u) &= f_0(I_{t_{k-1}}, I_{t_k}; \tau_k(u)), \\ \widehat{y}_{1,k}(u) &= f_1(I_{t_{k-1}}, I_{t_k}, E_{[t_k, \tau_k(u))}; \tau_k(u)). \end{aligned} \tag{3}$$

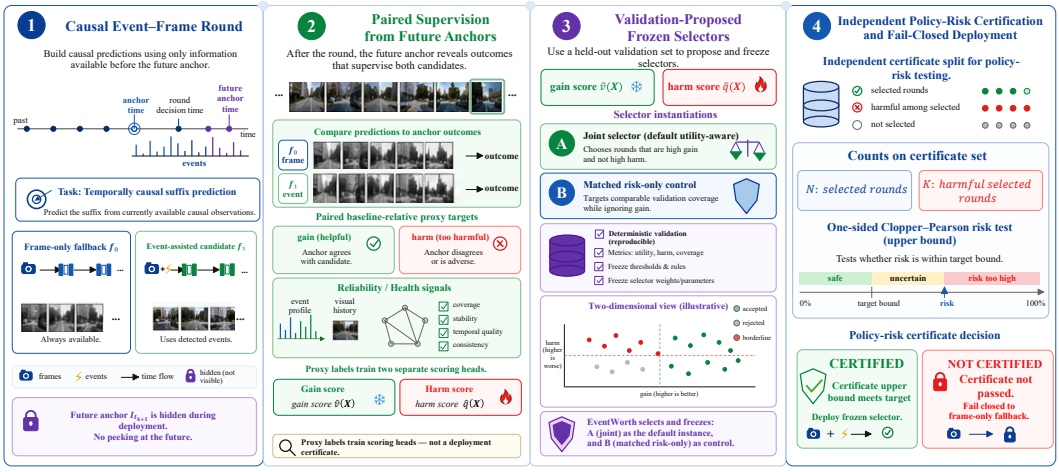

Figure 1: Overview of EventWorth. Future anchors provide paired supervision for training gain and harm scores, validation data propose a frozen selector, and an independent certificate split controls the selected harm risk. Joint gain–risk gating is the default utility-aware instantiation, while matched risk-only selection serves as a predeclared control. A failed certificate triggers fail-closed deployment to the frame-only fallback.

Neither branch observes $I_{t_{k+1}}$ before producing the suffix.

Let $\ell(\widehat{y}, y) \in [0, 1]$ be a normalized task loss. We define:

$$L_{j,k} = \int_0^1 \ell(\widehat{y}_{j,k}(u), y_k(u))\, du, \quad j \in \{0, 1\}, \tag{4}$$

$$r_k = L_{1,k} - L_{0,k}, \qquad g_k = -r_k, \qquad h_k = \mathbf{1}\{r_k > \delta_h\}, \tag{5}$$

where $\delta_h \in [0, 1]$ is the tolerated loss increase used to define a meaningful negative-transfer event. Positive g_k means that event use improves over the fallback. The population objective is to retain selected gain while constraining:

$$R_h(\pi) := \Pr(H = 1 \mid \pi(X) = 1) \le \alpha, \tag{6}$$

where π is any frozen event-use policy and $\alpha \in (0, 1)$ is the selected-harm tolerance. For any policy with positive selection probability, we also define:

$$G_{\text{sel}}(\pi) := \mathbb{E}[g \mid \pi(X) = 1], \tag{7}$$

$$G_{\text{sys}}(\pi) := \mathbb{E}[\pi(X)g] = \Pr(\pi(X) = 1)\, G_{\text{sel}}(\pi). \tag{8}$$

G_{sel} measures the average gain conditional on selection, whereas G_{sys} also accounts for how often the selector is activated. A policy can therefore achieve high selected gain but make only a small system-level contribution when its coverage is low. Both quantities are used for proposal and empirical evaluation. The formal certificate in Sec. 3.4 controls selected harm and applies to both joint gain–risk and risk-only selectors.

3.2 Paired Supervision and Four-Way Split

When $I_{t_{k+1}}$ arrives, it evaluates endpoint predictions generated without access to that anchor:

$$\widetilde{L}_{j,k} := \ell(\widehat{y}_{j,k}(1^-), I_{t_{k+1}}), \qquad j \in \{0, 1\}. \tag{9}$$

The paired proxy targets are:

$$\widetilde{r}_k = \widetilde{L}_{1,k} - \widetilde{L}_{0,k}, \qquad \widetilde{g}_k = -\widetilde{r}_k, \qquad \widetilde{h}_k = \mathbf{1}\{\widetilde{r}_k > \delta_h\}. \tag{10}$$

Assumption 1 (Paired replayability)**.** *Conditional on the logged causal inputs, both branch outputs and the future anchor are invariant to the deployment decision. Thus, both proxy losses are well defined regardless of which branch would have been deployed.*

Assumption 1 makes the delayed anchor paired full-information supervision rather than bandit feedback. We use four sequence-disjoint splits:

$$\mathcal{D}_{\text{train}}, \qquad \mathcal{D}_{\text{val}}, \qquad \mathcal{D}_{\text{cert}}, \qquad \mathcal{D}_{\text{test}}. \tag{11}$$

Their roles are fixed before any policy-risk test.

- $\mathcal{D}_{\text{train}}$ trains a gain score $\widehat{v}(X)$ from $\widetilde{g}$ and a harm score $\widehat{q}(X)$ from $\widetilde{h}$.
- $\mathcal{D}_{\text{val}}$ uses true suffix gain g and harm h to propose one frozen candidate for each predeclared selector.
- $\mathcal{D}_{\text{cert}}$ supplies the formal policy-risk test for each validation-frozen candidate.
- $\mathcal{D}_{\text{test}}$ reports development-test performance of the locked system; because this supplied split was inspected during development, it is not treated as fresh confirmatory evidence.

The prediction backbones, health encoder, score models, and selector thresholds are frozen before certification. No statistic from $\mathcal{D}_{\text{cert}}$ may influence training or validation proposal; proxy-fidelity summaries computed there are diagnostic only. The matched risk-only coverage target is determined from $\mathcal{D}_{\text{val}}$, or from a predeclared fallback, and never uses certificate or development-test outcomes.

A failed certificate is evaluated as reject-all, yielding zero event-use coverage, gain, and additional harm in system aggregation. Score outputs are point predictions; the certificate requires neither pointwise calibration nor the proxy-bias corrections analyzed in Appendix A.6.

3.3 Frozen selector instantiations and deterministic proposal

We consider two predeclared selector instantiations. The default utility-aware selector uses both scores:

$$\pi_{\tau}^{\text{joint}}(x) = \mathbf{1}\{\widehat{q}(x) \le \tau_q\}\mathbf{1}\{\widehat{v}(x) \ge \tau_v\}, \qquad \tau = (\tau_q, \tau_v), \tag{12}$$

whereas the matched risk-only control uses:

$$\pi_{\tau_q}^{\text{risk}}(x) = \mathbf{1}\{\widehat{q}(x) \le \tau_q\}. \tag{13}$$

Risk-only is the joint family with an inactive gain gate; the certificate below is valid for either frozen selector. All thresholds are proposed deterministically from $\mathcal{D}_{\text{val}}$. Joint search uses prespecified support, coverage, risk, and positive-utility criteria with fixed tie-breaking. Matched risk-only uses a predeclared validation-coverage target. The recorded analytic and learned proposal rules are detailed in Appendix B.3.

3.4 Independent policy-risk certification

The following certificate is selector agnostic. Fix any one predeclared validation-frozen selector and denote it generically by $\pi_{\tau^\star}$. Let the certificate split contain m labeled rounds $\{(X_i, H_i)\}_{i=1}^m$, where H_i is the true suffix harm label in Eq. (5). Define:

$$I_i = \pi_{\tau^\star}(X_i), \qquad N = \sum_{i=1}^m I_i, \qquad K = \sum_{i=1}^m I_i H_i. \tag{14}$$

For $0 \le K < N$, the one-sided Clopper–Pearson upper confidence limit is:

$$U_{\text{CP}}(K, N; \eta) = F^{-1}_{\text{Beta}(K+1, N-K)}(1-\eta), \tag{15}$$

where $\eta \in (0,1)$ is the failure budget assigned to this selector (Clopper & Pearson, 1934). We set $U_{\text{CP}} = 1$ when $N = 0$ or $K = N$. The policy-risk certificate decision is:

$$C = \mathbf{1}\{N \ge N_{\min}\}\mathbf{1}\{U_{\text{CP}}(K, N; \eta) \le \alpha\}, \tag{16}$$

and the final deployed action is:

$$a(x) = C\,\pi_{\tau^\star}(x). \tag{17}$$

The certificate depends only on the number of selected rounds N and the number of harmful selected rounds K. The magnitudes of the gain and harm scores do not enter the certificate once the selector is frozen.

Thus, a failed or underpowered policy-risk certificate disables event use. The risk target, minimum selected support, and certificate failure budgets are protocol parameters and are reported with the experiments. When M predeclared selectors are certified simultaneously on the same certificate split, we allocate the failure budget across the family; equal Bonferroni allocation assigns η/M to each selector. Appendix A.5 states the general family-wise guarantee.

Theorem 1 (Independent selected-risk policy certification)**.** *Assume that the frozen policy $\pi_{\tau^\star}$, including its score models and thresholds, is measurable with respect to $\sigma(\mathcal{D}_{\mathrm{train}}, \mathcal{D}_{\mathrm{val}})$ only; no statistic from $\mathcal{D}_{\mathrm{cert}}$ influences this policy. Assume further that $\mathcal{D}_{\mathrm{cert}}$ is independent of the training and validation data, and that the certificate rounds (X_i, H_i) are independent and identically distributed with the target deployment population. For a policy with positive selection probability, let:*

$$R_h(\pi_{\tau^\star}) = \Pr(H = 1 \mid \pi_{\tau^\star}(X) = 1)\,. \tag{18}$$

Then:

$$\Pr(N > 0 \text{ and } R_h(\pi_{\tau^\star}) > U_{\mathrm{CP}}(K, N; \eta)) \le \eta. \tag{19}$$

Consequently:

$$\Pr(C = 1 \text{ and } R_h(\pi_{\tau^\star}) > \alpha) \le \eta. \tag{20}$$

The guarantee is exact under the stated round-level i.i.d. certificate assumption. Sequence-disjoint partitioning prevents cross-split leakage, but rounds from the same video may still be dependent. Theorem 1 therefore controls round-weighted aggregate harm among selected rounds. It is not a pointwise safety statement, and the guarantee does not depend on the presence of a gain gate.

