INSIDER: MINING LATENT FORENSIC KNOWLEDGE FOR AI-GENERATED VIDEO DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

As AI-generated videos become increasingly realistic, reliably distinguishing authentic from AI-generated content is essential for maintaining trust in visual information. However, existing learning-based detectors often rely on large amounts of labeled data for task-specific training, while latent forensic knowledge within pretrained video models remains underutilized. To address this limitation, we propose INSIDER, an AI-generated-video detection framework that uses a small set of labeled source videos to mine forensic knowledge within a frozen video encoder. Specifically, we first use layer selection to identify candidate layers based on the source-validation performance of layer-wise linear probes. Next, we combine decision leverage (DL) and supervised sensitivity (SS) to assess how pooled neuron responses influence fixed probe scores and how sensitive the same probe's supervised authenticity loss is to those responses for each source class. Forensic neurons are then selected adaptively using the knee points of layer-wise score curves. Finally, sparse readout concatenates the selected responses across layers into a compact forensic representation, on which we train a lightweight linear classifier for authenticity detection. With the detector fixed during target evaluation, INSIDER outperforms the evaluated detection methods on all four benchmarks covering 57 generator subsets, achieving a benchmark-averaged macro AUC of 92.53%. Further experiments show that INSIDER outperforms direct final-layer readouts with each of four video encoders. In separate tests, the source-trained detector remains robust to video compression and substantial resolution changes.

1 INTRODUCTION

As AI-generated videos become increasingly realistic in appearance and motion, reliable authenticity detection is essential for maintaining trust in visual information. The task (Vahdati et al., 2024; Kundu et al., 2025) extends beyond facial manipulation to full scenes, camera motion, and interactions among multiple subjects. Identifying reliable evidence in both visual appearance and cross-frame relationships is therefore important for video-level authenticity judgments.

Existing detectors learn discriminative representations or use explicitly defined forensic cues. Learning-based methods use mouth dynamics (Haliassos et al., 2021), self-supervised talking-face representations (Haliassos et al., 2022), and alternating spatial and temporal optimization (Wang et al., 2023b), with recent work (Kundu et al., 2025; Nguyen et al., 2025) extending detection to full-frame manipulations and subtle spatiotemporal inconsistencies. Complementary approaches include second-order temporal discrepancy analysis in D3 (Zheng et al., 2025), spatial-temporal likelihood scoring in STALL (Ben Hayun et al., 2026), and explanations of visible artifacts (Li et al., 2026) in Skyra. Task-specific training can require substantial labeled data, while explicit-cue methods focus on predefined discrepancies or statistics. Although pretrained video models offer rich representations, using them directly for classification does not reveal which internal components provide forensic evidence or how to select them with limited supervision for video authenticity detection.

Video understanding pretraining provides a basis for this investigation. VideoMAE (Tong et al., 2022) and VideoMAE V2 (Wang et al., 2023a) reconstruct masked spatiotemporal content using visible regions and neighboring frames. We hypothesize that the resulting representations of appearance and cross-frame relationships retain signals useful for distinguishing authentic from AI-generated videos. However, strong responses to scenes or motion do not necessarily indicate forensic rele-

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.

vance, and dense representations may mix discriminative information with irrelevant components. We therefore examine neurons within specific layers through their pooled video responses, using limited supervision to assess their role in authenticity readouts and select useful evidence. This provides a way to reuse latent forensic knowledge while keeping the video encoder frozen.

Building on this motivation, we propose INSIDER, an AI-generated-video detection framework that mines forensic knowledge within a frozen video encoder using a small set of labeled source videos (Figure 1). Layer selection identifies candidate layers based on the source-validation accuracy of layer-wise linear probes. Within each candidate layer, decision leverage (DL) measures the mean absolute change in the fixed probe's logit after zeroing a pooled neuron response. Under the same fixed probe, supervised sensitivity (SS) measures the authenticity loss's sensitivity to each pooled response, using class-averaged loss sensitivities from authentic and AI-generated videos. We multiply the two factors and use the knee point of each layer's score curve to adaptively select forensic neurons. Finally, sparse readout concatenates their pooled responses across layers into a compact forensic representation and trains a separate lightweight linear classifier for authenticity detection. The selected neuron set and detector remain fixed during target evaluation. Our main contributions are as follows:

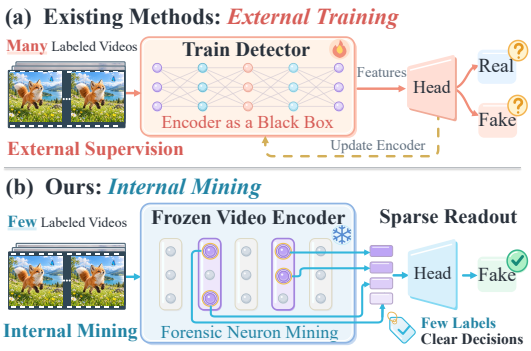

Figure 1: **Detection Paradigms.** (a) Conventional detectors use lots of labeled videos to train a black-box detection model. (b) INSIDER selects forensic neurons within a frozen encoder using few labeled videos.

- We propose INSIDER, an AI-generated-video detection framework that mines latent forensic knowledge within a frozen video encoder using limited labeled source videos, combining layer selection, forensic neuron mining, and sparse readout.
- We identify a neuron-scoring criterion that multiplies DL, the mean absolute logit change after removing a pooled response, by SS, the fixed probe's supervised loss sensitivity. Adaptive layer-wise selection produces a compact forensic representation.
- Across four benchmarks covering 57 generator subsets, INSIDER outperforms evaluated methods with a benchmark-averaged macro AUC of 92.53%. Evaluations demonstrate applicability across video encoders and robustness to compression and resolution changes.

2 Related Work

AI-Generated Video Detection. Early methods detect facial manipulation through mouth-motion analysis (Haliassos et al., 2021), self-supervised talking-face representations (Haliassos et al., 2022), and alternating spatial and temporal optimization (Wang et al., 2023b), with DeepfakeBench (Yan et al., 2023) and DF40 (Yan et al., 2024) broadening evaluation across manipulation families. Recent work (Kundu et al., 2025; Nguyen et al., 2025; Vahdati et al., 2024) extends detection to full-frame manipulations and scene-wide spatiotemporal inconsistencies. Complementary approaches use second-order temporal discrepancies in D3 (Zheng et al., 2025), spatial-temporal likelihoods in STALL (Ben Hayun et al., 2026), and explanations of visible artifacts (Li et al., 2026) in Skyra. Resources including GenVideo (Chen et al., 2026), EvalCrafter (Liu et al., 2024), VideoPhy-2 (Bansal et al., 2026), and VideoFeedback2 (He et al., 2025) further broaden generator and content coverage. Task-specific encoder adaptation can require substantial labeled data, while explicit-cue methods focus on selected discrepancies or statistics for video-level authenticity judgments.

Pretrained Representations and Neuron-Level Interpretability. VideoMAE (Tong et al., 2022) and VideoMAE V2 (Wang et al., 2023a) learn transferable representations of appearance and cross-frame relationships through masked video reconstruction. Linear probes (Alain & Bengio, 2017) assess the accessibility of task information in intermediate representations, while Network Dissection (Bau et al., 2017) and TCAV (Kim et al., 2018) relate internal units or directions to semantic concepts. Language-model studies identify neurons associated with linguistic properties (Dalvi