Corollary 1 (Fail-closed deployment)**.** *Under the conditions of Theorem 1, with probability at least $1-\eta$, either $C = 0$ and the event-assisted branch is never deployed, or $C = 1$ and its selected harm risk is at most α.*

Corollary 2 (Selected baseline-relative excess loss)**.** *Suppose $r = L_1 - L_0 \le 1$ and $H = \mathbf{1}\{r > \delta_h\}$. Under the conditions of Theorem 1,*

$$\Pr(C = 1 \text{ and } \mathbb{E}[r \mid \pi_{\tau^\star}(X) = 1] > \delta_h + (1 - \delta_h)\alpha) \le \eta. \tag{21}$$

Corollary 2 converts the binary harm certificate into a continuous fallback-relative excess-loss guarantee: the probability of certifying a selector whose expected selected excess loss exceeds $\delta_h + (1 - \delta_h)\alpha$ is at most η.

Gain is used for selector proposal and is evaluated empirically on the development-test split. The formal certificate controls selected harm. After the selector and certification rule are locked, development-test performance is summarized by empirical coverage together with the sample analogues of G_{sel} and G_{sys} in Eqs. (7)–(8).

Predeclared selector-level comparison. For the learned study, seeds and the joint–risk-only comparison criteria are fixed in advance. The complete decision rule is given in Appendix A.10.

4 Experiments

4.1 Experimental protocol

We use the Real Low-light Event Dataset (RLED) (Liu et al., 2024a), comprising 100 official training sequences and 12 official test sequences. The 112 sequences yield 1,680 rounds and are split sequence-wise into 31/21/48/12 train/validation/certificate/development-test sequences (465/315/720/180 rounds). Because the supplied test split was inspected during development, we report it as development-test rather than fresh confirmatory evidence.

We evaluate five analytic runs and five lightweight learned-predictor runs. All experiments use MAE on $[0, 1]$-normalized intensities with harm margin $\delta_h = 0.005$, risk target $\alpha = 0.05$, and $N_{\min} = 128$. A marginal certificate assigns $\eta = 0.05$ to one deployment selector. The predeclared joint–risk-only comparison uses 0.025 per selector, while the broader analytic ablation in Table 3 uses the archived eight-slot allocation of 0.00625 per slot. Reported deployment metrics apply fail-closed rejection. Full preprocessing, architectures, optimization, proposal rules, seeds, and audit details are given in Appendix B.

4.2 CERTIFIED RISK–UTILITY

Table 1 compares empirical utility with certificate outcomes. Always-on event use attains the largest unconstrained system gain of 0.0301, with an observed harm rate of 4.44%. Its mean certificate upper bound is 9.97%, and none of the five runs is certified. EventWorth joint retains 50.44% coverage, 1.30% observed harm, and $G_{\mathrm{sys}} = 0.0220 \pm 0.0019$, while passing all five marginal single-policy certificates.

The selective baselines exhibit different risk–utility tradeoffs. Reliability-based selection reduces observed harm to 0.21%, but its system gain falls to 0.0132 and four of five runs are certified. Gain-only reaches $G_{\mathrm{sys}} = 0.0224$, close to EventWorth joint, but with 2.40% observed harm and four certified runs. Matched risk-only achieves 0.98% harm and $G_{\mathrm{sys}} = 0.0185$, again with four certified runs. Compared with gain-only, EventWorth joint retains nearly the same system gain while reducing observed harm and passing one additional marginal certificate.

Table 1: Analytic risk–utility results. Mean±SD over five seeds. Development-test metrics describe proposed policies; U_{CP} uses $\eta = 0.05$.

	Development test				Certificate	
Selector	Cov.	Harm	G_{sel} ($\times 10^{-3}$)	G_{sys} ($\times 10^{-3}$)	U_{CP}	Pass
Frame-only	0	—	—	0	—	—
Always-on	100.00 ± 0.00	4.44 ± 0.00	30.1 ± 0.0	30.1 ± 0.0	9.97 ± 2.27	0/5
Event density	53.33 ± 4.76	3.14 ± 0.27	35.8 ± 0.4	19.1 ± 1.6	9.50 ± 1.33	0/5
Reliability	45.33 ± 4.21	0.21 ± 0.48	29.0 ± 1.4	13.2 ± 1.8	3.18 ± 1.31	4/5
Gain-only	53.67 ± 5.09	2.40 ± 1.93	42.2 ± 6.4	22.4 ± 2.1	4.04 ± 2.10	4/5
Matched risk-only	46.00 ± 6.05	0.98 ± 0.57	40.4 ± 5.8	18.5 ± 3.1	3.39 ± 2.27	4/5
EventWorth joint	50.44 ± 4.05	1.30 ± 0.37	44.0 ± 5.4	22.0 ± 1.9	2.83 ± 0.98	5/5

4.3 ABLATIONS OF SUPERVISION, GATING, AND CERTIFICATION

Paired supervision. Table 2 changes only the training target while holding predictors, causal features, score models, optimization, and proposal protocol fixed. Shuffling the paired labels reduces gain correlation to near zero and harm ranking to chance, while reliability-derived supervision recovers only part of the task-relative signal. In contrast, future-anchor supervision reaches $\rho_g = 0.72$, harm AUROC $= 0.873$, and $G_{\mathrm{sys}} = 0.0220$, substantially narrowing the gap to the non-causal suffix-supervised oracle. Full control definitions are given in Appendix B.4.

Relative to reliability-derived supervision, the paired future-anchor target raises gain correlation from 0.45 to 0.72 and harm AUROC from 0.74 to 0.873, while increasing system gain from 0.0128 to 0.0220. This indicates that the delayed anchor provides task-relative information that is not captured by a generic event-reliability target alone.

Selector gates and certification. Gain-only and risk-only remove one gate at a time while reusing the same frozen scores. Table 3 instead holds the proposed selector fixed and changes only whether the independent certificate is enforced.

For the joint selector, family-wise certification reduces mean coverage from 50.44% to 40.89% and system gain from 0.0220 to 0.0174 because one of the five proposals is rejected. Gain-only retains comparable unconstrained utility but higher observed harm, whereas matched risk-only is more conservative and sacrifices utility. The unmatched risk-only selector reaches substantially higher coverage and therefore represents a different operating point. At comparable coverage, matched risk-only yields 38.22% deployed coverage and $G_{\mathrm{sys}} = 0.0153$, compared with 40.89% and 0.0174 for joint gating. Joint gating is therefore used as the utility-aware default, while the certificate itself remains selector agnostic.

4.4 CERTIFICATION STABILITY AND PREDICTOR-SCOPE ANALYSIS

Table 4 reports certificate outcomes for two predictor settings with their recorded frozen proposal protocols. The analytic setting yields proposals in 5/5 seeds and certifies all five, whereas the

Table 2: Controlled supervision ablation. Only training targets change; all predictors, features, score models, and proposal protocols are fixed. ρ_g measures gain-score correlation and AUROC measures harm ranking; the remaining columns report development-test proposal metrics.

Training supervision	$\rho_g \uparrow$	Harm AUROC↑	Cov.↑	Harm↓	$G_{\text{sys}} \uparrow$
Shuffled paired labels	0.01	0.50	31.2	4.6	0.0035
Reliability-derived labels	0.45	0.74	43.8	2.2	0.0128
Paired future-anchor proxy	**0.72**	**0.873**	**50.44**	**1.30**	**0.0220**
True suffix labels (oracle)	0.82	0.93	52.5	0.8	0.0250

Table 3: Gate and certification ablations. Five-seed means under the eight-slot audit. Proposed metrics precede the certificate mask; deployed metrics apply fail-closed filtering. SDs are in Appendix C.

	Proposed policy		Certificate	Deployed system	
Selector	Cov.	G_{sys}	Pass	Cov.	G_{sys}
Neither gate (always-on)	100.00	0.0301	0/5	0.00	0.0000
Gain only	53.67	0.0224	3/5	32.56	0.0140
Risk only (unmatched)	83.33	0.0287	4/5	67.44	0.0228
Risk only (matched)	46.00	0.0185	4/5	38.22	0.0153
Joint gain–risk	50.44	0.0220	4/5	40.89	0.0174

lightweight learned setting yields proposals in 3/5 seeds and certifies one. Because the proposal protocols differ, these results characterize the scope of the certification pipeline rather than isolate an architecture effect.

Figure 2 resolves these outcomes at seed level. Under the stricter simultaneous audit, analytic joint acceptance decreases from 5/5 to 4/5. In the learned setting, joint and matched risk-only pass 1/5 and 2/5 seeds, respectively. Neither satisfies the predeclared selector-level evidence rule; full per-seed records are provided in Appendix C.

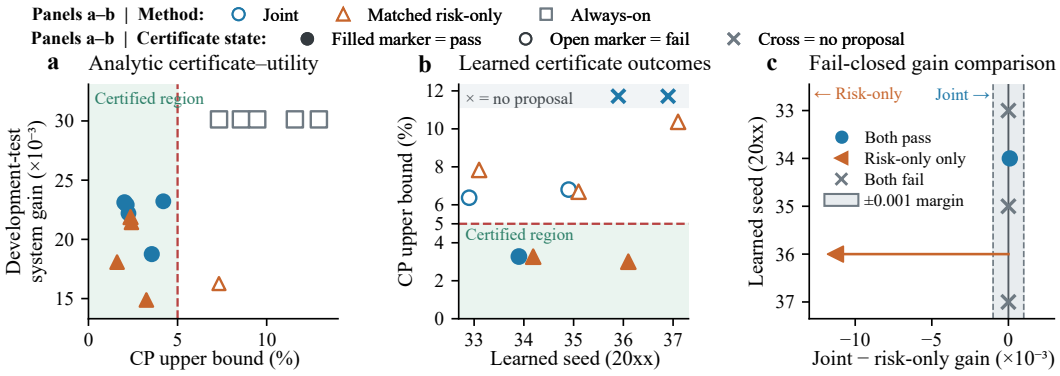

Figure 2: Risk evidence and selector scope. **a**, Analytic certificate bounds versus development-test system gain at $\eta = 0.05$. **b**, Learned certificate outcomes at $\eta = 0.025$. **c**, Fail-closed joint-minus-risk-only gain relative to the ± 0.001 predeclared margin. Filled/open markers denote pass/fail and crosses denote missing proposals.

4.5 Delayed supervision and qualitative evidence

On the validation split, endpoint-proxy and suffix gains have Spearman correlations of 0.9608 ± 0.0090 and 0.9620 ± 0.0079 for the analytic and learned settings, respectively. The corresponding causal harm scores achieve AUROCs of 0.8726 ± 0.0306 and 0.8843 ± 0.0250. The close proxy–suffix ordering indicates that the delayed anchor provides a stable training signal, while the causal scores retain useful harm discrimination using only information available at the decision time.

Table 4: Predictor-scope analysis. Five-seed proposal/certificate counts and fail-closed deployment metrics. The analytic and learned settings use different frozen validation-proposal protocols (Appendix B.3). Analytic certificate counts use the marginal single-selector budget ($\eta = 0.05$), whereas learned counts use the predeclared two-selector budget ($\eta = 0.025$ per selector).

Predictor pair	Proposal	Cert. pass	Deployed Cov.	Deployed G_{sys}
Analytic predictor	5/5	5/5	50.44	0.0220
Lightweight residual CNN	3/5	1/5	7.56	0.00295

Figure 3 illustrates the resulting decisions. The first example accepts an event-assisted prediction that improves the suffix loss, whereas the second rejects a prediction for which event use is harmful. These examples visualize selective deployment rather than pointwise correctness; certified policies may still make individual selection errors, as shown in Appendix F.

We further evaluate the frozen analytic selectors under 120 nonzero event perturbation settings. Observed harm remains at most 5% in 99 settings, and system gain remains positive in 114 settings. Polarity inversion is the dominant failure mode: at the highest severity, mean selected harm rises to 55.7% and system gain falls to -0.001. In contrast, dropout, timestamp jitter, background activity, hot pixels, and threshold thinning retain positive mean system gain at their highest tested severity. Full perturbation results are reported in Appendix E.

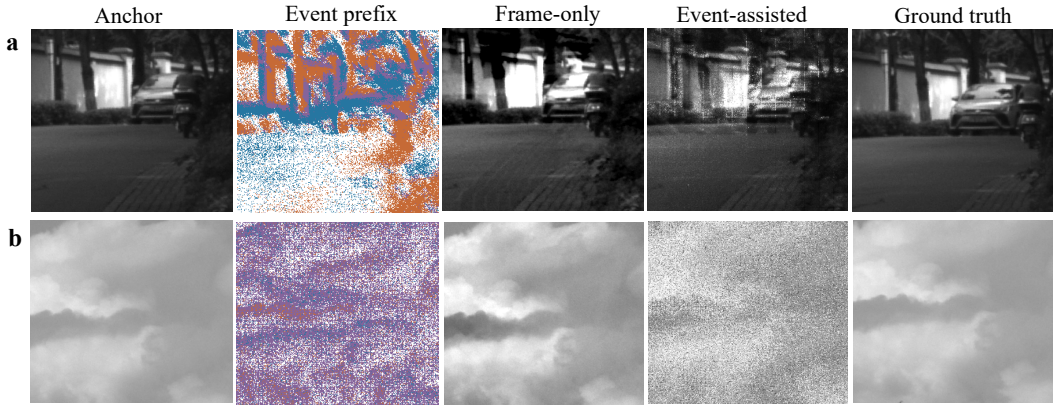

Figure 3: Selective event use on development-test sequences. **a**, Accepted beneficial round; **b**, rejected harmful round. Green outlines indicate the selected branch. Full-frame context, event-color definitions, and failure cases are provided in Appendix F.

Limitations and future work. The current certificate controls round-weighted selected harm under the stated round-level i.i.d. assumption. Sequence-cluster bootstrap diagnostics in Appendix D quantify sensitivity to within-video dependence, while formal sequence-level certification and deployment shift remain directions for future work.

5 Conclusion

We introduced EventWorth to determine when event information should assist temporally causal video prediction relative to a frame-only fallback. Delayed visual anchors provide paired baseline-relative supervision, validation freezes the selector, and an independent certificate determines whether event assistance is enabled. In the analytic setting, EventWorth passes all five marginal certificates with 50.44% development-test coverage, 1.30% harm, and 0.0220 system gain. Paired future-anchor supervision also produces stronger gain and harm signals than reliability-derived or shuffled training targets. The resulting system uses event assistance selectively and retains the frame-only branch whenever the certificate is not met.

AI USE STATEMENT

Generative AI tools were used to assist in deriving and drafting the mathematical proofs of the selected-risk certification result and its associated corollaries. In particular, AI assistance was used in developing proof arguments and in writing the corresponding proof steps. The authors reviewed and verified the AI-assisted derivations against the stated assumptions and mathematical claims. We take full responsibility for the correctness and final content of all theoretical results and proofs presented in this work.

REPRODUCIBILITY STATEMENT

The assumptions and complete proofs supporting the theoretical guarantees are provided in Appendix A. Appendix B documents the data partitioning, causal preprocessing, predictor and score model definitions, optimization settings, deterministic proposal protocols, certificate budgets, and random seeds. The subsequent experimental appendices provide the corresponding per-seed decision records, ablation protocols, diagnostic evaluations, and qualitative-example provenance. Together, these materials specify the theoretical and experimental procedures underlying the reported results. An anonymized implementation and the corresponding configuration files are provided in the supplementary material.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online least squares estimation with self-normalized processes: An application to bandit problems. *arXiv preprint arXiv:1102.2670*, 2011. URL https://arxiv.org/abs/1102.2670.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/f3549ef9b5ff520a7e41ff3cc306ab2b-Abstract-Conference.html.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):43:1–43:34, 2021. doi: 10.1145/3478535.

Jingxi Chen, Brandon Y. Feng, Haoming Cai, Tianfu Wang, Levi Burner, Dehao Yuan, Cornelia Fermüller, Christopher A. Metzler, and Yiannis Aloimonos. Repurposing pre-trained video diffusion models for event-based video interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12456–12466, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Chen_Repurposing_Pre-trained_Video_Diffusion_Models_for_Event-based_Video_Interpolation_CVPR_2025_paper.html.

Hoonhee Cho, Taewoo Kim, Yuhwan Jeong, and Kuk-Jin Yoon. TTA-EVF: Test-time adaptation for event-based video frame interpolation via reliable pixel and sample estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25701–25711, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Cho_TTA-EVF_Test-Time_Adaptation_for_Event-based_Video_Frame_Interpolation_via_Reliable_CVPR_2024_paper.html.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Siyi Du, Xinzhe Luo, Declan P. O'Regan, and Chen Qin. Inference-time dynamic modality selection for incomplete multimodal classification. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=PWhDUWRVhM.

Yuxing Duan. LED: A large-scale real-world paired dataset for event camera denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25637–25647, 2024.

Burak Ercan, Onur Eker, Aykut Erdem, and Erkut Erdem. EVREAL: Towards a comprehensive benchmark and analysis suite for event-based video reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3943–3952, 2023. URL https://openaccess.thecvf.com/content/CVPR2023W/EventVision/html/Ercan_EVREAL_Towards_a_Comprehensive_Benchmark_and_Analysis_Suite_for_Event-Based_CVPRW_2023_paper.html.

Shai Feldman, Liran Ringel, Stephen Bates, and Yaniv Romano. Achieving risk control in online learning settings. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=5Y04GWvoJu.

Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2022. doi: 10.1109/TPAMI.2020.3008413.

Qiyao Gao, Peiqi Duan, Hanyue Lou, Minggui Teng, Ziqi Cai, Xu Chen, and Boxin Shi. Unified reconstruction of static and dynamic scenes from events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27914–27923, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Gao_Unified_Reconstruction_of_Static_and_Dynamic_Scenes_from_Events_CVPR_2025_paper.html.

Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26876–26885, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Gao_Embracing_Unimodal_Aleatoric_Uncertainty_for_Robust_Multimodal_Fusion_CVPR_2024_paper.html.

Zongbo Han, Fan Yang, Junzhou Huang, Changqing Zhang, and Jianhua Yao. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20707–20717, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Han_Multimodal_Dynamics_Dynamical_Fusion_for_Trustworthy_Multimodal_Classification_CVPR_2022_paper.html.

Weihua He, Kaichao You, Zhendong Qiao, Xu Jia, Ziyang Zhang, Wenhui Wang, Huchuan Lu, Yaoyuan Wang, and Jianxing Liao. Timereplayer: Unlocking the potential of event cameras for video interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17804–17813, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/He_TimeReplayer_Unlocking_the_Potential_of_Event_Cameras_for_Video_Interpolation_CVPR_2022_paper.html.

Yifei He, Runxiang Cheng, Gargi Balasubramaniam, Yao-Hung Hubert Tsai, and Han Zhao. Efficient modality selection in multimodal learning. *Journal of Machine Learning Research*, 25(47):1–39, 2024. URL https://jmlr.org/papers/v25/23-0439.html.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Benjamin Howson, Ciara Pike-Burke, and Sarah Filippi. Delayed feedback in generalised linear bandits revisited. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 6095–6119, 2023. URL https://proceedings.mlr.press/v206/howson23b.html.

Adam Tauman Kalai and Varun Kanade. Towards optimally abstaining from prediction with OOD test examples. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/6a26c75d6a576c94654bfc4dda548c72-Abstract.html.

Nikos Karampatziakis, Paul Mineiro, and Aaditya Ramdas. Off-policy confidence sequences. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5301–5310, 2021. URL https://proceedings.mlr.press/v139/karampatziakis21a.html.

Junho Kim, Inwoo Hwang, and Young Min Kim. Ev-TTA: Test-time adaptation for event-based object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17745–17754, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Kim_Ev-TTA_Test-Time_Adaptation_for_Event-Based_Object_Recognition_CVPR_2022_paper.html.

Taewoo Kim, Yujeong Chae, Hyun-Kurl Jang, and Kuk-Jin Yoon. Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18032–18042, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Kim_Event-Based_Video_Frame_Interpolation_With_Cross-Modal_Asymmetric_Bidirectional_Motion_Fields_CVPR_2023_paper.html.

Guoqiang Liang, Kanghao Chen, Hangyu Li, Yunfan Lu, and Lin Wang. Towards robust event-guided low-light image enhancement: A large-scale real-world event-image dataset and novel approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23–33, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liang_Towards_Robust_Event-guided_Low-Light_Image_Enhancement_A_Large-Scale_Real-World_Event-Image_CVPR_2024_paper.html.

Haoyue Liu, Shihan Peng, Lin Zhu, Yi Chang, Hanyu Zhou, and Luxin Yan. Seeing motion at nighttime with an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25648–25658, 2024a.

Yuhan Liu, Yongjian Deng, Hao Chen, and Zhen Yang. Video frame interpolation via direct synthesis with the event-based reference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8477–8487, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_Video_Frame_Interpolation_via_Direct_Synthesis_with_the_Event-based_Reference_CVPR_2024_paper.html.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18177–18186, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Ma_Are_Multimodal_Transformers_Robust_to_Missing_Modality_CVPR_2022_paper.html.

Harsh Maheshwari, Yen-Cheng Liu, and Zsolt Kira. Missing modality robustness in semi-supervised multi-modal semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1020–1030, 2024. URL https://openaccess.thecvf.com/content/WACV2024/html/Maheshwari_Missing_Modality_Robustness_in_Semi-Supervised_Multi-Modal_Semantic_Segmentation_WACV_2024_paper.html.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8238–8247, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Peng_Balanced_Multimodal_Learning_via_On-the-Fly_Gradient_Modulation_CVPR_2022_paper.html.

Andrea Pugnana and Salvatore Ruggieri. AUC-based selective classification. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 2494–2514, 2023. URL https://proceedings.mlr.press/v206/pugnana23a.html.