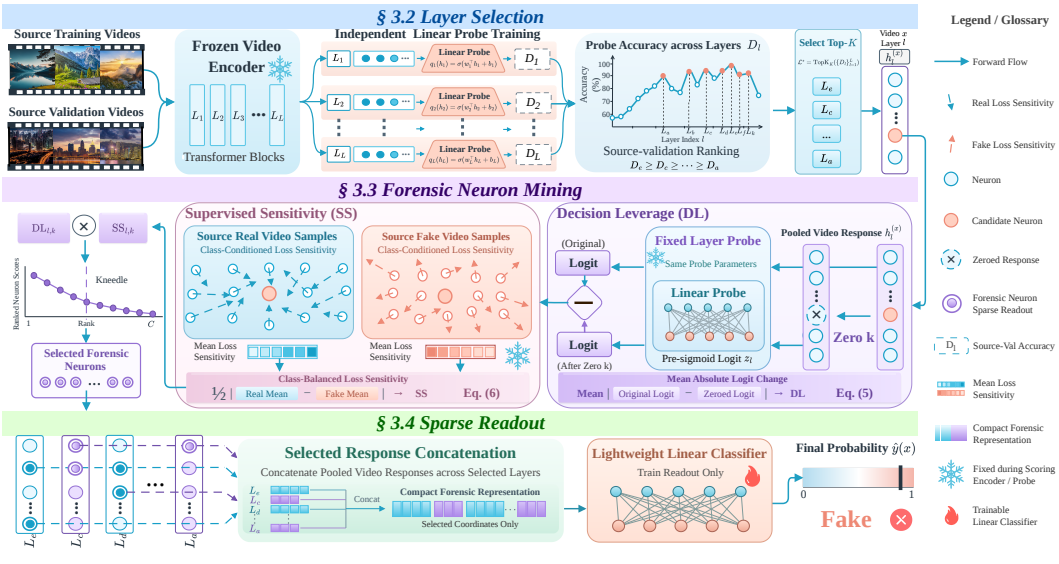

Figure 2: **Overview of INSIDER.** Source-validation probe accuracy selects layers. DL and SS score pooled source-training responses through fixed layer probes. Adaptive neuron selection supplies a compact representation for a separately trained linear readout.

et al., 2019; Durrani et al., 2020) and factual knowledge (Dai et al., 2022), with interventions (Meng et al., 2022) connecting internal components to model behavior. However, semantic relevance alone does not establish a neuron's usefulness for video authenticity detection.

3 Proposed Method

3.1 Overview

Figure 2 presents INSIDER's layer selection, forensic neuron mining, and sparse readout. Labels $y = 0, 1$ denote authentic and AI-generated videos. Source-training data $\mathcal{D}_s^{\text{tr}}$ support probe fitting, neuron scoring, and classifier training; disjoint source-validation data $\mathcal{D}_s^{\text{val}}$ support layer selection. The two scoring factors share cached source-training responses and the same fixed probe, so neuron scores refer to the same authenticity readout within each candidate encoder layer.

Video x's layer-l tensor $\mathbf{F}_l^{(x)} \in \mathbb{R}^{T\times P\times C}$ has T temporal positions, P spatial positions, and C channels, indexed by t, p, and k. Channel k has the pooled video response

$$h_{l,k}^{(x)} = \frac{1}{TP}\sum_{t=1}^{T}\sum_{p=1}^{P} F_l^{(x)}(t,p,k). \tag{1}$$

Each layer-channel pair is a candidate neuron. The layer's pooled response vector $\mathbf{h}_l^{(x)} \in \mathbb{R}^C$ retains appearance and cross-frame context encoded before pooling. A linear probe maps this vector to a video-level logit for AI-generated content. This lets us assess authenticity information at each depth:

$$z_l^{(x)} = \mathbf{w}_l^\top \mathbf{h}_l^{(x)} + b_l. \tag{2}$$

Probe q_l uses coefficients $\mathbf{w}_l$ and intercept b_l, fitted on $\mathcal{D}_s^{\text{tr}}$ and fixed for neuron scoring. It predicts the AI-generated-content probability $p_l^{(x)} = \sigma(z_l^{(x)})$, where σ denotes the sigmoid function.

3.2 Layer Selection

Across the encoder's L layers, independent probes share an architecture and training protocol. Probe q_l has source-validation accuracy, indicating how accessible authenticity information is at depth l:

$$D_l = \text{ACC}_{\text{src-val}}(q_l). \tag{3}$$

With a layer budget K, we select the subset with the largest sum of source-validation accuracies:

$$\mathcal{L}^\star = \arg \max_{\substack{\mathcal{L}\subseteq\{1,\dots,L\} \\ |\mathcal{L}|\le K}} \sum_{l\in\mathcal{L}} D_l. \tag{4}$$

Here, $\mathcal{L}$ and $\mathcal{L}^\star$ are candidate and selected layer sets. Top-K ranking allows noncontiguous depths. We set $K = 9$ to balance detection performance and the size of the forensic representation, with layer-count sensitivity reported in Appendix C. The selected set stays fixed for scoring and evaluation. Indices are one-based in equations and zero-based in the implementation and tables for the same encoder blocks. This makes the selected depths directly traceable in the reported ablations.