Lei Sun, Christos Sakaridis, Jingyun Liang, Peng Sun, Jiezhang Cao, Kai Zhang, Qi Jiang, Kaiwei Wang, and Luc Van Gool. Event-based frame interpolation with ad-hoc deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18043–18052, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Sun_Event-Based_Frame_Interpolation_With_Ad-Hoc_Deblurring_CVPR_2023_paper.html.

Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16155–16164, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Tulyakov_Time_Lens_Event-Based_Video_Frame_Interpolation_CVPR_2021_paper.html.

Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17755–17764, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Tulyakov_Time_Lens_Event-Based_Frame_Interpolation_With_Parametric_Non-Linear_Flow_and_CVPR_2022_paper.html.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12695–12705, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_What_Makes_Training_Multi-Modal_Classification_Networks_Hard_CVPR_2020_paper.html.

Ziyun Wang, Friedhelm Hamann, Kenneth Chaney, Wen Jiang, Guillermo Gallego, and Kostas Daniilidis. Event-based continuous color video decompression from single frames. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 5007–5017, 2025. URL https://openaccess.thecvf.com/content/CVPR2025W/EventVision/html/Wang_Event-based_Continuous_Color_Video_Decompression_from_Single_Frames_CVPRW_2025_paper.html.

Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2563–2572, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Weng_Event-Based_Video_Reconstruction_Using_Transformer_ICCV_2021_paper.html.

Zihui Xue and Radu Marculescu. Dynamic multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2575–2584, 2023. URL https://openaccess.thecvf.com/content/CVPR2023W/MULA/html/Xue_Dynamic_Multimodal_Fusion_CVPRW_2023_paper.html.

Yixin Yang, Jinxiu Liang, Bohan Yu, Yan Chen, Jimmy S. Ren, and Boxin Shi. Latency correction for event-guided deblurring and frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24977–24986, 2024a. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yang_Latency_Correction_for_Event-guided_Deblurring_and_Frame_Interpolation_CVPR_2024_paper.html.

Zequn Yang, Yake Wei, Ce Liang, and Di Hu. Quantifying and enhancing multi-modal robustness with modality preference. In *International Conference on Learning Representations*, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/45a9d4290ae75ceb752b2a61be1d2baa-Abstract-Conference.html.

Ahmed Zaoui, Christophe Denis, and Mohamed Hebiri. Regression with reject option and application to kNN. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e8219d4c93f6c55c6b10fe6bfe997c6c-Abstract.html.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41753–41769, 2023. URL https://proceedings.mlr.press/v202/zhang23ar.html.

Xiang Zhang and Lei Yu. Unifying motion deblurring and frame interpolation with events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17765–17774, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Zhang_Unifying_Motion_Deblurring_and_Frame_Interpolation_With_Events_CVPR_2022_paper.html.

Yunhao Zou, Yinqiang Zheng, Tsuyoshi Takatani, and Ying Fu. Learning to reconstruct high speed and high dynamic range videos from events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2024–2033, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zou_Learning_To_Reconstruct_High_Speed_and_High_Dynamic_Range_Videos_CVPR_2021_paper.html.

A Proofs and Additional Certification Details

A.1 Proof of Theorem 1

Proof. Condition on $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{val}}$. Under the theorem assumptions, this fixes $\tau^\star$ while leaving the certificate rounds independent and identically distributed. Let:

$$\rho = \Pr(H = 1 \mid \pi_{\tau^\star}(X) = 1)\,. \tag{22}$$

For each certificate round, $I_i = \pi_{\tau^\star}(X_i)$ is a deterministic function of X_i. Conditional on any realized selected index set, the corresponding harm labels are independent Bernoulli variables with common success probability ρ. Hence, after averaging over selected index sets of size n,

$$K \mid N = n \sim \text{Binomial}(n, \rho). \tag{23}$$

By the defining coverage property of the one-sided Clopper–Pearson limit:

$$\Pr_{\rho}(\rho > U_{\text{CP}}(K, n; \eta) \mid N = n) \le \eta \tag{24}$$

for every $n > 0$ and every $\rho \in [0, 1]$. Averaging over the random selected count N gives the conditional form of Eq. (19). Since this conditional bound holds for every realization of $(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{val}})$ for which the frozen policy is defined, averaging once more over the training and validation data yields the unconditional statement in Eq. (19).

If $C = 1$, then $N \ge N_{\min} > 0$ and $U_{\text{CP}}(K, N; \eta) \le \alpha$. Therefore,

$$\{C = 1,\ \rho > \alpha\} \subseteq \{N > 0,\ \rho > U_{\text{CP}}(K, N; \eta)\}. \tag{25}$$

Applying Eq. (19) proves Eq. (20). □

A.2 Proof of Corollary 1

Proof. When $C = 0$, Eq. (17) gives $a(x) = 0$ for every context, so the event-assisted branch introduces no additional loss relative to the frame-only fallback. When $C = 1$, Theorem 1 implies $R_h(\pi_{\tau^\star}) \le \alpha$ except on an event of probability at most η. □

A.3 Proof of Corollary 2

Proof. If $H = 0$, then $r \le \delta_h$. If $H = 1$, bounded losses imply $r \le 1$. Hence:

$$r \le \delta_h + (1 - \delta_h)H. \tag{26}$$

Conditioning on $\pi_{\tau^\star}(X) = 1$ gives:

$$\mathbb{E}[r \mid \pi_{\tau^\star}(X) = 1] \le \delta_h + (1 - \delta_h)R_h(\pi_{\tau^\star}). \tag{27}$$

On the accepted policy-risk certificate event outside the failure event of Theorem 1, the right-hand side is at most $\delta_h + (1 - \delta_h)\alpha$. □

A.4 Zero-harm policy-risk certificate size

When $K = 0$, the one-sided exact upper limit has the closed form:

$$U_{\text{CP}}(0, N; \eta) = 1 - \eta^{1/N}. \tag{28}$$

Therefore, a zero-harm candidate satisfies: $U_{\text{CP}}(0, N; \eta) \le \alpha$ whenever

$$N \ge \left\lceil \frac{\log \eta}{\log(1 - \alpha)} \right\rceil. \tag{29}$$

For $\alpha = \eta = 0.05$, this threshold is 59 selected certificate rounds. The implementation additionally requires $N \ge 128$ to avoid accepting an operating point with weak selected-sample support.

A.5 MULTIPLE PREDECLARED SELECTORS

Theorem 1 applies to any one selector fixed without using $\mathcal{D}_{\mathrm{cert}}$. If a single predeclared deployment selector is tested, it may use the full failure budget η. More generally, suppose M selectors $\pi_1, \ldots, \pi_M$ are all fixed using only training and validation data. Assign nonnegative failure budgets $\eta_1, \ldots, \eta_M$ satisfying:

$$\sum_{j=1}^{M} \eta_j \leq \eta. \tag{30}$$

If selector j is accepted only when $U_{\mathrm{CP}}(K_j, N_j; \eta_j) \leq \alpha$, then a union bound and Theorem 1 give:

$$\Pr(\exists j : \ C_j = 1 \text{ and } R_h(\pi_j) > \alpha) \leq \sum_{j=1}^{M} \eta_j \leq \eta. \tag{31}$$

Equal Bonferroni allocation uses $\eta_j = \eta/M$. Thus a single-selector certificate with $\eta = 0.05$ uses the full budget, whereas the predeclared joint-versus-risk-only comparison uses $M = 2$ and 0.025 per selector. The same rule extends to larger simultaneous method families. Multiplicity correction is needed for the family-wise comparison, not because either selector intrinsically requires a smaller failure budget.

A.6 FUTURE-ANCHOR PROXY FIDELITY

The final policy-risk certificate uses true suffix harm labels on $\mathcal{D}_{\mathrm{cert}}$, so proxy fidelity is not an assumption of Theorem 1. It remains relevant because the score models are trained on future-anchor proxy targets.

Let:

$$r_k(u) = \ell(\widehat{y}_{1,k}(u), y_k(u)) - \ell(\widehat{y}_{0,k}(u), y_k(u)). \tag{32}$$

If r_k is L_k^{reg}-Lipschitz on $[0, 1)$ and $|\widetilde{r}_k - r_k(1^-)| \leq \varepsilon_k$, then:

$$|\widetilde{r}_k - r_k| \leq \frac{L_k^{\mathrm{reg}}}{2} + \varepsilon_k, \tag{33}$$

where $r_k = \int_0^1 r_k(u)\, du$. This follows by integrating $|r_k(1^-) - r_k(u)| \leq L_k^{\mathrm{reg}}(1 - u)$.

A probabilistic version makes the effect on proxy labels explicit. Suppose there are $\mathcal{G}_k$-measurable $\bar{b}_k$ and ζ_k such that:

$$\Pr\big(|\widetilde{r}_k - r_k| > \bar{b}_k \mid \mathcal{G}_k\big) \leq \zeta_k, \tag{34}$$

and suppose:

$$\Pr(|r_k - \delta_h| \leq z \mid \mathcal{G}_k) \leq C_{\mathrm{m}} z^{\kappa}. \tag{35}$$

Then:

$$\Pr(\widetilde{h}_k \neq h_k \mid \mathcal{G}_k) \leq \min\big\{1,\, C_{\mathrm{m}} \bar{b}_k^{\kappa} + \zeta_k\big\}, \tag{36}$$

$$|\mathbb{E}[\widetilde{g}_k - g_k \mid \mathcal{G}_k]| \leq \bar{b}_k + 2\zeta_k. \tag{37}$$

Proof of Eqs. (36)–(37). Let:

$$\mathcal{E}_k = \big\{|\widetilde{r}_k - r_k| \leq \bar{b}_k\big\}. \tag{38}$$

On $\mathcal{E}_k$, if the proxy and suffix harm labels disagree, then $\widetilde{r}_k$ and r_k lie on opposite sides of the threshold δ_h. Hence:

$$\{\widetilde{h}_k \neq h_k\} \cap \mathcal{E}_k \subseteq \{|r_k - \delta_h| \leq \bar{b}_k\}. \tag{39}$$

Conditioning on $\mathcal{G}_k$ therefore gives:

$$\begin{aligned} \Pr(\widetilde{h}_k \neq h_k \mid \mathcal{G}_k) &\leq \Pr(|r_k - \delta_h| \leq \bar{b}_k \mid \mathcal{G}_k) + \Pr(\mathcal{E}_k^c \mid \mathcal{G}_k) \\ &\leq C_{\mathrm{m}} \bar{b}_k^{\kappa} + \zeta_k. \end{aligned} \tag{40}$$

Since the left-hand side is a probability, taking the minimum with one proves Eq. (36).

For gain, $\widetilde{g}_k - g_k = -(\widetilde{r}_k - r_k)$. The normalized branch losses imply $r_k, \widetilde{r}_k \in [-1, 1]$, so $|\widetilde{r}_k - r_k| \le 2$ everywhere. Consequently,

$$\begin{aligned} |\mathbb{E}[\widetilde{g}_k - g_k \mid \mathcal{G}_k]| &\le \mathbb{E}[|\widetilde{r}_k - r_k| \mid \mathcal{G}_k] \\ &\le \bar{b}_k \Pr(\mathcal{E}_k \mid \mathcal{G}_k) + 2 \Pr(\mathcal{E}_k^c \mid \mathcal{G}_k) \\ &\le \bar{b}_k + 2\zeta_k, \end{aligned} \tag{41}$$

which proves Eq. (37). □

These bounds explain when the delayed endpoint is a useful training proxy. They are reported as proxy-fidelity diagnostics and are not added to the deployed thresholds or the policy-risk certificate.

A.7 Online confidence-sequence baselines

The code also contains `linear_eventworth` and `static_eventworth` diagnostic variants. A representative online linear variant forms:

$$M_k = \lambda I + \sum_{i \in \mathcal{A}_k} \phi_i \phi_i^\top, \qquad \widehat{\theta}_k = M_k^{-1} \sum_{i \in \mathcal{A}_k} \phi_i \widetilde{h}_i, \tag{42}$$

and uses a self-normalized radius proportional to $\|\phi_k\|_{M_k^{-1}}$ (Abbasi-Yadkori et al., 2011). Such variants study the conservativeness of sequential pointwise confidence bounds under delayed proxy feedback. These variants are included only as ablations or diagnostics; they are not part of the final EventWorth policy and do not enter Theorem 1.

A.8 Round-level versus sequence-level dependence

The exact theorem treats certificate rounds as independent and identically distributed. In video data, rounds within one sequence may be correlated. Sequence-disjoint splitting removes direct overlap across train, validation, certificate, and test sets, but it does not turn within-sequence rounds into independent observations. Sequence-cluster bootstrap intervals can reveal whether observed harms are dominated by a small number of sequences, but they are diagnostic rather than a sequence-level policy-risk certificate. A formal sequence-level claim would require a cluster-aware certification rule and is outside the current implementation.

A.9 Validation proposal statistics

For a selector π evaluated on m_{val} validation rounds, define:

$$\begin{aligned} N_{\mathrm{val}}(\pi) &= \sum_{i=1}^{m_{\mathrm{val}}} \pi(X_i), & K_{\mathrm{val}}(\pi) &= \sum_{i=1}^{m_{\mathrm{val}}} \pi(X_i) h_i, \\ \widehat{\mathrm{Cov}}_{\mathrm{val}}(\pi) &= \frac{N_{\mathrm{val}}(\pi)}{m_{\mathrm{val}}}, & \widehat{G}_{\mathrm{sys,val}}(\pi) &= \frac{1}{m_{\mathrm{val}}} \sum_{i=1}^{m_{\mathrm{val}}} \pi(X_i) g_i, \\ \widehat{G}_{\mathrm{sel,val}}(\pi) &= \frac{\sum_{i=1}^{m_{\mathrm{val}}} \pi(X_i) g_i}{N_{\mathrm{val}}(\pi) \vee 1}. \end{aligned} \tag{43}$$

The certification theorem does not require a particular validation-search algorithm; it requires only that the resulting selector be fixed without using $\mathcal{D}_{\mathrm{cert}}$. In all experiments, selector thresholds are chosen deterministically from prespecified validation grids using minimum selected support, coverage and utility criteria, and fixed tie-breaking rules. The matched risk-only threshold is chosen on the same validation split to approximate the joint selector's validation coverage, with a predeclared fallback target when no joint proposal exists. No certificate or development-test statistic enters either proposal.

The two recorded predictor studies use different frozen proposal protocols: the analytic suite predates the later validation-CP feasibility screen, whereas the learned suite uses that stricter screen. Their exact threshold grids, coverage ranges, validation budgets, fallback target, and tie-breaking orders are reported in Appendix B.3. This distinction is implementation-specific and does not alter Theorem 1.

A.10 SELECTOR-LEVEL COMPARISON PROTOCOL

Certification answers whether a frozen selector may be deployed; it does not by itself justify a claim that one selector architecture is universally necessary. We therefore separate *policy certification* from *selector-level claim selection*. For the learned-predictor comparison, the seeds

$$\mathcal{S} = \{2033, 2034, 2035, 2036, 2037\}$$

and the decision rule below are fixed before running these seeds and before observing any corresponding outcomes.

For selector $a \in \{\text{joint}, \text{risk}\}$, let $C_s^a \in \{0, 1\}$ denote the family-wise-corrected certificate decision on seed s. Let $\widehat{G}^a_{\text{sys},s}$ and $\widehat{H}^a_s$ denote its fail-closed development-test system gain and empirical development-test harm rate. Consistent with Sec. 3.2, a failed certificate is evaluated as reject-all and therefore contributes zero event-use gain and zero additional event-induced harm. Define the five-seed means:

$$\overline{G}^a_{\text{sys}} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \widehat{G}^a_{\text{sys},s}, \qquad \overline{H}^a = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \widehat{H}^a_s. \tag{44}$$

We also define seed-level support counts. Joint support requires a per-seed system-gain advantage of at least $\Delta_G = 0.001$,

$$S_{\text{joint}} = \sum_{s \in \mathcal{S}} \mathbf{1}\Big\{ \widehat{G}^{\text{joint}}_{\text{sys},s} \geq \widehat{G}^{\text{risk}}_{\text{sys},s} + \Delta_G \Big\}. \tag{45}$$

Risk-only support uses the predeclared noninferiority margin $\Delta_{\text{NI}} = 0.001$ and zero test-harm tolerance,

$$S_{\text{risk}} = \sum_{s \in \mathcal{S}} \mathbf{1}\Big\{ \widehat{G}^{\text{risk}}_{\text{sys},s} \geq \widehat{G}^{\text{joint}}_{\text{sys},s} - \Delta_{\text{NI}}, \quad \widehat{H}^{\text{risk}}_s \leq \widehat{H}^{\text{joint}}_s \Big\}. \tag{46}$$

The outcome `retain_joint` requires all four conditions:

$$\sum_{s \in \mathcal{S}} C^{\text{joint}}_s \geq 4, \qquad S_{\text{joint}} \geq 3, \qquad \overline{G}^{\text{joint}}_{\text{sys}} \geq \overline{G}^{\text{risk}}_{\text{sys}} + 0.001, \qquad \overline{H}^{\text{joint}} \leq \alpha. \tag{47}$$

If this rule is not met, the outcome `promote_risk_only` requires:

$$\sum_{s \in \mathcal{S}} C^{\text{risk}}_s \geq 4, \qquad S_{\text{risk}} \geq 3, \qquad \overline{G}^{\text{risk}}_{\text{sys}} \geq \overline{G}^{\text{joint}}_{\text{sys}} - 0.001, \qquad \overline{H}^{\text{risk}} \leq \overline{H}^{\text{joint}}. \tag{48}$$

The last inequality corresponds to the prespecified `test_harm_tolerance`=0. If neither criterion is satisfied, no architecture-level preference is reported. The selector comparison affects only the empirical comparison between joint and risk-only policies; the policy-risk certificate is unchanged.

This selector-level rule is an empirical claim protocol, not an additional risk theorem. It changes only the strength of the architecture-level conclusion and does not alter Theorem 1, the policy-risk certificate, or any per-seed deployment decision.

A.11 IMPLEMENTATION NOMENCLATURE

For backward compatibility, the configuration key `online` denotes $\mathcal{D}_{\text{val}}$ for the frozen primary policy; it does not imply online score updates. Future-anchor agreement outputs are *proxy-fidelity diagnostics*, whereas the Clopper–Pearson test defining C is the *policy-risk certificate*. The matched risk-only coverage target is also determined on $\mathcal{D}_{\text{val}}$ only.

B EXPERIMENTAL PROTOCOLS AND REPRODUCIBILITY

B.1 DATA PARTITION AND EVALUATION UNITS

RLED contains 100 official training sequences and 12 official test sequences (Liu et al., 2024a). We use all 112 sequences, with 15 rounds constructed per sequence. The realized train/validation/certificate/development-test partition contains 31/21/48/12 sequence-disjoint sequences, or 465/315/720/180 rounds. The same 12 supplied test sequences are reused across seeds

Table 5: Recorded experimental configuration. Configured caps and realized values are distinguished where relevant. These values define the reported experiments; earlier online-calibration parameters do not define the final policy-risk certificate.

Setting	Recorded value
Sequences / rounds per run	112 / 1,680
Train / validation / certificate / dev-test	31 / 21 / 48 / 12 sequences; 465 / 315 / 720 / 180 rounds
Crop; anchor / round stride	256×256 center crop; 6 / 6 frames
Decision ratio; realized suffix targets	0.35 of the interval; offsets 3, 4, 5 (three targets)
Loss; harm margin	MAE on $[0, 1]$-normalized intensities; $\delta_h = 0.005$
Risk target; selected-sample requirements	$\alpha = 0.05$; $N_{\text{val}} \geq 64$, $N_{\text{cert}} \geq 128$
Threshold grids	41 harm-score quantiles; 21 gain-score quantiles plus $\gamma = 0$
Learned validation proposal	$\eta_{\text{val}} = 0.05$; coverage $[0.25, 0.75]$; fallback target 0.35
Analytic validation proposal	primary coverage floor 0.40; legacy reserve interval $[0.55, 0.75]$
Certificate failure budgets	marginal single selector 0.05; learned pair 0.025 each; eight-slot audit 0.00625 each
Analytic event branch	contrast scale 0.20; event blend 0.75; hot-pixel quantile 0.995
Health score model	hidden width 64; embedding 8; BCE + smooth-L1 heads; 100 epochs; batch 128
Health optimizer / augmentation	AdamW, 10^{-3}, decay 10^{-4}; one corrupted copy per clean training round
Learned residual branches	64×64; width 16; 3 convolutions; residual scale 0.15; 12 epochs; batch 32
Branch optimizer / cap	Adam, 10^{-3}, decay 10^{-5}; at most 1,536 examples per branch
Analytic / learned seeds	2028–2032 / 2033–2037
Sequence bootstrap	10,000 whole-sequence resamples per split and seed
Perturbation grid	analytic only; 6 operators $\times$ 4 severities $\times$ 5 seeds

and were inspected during development; they are therefore reported as development-test evidence. Sequence-disjoint splitting prevents direct overlap across the four roles but does not imply independence among rounds from the same video. Exact run seeds and all protocol constants are summarized in Table 5.

B.2 Causal preprocessing, predictors, and score models

Images are normalized to $[0, 1]$, center-cropped to 256×256, and events are spatially aligned and time-sorted after cropping. Consecutive anchor-delimited rounds use two past anchors; the future anchor is withheld during prediction and is used only after the round for paired supervision. With the recorded stride and decision ratio, each round realizes three suffix targets, and the suffix objective is their mean MAE on $[0, 1]$-normalized intensities.

The analytic frame branch uses clipped linear extrapolation from the two past frames. Its event-assisted counterpart adds reliability-weighted signed event evidence to the current log intensity and blends the result with the fallback. The learned setting adds a small residual CNN to each analytic branch while preserving the same causal inputs. For the learned branches, the 256×256 image crop, event features, and training targets are bilinearly downsampled to 64×64. The CNN predicts a residual $0.15 \tanh f_\theta(\cdot)$ at this resolution. At inference, only this residual is bilinearly upsampled to 256×256 and added to the corresponding analytic prediction; losses are evaluated on the final clipped 256×256 output. All resizing uses `align_corners=False`.

The selector score model standardizes causal sensor, reliability, observability, and cross-modal features and learns separate gain and harm heads. Predictor branches and score heads are frozen before validation proposal. Architecture, optimization, and augmentation details are consolidated in Table 5; they are not repeated in the main text.

Table 6: Analytic joint proposals and independent certificates. U_{val} and $U_{\mathrm{cert}}^{(1)}$ use failure level 0.05; $U_{\mathrm{cert}}^{(8)}$ uses 0.00625. All certificate samples satisfy $N \geq 128$. The archived analytic validation screen used empirical harm, not $U_{\mathrm{val}} \leq 0.05$.

Seed	N_{val}	K_{val}	U_{val}	N_{cert}	K_{cert}	$U_{\mathrm{cert}}^{(1)}$	$U_{\mathrm{cert}}^{(8)}$
2028	126	3	0.0604	310	7	0.0420	0.0533
2029	128	6	0.0904	367	7	0.0355	0.0451
2030	126	2	0.0491	424	4	0.0215	0.0287
2031	132	5	0.0780	381	3	0.0202	0.0278
2032	130	4	0.0690	343	3	0.0224	0.0308

B.3 Recorded validation-proposal protocols

The archived analytic suite predates the stricter learned validation screen. Its primary proposal requires at least 64 selected validation rounds, empirical harm at most 0.05, positive selected mean gain, and validation coverage at least 0.40. When available, zero-observed-harm candidates are preferred; the recorded ranking then uses the gain threshold, validation risk upper bound, and validation system gain. All five reported analytic proposals come from this primary `conservative_high_gain` regime. In particular, the analytic proposals should not be described as satisfying the later requirement $U_{\mathrm{CP}}^{\mathrm{val}} \leq 0.05$.

The learned proposal uses at least 64 selected validation rounds, coverage $[0.25, 0.75]$, $U_{\mathrm{CP}}^{\mathrm{val}} \leq 0.05$, and positive selected and system gain. Feasible candidates maximize validation system gain, with fixed tie-breaking by lower validation risk upper bound, higher coverage, and higher selected gain. The matched risk-only selector targets the joint validation coverage, or 0.35 when no feasible joint proposal exists. Both selectors are frozen before certification and development-test evaluation.

Theorem 1 applies to either recorded validation-frozen proposal under its stated sampling assumptions. The analytic and learned studies therefore instantiate the same certification theory but not an identical proposal algorithm; differences between them should not be attributed solely to predictor architecture. Table 6 records the resulting validation proposals and independent certificate outcomes for all five analytic seeds.

B.4 Controlled Supervision Ablations

The controlled supervision ablation in Table 2 keeps the predictor pair, causal features, score architecture, optimization, validation proposal, and evaluation protocol fixed, and changes only the targets used to train the gain and harm scoring heads. This isolates whether performance depends on the task-relative supervision supplied by delayed future anchors rather than on architecture or optimization changes.

Shuffled paired-label control. To test whether the learned scores depend on the round-specific association between causal features and delayed supervision, we construct a sequence-blocked derangement on $\mathcal{D}_{\mathrm{train}}$. Training sequences are first grouped by domain and frame rate. Within each group, we draw a permutation d over sequences with no fixed points, so that every receiving sequence s is assigned a distinct donor sequence $d(s) \neq s$. For round position j, the receiving round retains its original causal inputs and features but receives the paired proxy regret from the corresponding round of its donor sequence:

$$\widetilde{r}_{s,j}^{\mathrm{shuf}} = \widetilde{r}_{d(s),j}, \qquad \widetilde{g}_{s,j}^{\mathrm{shuf}} = -\widetilde{r}_{s,j}^{\mathrm{shuf}}, \qquad \widetilde{h}_{s,j}^{\mathrm{shuf}} = \mathbf{1}\left\{\widetilde{r}_{s,j}^{\mathrm{shuf}} > \delta_h\right\}. \tag{49}$$

The gain and harm targets are therefore transferred jointly and remain internally consistent, while their association with the receiving round is removed. Because the mapping is a within-group permutation, the empirical proxy-label distribution within each domain/frame-rate group and round position is preserved, while round-specific feature–label correspondence is destroyed.

The derangement is applied only to $\mathcal{D}_{\mathrm{train}}$. Predictor outputs, causal features, score architecture, optimization, validation proposal, independent certification, and development-test evaluation remain

unchanged. One deterministic derangement is fixed in advance for each experimental seed and is never selected or modified using validation, certificate, or development-test outcomes.

Marginal-matched reliability control. Let $R_{s,j} \in [0,1]$ denote the causal reliability statistic computed from events available at the decision time of round j in sequence s. This is the same `reliability_mean` statistic used by the reliability selector.

To construct a reliability-based supervision control without changing the target marginal, we perform a monotone rearrangement within each domain–frame-rate stratum of $\mathcal{D}_{\mathrm{train}}$. Rounds are sorted by reliability,

$$R_{(1)} \le \cdots \le R_{(n)},$$

while the observed future-anchor proxy regrets in the same stratum are sorted in descending order,

$$\widetilde{r}_{[1]} \ge \cdots \ge \widetilde{r}_{[n]}.$$

We then assign

$$\widetilde{r}^{\mathrm{rel}}_{(k)} = \widetilde{r}_{[k]}, \qquad \widetilde{g}^{\mathrm{rel}}_{(k)} = -\widetilde{r}^{\mathrm{rel}}_{(k)}, \qquad \widetilde{h}^{\mathrm{rel}}_{(k)} = \mathbf{1}\{\widetilde{r}^{\mathrm{rel}}_{(k)} > \delta_h\}.$$

Thus, lower reliability is paired with larger event-induced regret while the multiset of gain targets, harm prevalence, and gain–harm consistency are preserved within each stratum. Future-anchor values determine only the stratum-level target marginal; the round-wise assignment is determined by the causal reliability ranking. Reliability ties are broken deterministically by sequence ID and round ID.

Only the score-head targets on $\mathcal{D}_{\mathrm{train}}$ are changed. Predictor outputs, causal features, feature normalization, score architecture, optimization, validation proposal, certification, and development-test evaluation are otherwise identical to the paired-proxy condition.

True-suffix oracle supervision. As a non-causal upper-reference condition, we replace the delayed proxy targets with the true suffix quantities defined in Eq. (5):

$$g^{\mathrm{oracle}}_{s,j} = g_{s,j}, \qquad h^{\mathrm{oracle}}_{s,j} = h_{s,j}.$$

This condition uses information unavailable at deployment and is therefore not an EventWorth variant. It is included only to measure the gap between future-anchor proxy supervision and direct suffix supervision. All remaining training, proposal, certification, and evaluation procedures are unchanged.

B.5 Baselines and the meaning of a pass count

Density and reliability controls select high event-rate or reliability scores using validation quantiles. Analytic controls target 0.40 validation coverage; learned controls target joint validation coverage or the fallback target. Gain-only removes the risk condition from the joint thresholds. Unmatched risk-only removes the gain condition while retaining the joint risk threshold; matched risk-only chooses its own validation quantile. The hindsight oracle uses realized suffix gain and harm and is evaluation-only. It is not deployable and does not receive a policy-risk certificate. The random control uses a seed-and-round-ID hash ranking to select a fixed fraction of each batch. Because this exact-quota rule is not a pointwise frozen selector of the form assumed in Theorem 1, its logged CP calculation is descriptive and is excluded from the theorem's guarantee.

Single-policy pass counts are marginal assessments and do not jointly guarantee all compared policies. The eight-slot audit allocates 0.00625 to each recorded slot, while the predeclared learned comparison allocates 0.025 to each of two policies. These budgets provide within-run selector-family control and do not extend across seeds or predictor settings. For a rejected policy, the number of event-induced harms and total event-induced gain are zero. Conditional harm and selected mean gain are undefined when no round is selected; any zero used in system aggregation is an explicit computational convention.

C Extended clean results and selector comparison

Table 7 reports all analytic proposed-policy metrics, including selected mean gain. Always-off has zero event use and zero gain relative to itself. The oracle provides descriptive headroom rather than

Table 7: Full analytic proposed-policy results; mean ± sample SD across five seeds. Development-test outcomes are shown before family-corrected deployment filtering. The two pass counts use marginal and eight-slot budgets. The random control's logged pass calculations are descriptive only.

Selector	Coverage (%)	Harm (%)	$\widehat{G}_{\mathrm{sel}}$	$\widehat{G}_{\mathrm{sys}}$	Pass: single; eight
Frame-only fallback	0.00 ± 0.00	0.00 ± 0.00	0.0000 ± 0.0000	0.0000 ± 0.0000	–
Always use events	100.00 ± 0.00	4.44 ± 0.00	0.0301 ± 0.0000	0.0301 ± 0.0000	0/5; 0/5
Random matched	40.00 ± 0.00	5.00 ± 3.04	0.0296 ± 0.0027	0.0119 ± 0.0011	0/5; 0/5
Event density	53.33 ± 4.76	3.14 ± 0.27	0.0358 ± 0.0004	0.0191 ± 0.0016	0/5; 0/5
Reliability	45.33 ± 4.21	0.21 ± 0.48	0.0290 ± 0.0014	0.0132 ± 0.0018	4/5; 4/5
Gain-only	53.67 ± 5.09	2.40 ± 1.93	0.0422 ± 0.0064	0.0224 ± 0.0021	4/5; 3/5
Risk-only (unmatched)	83.33 ± 9.60	1.18 ± 0.66	0.0344 ± 0.0013	0.0287 ± 0.0029	4/5; 4/5
Risk-only (matched)	46.00 ± 6.05	0.98 ± 0.57	0.0404 ± 0.0058	0.0185 ± 0.0031	4/5; 4/5
EventWorth joint	50.44 ± 4.05	1.30 ± 0.37	0.0440 ± 0.0054	0.0220 ± 0.0019	5/5; 4/5
Hindsight oracle	92.78 ± 0.00	0.00 ± 0.00	0.0338 ± 0.0000	0.0314 ± 0.0000	–

Table 8: Analytic operational results under the eight-slot allocation. Failed certificates are replaced by the frame-only fallback before averaging all five runs.

Selector	Pass	Deployed coverage (%)	Deployed $\widehat{G}_{\mathrm{sys}}$
Always use events	0/5	0.00 ± 0.00	0.0000 ± 0.0000
Event density	0/5	0.00 ± 0.00	0.0000 ± 0.0000
Reliability	4/5	34.89 ± 19.58	0.0100 ± 0.0057
Gain-only	3/5	32.56 ± 29.93	0.0140 ± 0.0128
Risk-only (unmatched)	4/5	67.44 ± 38.84	0.0228 ± 0.0131
Risk-only (matched)	4/5	38.22 ± 21.85	0.0153 ± 0.0090
EventWorth joint	4/5	40.89 ± 23.17	0.0174 ± 0.0099

a fair deployable comparator. Table 8 applies the eight-slot certificate decisions before averaging, including every failed run. Joint retains more deployed system gain than the matched risk-only control in this setting, but unmatched risk-only retains more than joint.

C.1 Complete learned decision ledger

Table 9 records every learned seed under the two-selector failure budget. Seeds 2033 and 2035 have joint certificate upper bounds 0.0637 and 0.0680, exceeding 0.05 despite feasible validation proposals. Seeds 2036 and 2037 have no joint proposal. In seed 2034, both policies pass; joint's system-gain advantage is only 0.0000904, below the required 0.001 margin. Risk-only additionally passes in seed 2036, but its observed selected harm of 0.01493 exceeds joint's reject-all convention of zero. Consequently, that seed does not support the specified risk-only noninferiority claim. The resulting support counts are joint 0/5 and risk-only 1/5. Their acceptance counts, 1/5 and 2/5, also fall below 4/5. Neither selector satisfies the predeclared promotion rule. Support is counted only when the corresponding selector passes its certificate; joint support also requires that seed's observed test harm to be at most α. Double-failure seeds do not count as risk-only support merely because both logged gains and harms are zero.

D Proxy supervision and sequence dependence

Figure 4 summarizes the informativeness of delayed paired supervision across the analytic and learned predictor settings. On the 315 validation rounds per run, endpoint-proxy and suffix gains have Spearman correlations 0.9608 ± 0.0090 for analytic predictors and 0.9620 ± 0.0079 for learned predictors. Proxy-regret AUROC for suffix harm is 0.9814 ± 0.0138 and 0.9830 ± 0.0072, respectively. These proxy-regret values require the delayed future anchor and are not deployable gate scores. The corresponding causal health harm scores have AUROCs 0.8726 ± 0.0306 and 0.8843 ± 0.0250. Positive suffix-gain fractions are 0.8425 ± 0.0502 and 0.6419 ± 0.0664. Thus delayed anchors provide informative supervision in both settings, while good ranking alone does not guarantee a low-risk selected tail or a passing independent certificate. These are diagnostics for the proxy-fidelity analysis; they do not estimate a uniform proxy error bound or discharge its assumptions.

Table 9: Learned comparison with $\eta = 0.025$ per selector. K/N counts certificate harms and selections; U is the certificate upper bound. Coverage and gain are development-test outcomes after fail-closed deployment. Dashes indicate no joint proposal, not a successful zero-harm certificate.

	Joint					Matched risk-only				
Seed	K/N	U	Decision	Cov. (%)	G_{sys}	K/N	U	Decision	Cov. (%)	G_{sys}
2033	9/264	0.0637	fail	0.00	0.00000	13/279	0.0784	fail	0.00	0.00000
2034	2/218	0.0327	pass	37.78	0.01476	2/217	0.0329	pass	37.22	0.01467
2035	18/413	0.0680	fail	0.00	0.00000	18/420	0.0669	fail	0.00	0.00000
2036	–	–	no proposal	0.00	0.00000	2/238	0.0300	pass	37.22	0.01129
2037	–	–	no proposal	0.00	0.00000	18/269	0.1037	fail	0.00	0.00000

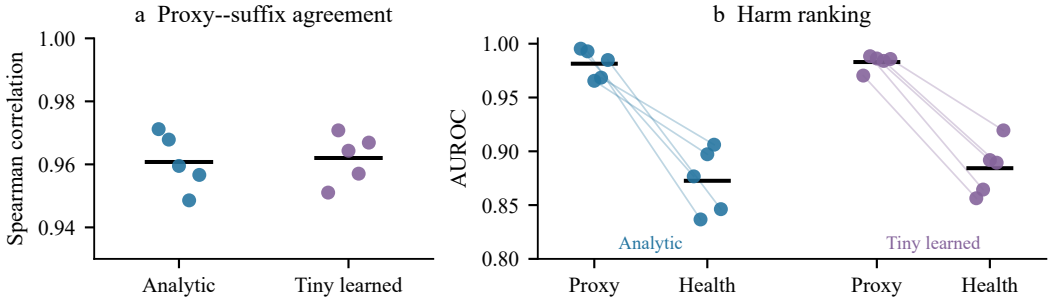

Figure 4: Informativeness of delayed paired supervision. **a**, Validation proxy–suffix gain correlation. **b**, Suffix-harm AUROC using delayed proxy regret or causal health scores. Points represent all five seeds per setting; bars mark means and lines connect paired rankings. Proxy quality alone is not a deployment certificate.

We additionally resample complete sequences, with replacement, for 10,000 replicates per seed and split. Each resample retains all rounds within each sampled sequence and recomputes round-weighted metrics. Figure 5 shows analytic development-test results. The 2.5th bootstrap percentile of system gain is positive in every seed, ranging from 0.01195 to 0.01405. Development-test harm's 97.5th percentile ranges from 0.03571 to 0.05128; it exceeds 0.05 in seed 2031. The certificate-split 97.5th percentile also reaches 0.05643 in seed 2028. These percentile diagnostics reveal sensitivity to sequence clustering. They are neither exact binomial certificates nor a proof of round independence.

E Perturbation diagnostics and predictor training

The analytic suite evaluates six event perturbations at severities 0.25, 0.50, 0.75, and 1.00 for every seed. Each cell contains 180 development-test rounds. Clean severity-zero copies are excluded when counting the 120 nonzero cells. Frames and targets are held fixed; event perturbations change causal inputs and branch outputs. The selector remains frozen and is not recertified on perturbed outcomes. Figure 6 shows means across all five seeds. Joint has observed harm at most 0.05 in 99/120 cells and positive system gain in 114/120. The perturbation sweep provides empirical robustness measurements rather than a distribution-shift certificate. Learned predictors have no corresponding stress sweep in the final experiment.

The operators are implementation-level diagnostics. For severity s, dropout removes each event with probability $0.85s$; polarity inversion flips its sign with probability s. The operator named `threshold_drift` removes events with probability $0.65s$; it is an event-thinning proxy, not a physical contrast-threshold simulator. Timestamp jitter adds Gaussian noise with standard deviation $0.06sD_0$, where $D_0 = \max(\max(0, \max_i t_i) - \min(0, \min_i t_i), 10^{-6})$ in the implementation. This scale is zero-referenced, rather than necessarily the current interval duration. Background activity inserts $\lfloor 0.12s \max(n, HW) \rfloor$ random-location events. Hot-pixel noise inserts $\lfloor 0.18s \max(n, 100) \rfloor$ events at $\max(1, \lfloor 1 + 4s \rfloor)$ sampled pixel locations. For both insertions, timestamps are sampled from the implementation's zero-inclusive range, ending at least one time unit after its lower endpoint;

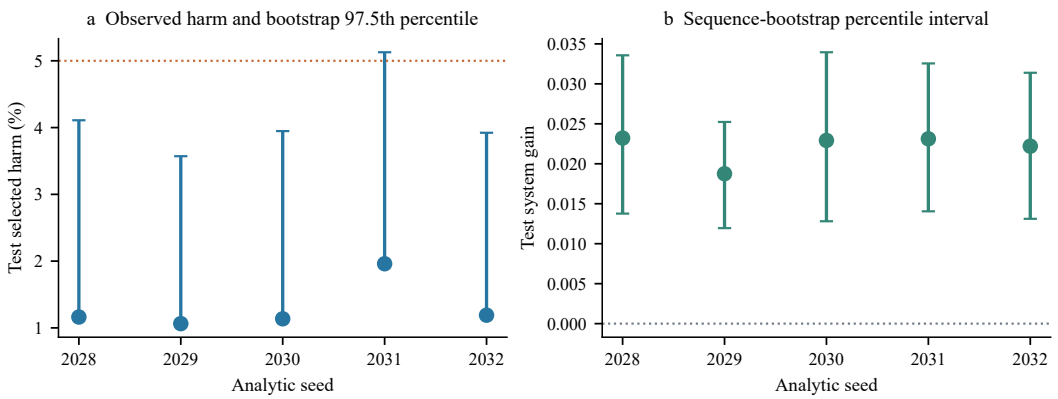

Figure 5: Sequence-resampling diagnostics for analytic joint policies. **a**, Observed development-test harm with an upper whisker to its bootstrap 97.5th percentile; the dotted line is the 5% reference. The lower end is the observation, not a lower confidence limit. **b**, Observed system gain and the bootstrap 2.5th–97.5th percentile interval. Each seed uses the same 12 development-test sequences and 10,000 whole-sequence resamples. These diagnostics do not replace the certificate's sampling assumptions.

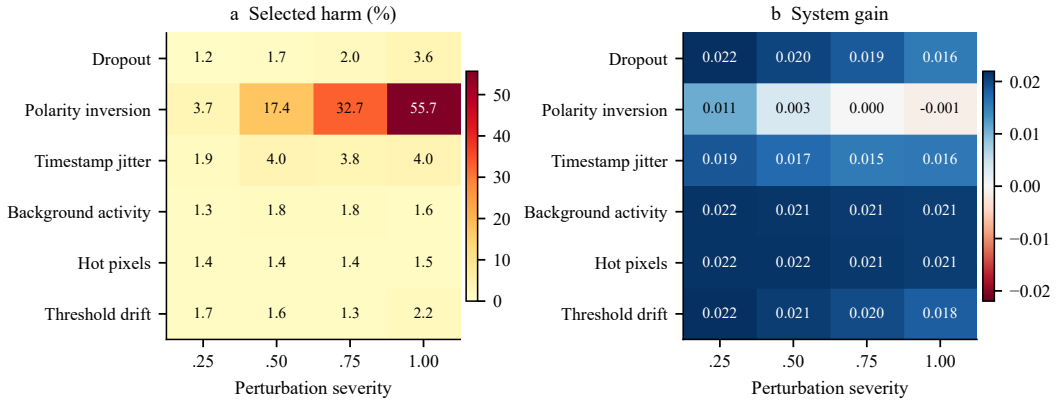

Figure 6: Frozen analytic joint policies under event perturbations. Cells average five seeds: **a**, selected harm in percent; **b**, system gain in normalized-MAE units. The same held-out sequences are reused across conditions. These are descriptive responses to the specified operators, without shifted-distribution certification.

later causal slicing may exclude inserted events outside the prediction interval. Thus severity indexes the implemented operators and should not be interpreted as a calibrated sensor-noise magnitude.

Figure 7 reports both residual CNN training losses for every learned seed. Training losses decrease, but this observation does not imply successful certification or establish cross-predictor robustness. Appendix F supplies analytic reconstructions replayed from the recorded configuration and checked against the original losses. No learned reconstructions are reported. Perceptual quality, inference speed, and computational cost are not evaluated by the selector-coverage metric.

F Qualitative examples and source-image provenance

Figure 8 expands the qualitative decisions shown in the main paper, while Figure 9 provides their full-frame source context. The qualitative export replays the analytic seed-2028 development-test run reused by the final experiment. It uses the original frozen scores and thresholds, without re-training, threshold adjustment, or additional certification. The selected policy passes at $\eta = 0.05$, with certificate bound 0.0419933. It does not pass the stricter eight-slot certificate. These examples

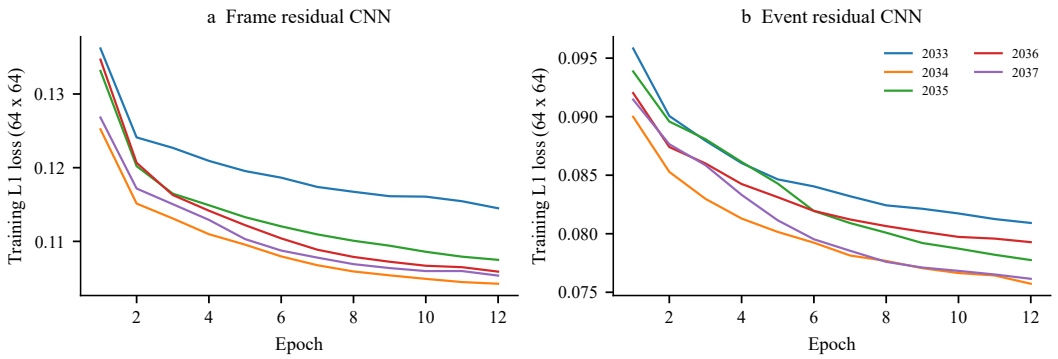

Figure 7: Lightweight residual predictor training. **a**, Frame-only and **b**, event-assisted branch training L1 losses at 64×64, for all five learned seeds over 12 epochs. These are training diagnostics, not held-out reconstruction scores.

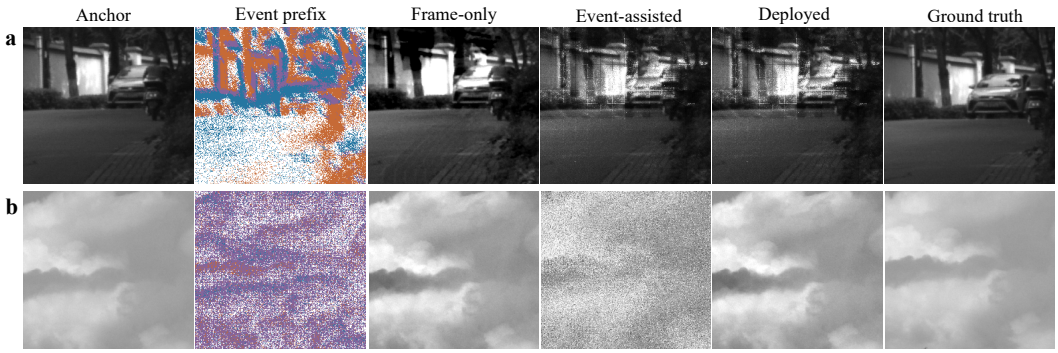

Figure 8: Real event-use decisions on the development-test sequences. **a**, Accepted beneficial prediction; **b**, rejected harmful prediction. Images show the middle suffix target; reported gain averages all three targets. Event-prefix colors indicate positive (blue), negative (orange), or both polarities (purple). Scores are not pointwise risk bounds. Both examples use analytic seed 2028's single-policy certificate, with $U_{\mathrm{CP}} = 0.0420$. Appendix F provides selection rules, full-frame context, and failure cases.

correspond to the single-policy certificate at $\eta = 0.05$, not the stricter eight-slot audit, and illustrate decisions on the evaluated crops.

The original development-test log contains 180 rounds, of which 86 are selected. These include 85 beneficial rounds and one harmful round. The 94 rejected rounds include 82 beneficial, seven harmful, and five small-loss rounds. Here, beneficial means $g > 0$, harmful means $g < -0.005$, and small loss means $-0.005 \le g \le 0$. Thus the policy can miss useful predictions and accept a harmful one despite passing its policy-level certificate. The observed selected harm is $1/86 = 1.16\%$, not zero.

To avoid choosing only visually favorable examples, the export covers every nonempty outcome class. Within each selected/rejected and outcome class, it retains up to two rounds nearest the class median gain. Ties are broken by round identifier; nine cases are retained in total. The main qualitative figure uses the closest-to-median accepted-benefit and rejected-harm cases. Figures 10 and 11 show the other seven. This outcome-stratified selection is illustrative, not an unbiased sample or a new performance test.

Each displayed prediction uses the middle of its three suffix targets. Reported round gains average all three targets, so they need not equal the displayed frame's loss difference. Both branches share the same spatial crop and intensity display range, $[0, 1]$. No per-image contrast enhancement, denoising, or perceptual post-processing is applied. Event-prefix images encode polarity occupancy before the decision, not the full event stream used by later predictions. Blue, orange, and purple pixels indicate

a city_0036_25fps (1120 × 660)



b valley_0028_10fps (1080 × 608)



Figure 9: Full-frame context for the accepted and rejected examples. Original anchor previews from the city and valley sequences are shown at their native spatial extent. Orange rectangles locate the 256×256 center crops used by the experiment. Predictions in Figure 3 are evaluated only within these crops.

positive, negative, and both polarities, respectively. The deployed column is an exact copy of the selected branch.

Independent recomputation from the exported floating-point arrays reproduces all nine recorded gains within 2.98×10^{-8}. The source package retains every suffix prediction, target, anchor, absolute-error array, and event-prefix count array. It also preserves the original uncropped HDF5 image arrays as NPY files, without grayscale conversion or normalization. Normalized full-frame PNGs are previews, not substitutes for these native arrays. Per-case metadata records the source HDF5 path and image key, frame index, timestamps, scores, decision, and replay discrepancy. The retained originals allow later panel rearrangement without modifying the underlying scientific evidence.

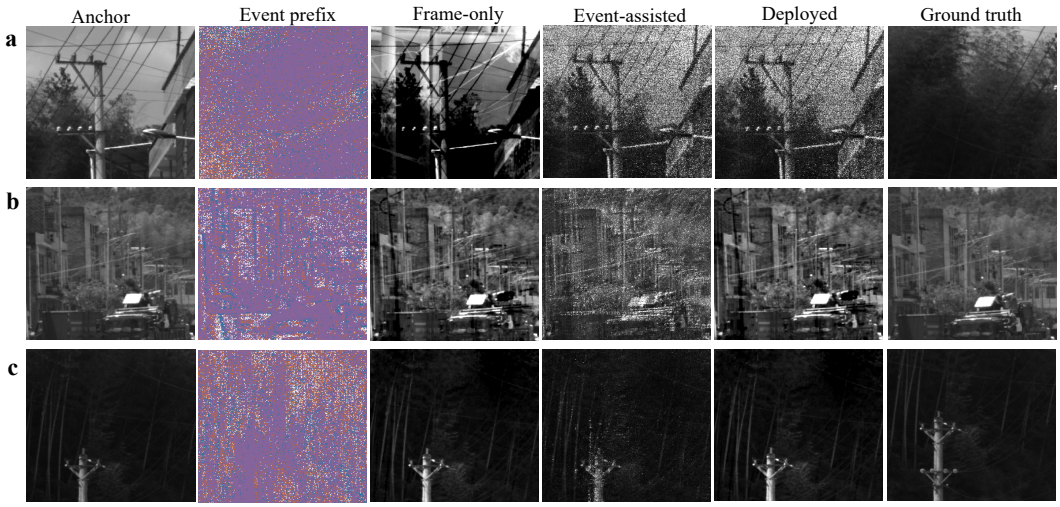

Figure 10: Errors and missed opportunities under a certified policy. **a**, The only accepted harmful round in analytic seed 2028. **b,c**, Rejected rounds with positive event-assisted gain. These cases demonstrate why policy certification does not imply correct decisions for every round. Displays, scores, and round-averaged gains follow Figure 3. MAE gain quantifies reconstruction error and does not measure perceptual quality.

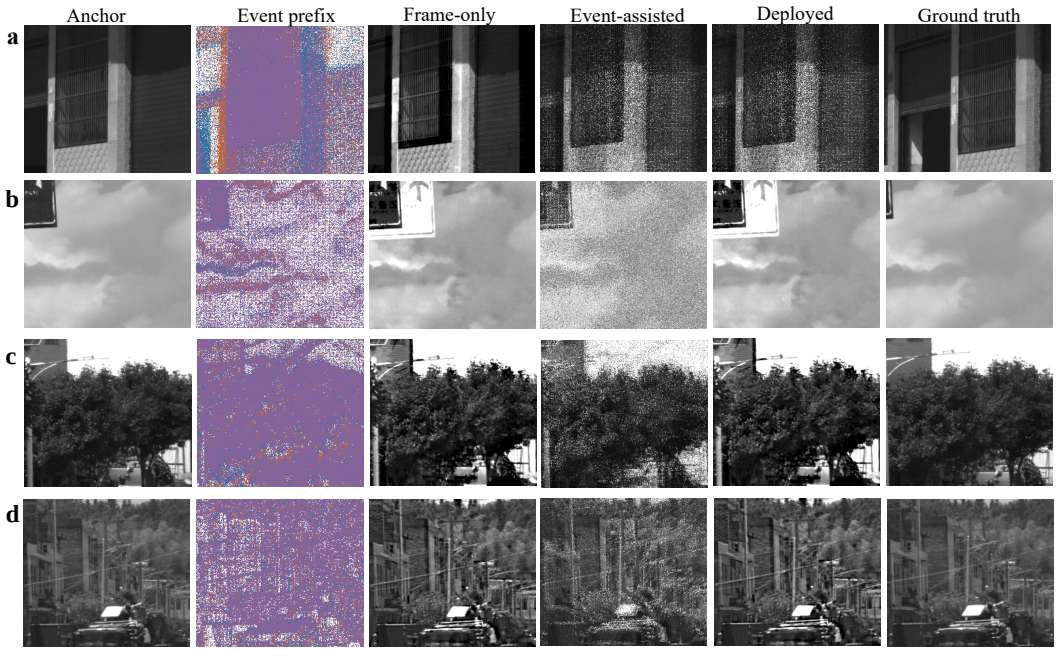

Figure 11: Additional outcome-stratified development-test examples. **a**, Accepted beneficial prediction; **b**, rejected harmful prediction; **c,d**, rejected predictions with losses below the meaningful-harm threshold. Together with Figure 3 and Figure 10, these panels show all nine exported cases. Scores determine selection; measured gains describe outcomes after the decision.