3.3 Forensic Neuron Mining

Decision Leverage. Layer selection identifies informative representations, but a strong response within those layers may still primarily describe scene content or motion. DL examines whether a candidate neuron's pooled response affects the fixed probe's video-level decision. We set that coordinate to zero while keeping the other coordinates and the probe unchanged. Let $\mathbf{h}^{(x)}_{l,k\leftarrow 0}$ denote the resulting vector. DL measures the average absolute logit change:

$$\mathrm{DL}_{l,k} = \mathbb{E}_x \left| z_l\left(\mathbf{h}^{(x)}_l\right) - z_l\left(\mathbf{h}^{(x)}_{l,k\leftarrow 0}\right) \right|. \tag{5}$$

Removal occurs after pooling, leaving the upstream encoder response tensor unchanged. DL first averages absolute logit changes within each batch of source-training videos and then gives equal weight to the batch means. A larger DL identifies a pooled response with greater average influence on the fixed probe's score when that response is removed from the pooled representation.

Supervised Sensitivity. SS measures the sensitivity of the fixed probe's supervised authenticity loss to a candidate neuron's pooled response. It uses the same source-training responses and fixed layer probe as DL. Let $\mathcal{L}_{\mathrm{BCE}}(x, y)$ denote the per-video binary cross-entropy loss evaluated with the observed label. Let $\mathcal{R} = \mathcal{D}^{\mathrm{tr}}_{s,0}$ and $\mathcal{F} = \mathcal{D}^{\mathrm{tr}}_{s,1}$ denote the authentic and AI-generated subsets of the scoring data. We measure loss sensitivity separately within each class and define SS as

$$\mathrm{SS}_{l,k} = \frac{1}{2} \left| \mathbb{E}_{x\in\mathcal{R}} \frac{\partial \mathcal{L}_{\mathrm{BCE}}}{\partial h^{(x)}_{l,k}} - \mathbb{E}_{x\in\mathcal{F}} \frac{\partial \mathcal{L}_{\mathrm{BCE}}}{\partial h^{(x)}_{l,k}} \right|. \tag{6}$$

Each class mean uses all source-training videos of its class. For the fixed linear probe with binary cross-entropy, the nonzero class means have opposite signs, so SS equally weights their magnitudes. Within a layer, SS is proportional to the absolute probe coefficient, scaled by a shared factor from class-averaged prediction residuals. Coupling SS with DL combines supervised coefficient weighting and response-dependent logit influence.

Coupled Selection. The two factors define the coupled neuron score $S^{\text{DL-SS}}_{l,k}$:

$$S^{\text{DL-SS}}_{l,k} = \mathrm{DL}_{l,k} \cdot \mathrm{SS}_{l,k}\,. \tag{7}$$

We rank the C candidate neurons within each retained layer in native pooled coordinates. Kneedle (Satopää et al., 2011) on the normalized descending score curve sets the retained count K_l adaptively under predefined size constraints. Let π_l map rank j to its neuron index in descending score order. The retained coordinates in layer l form the set

$$\mathcal{A}_l = \{\pi_l(j) \mid 1 \le j \le K_l\}\,. \tag{8}$$

Layer-wise selection preserves evidence from different depths without comparing absolute score scales. The resulting index sets define a common sparse representation for training the final classifier and evaluating the same detector on videos from different target generators.

3.4 Sparse Readout

For compact video authenticity detection, we concatenate the selected response subvectors $\mathbf{h}^{(x)}_l[\mathcal{A}_l]$ across retained layers and train a separate lightweight linear classifier:

$$\widehat{y}(x) = \sigma\left(\mathbf{w}_f^\top \,\mathrm{Concat}_{l\in\mathcal{L}^\star}\left(\mathbf{h}^{(x)}_l[\mathcal{A}_l]\right) + b_f\right). \tag{9}$$

Table 1: **Cross-Benchmark Comparison.** Scores are percentages. Superscripts give signed differences from the best baseline in each column (percentage points); reference values have no superscript. Bold and dashed underlining mark the best and second-best results, respectively.

Detection Level	Method	GenVideo			EvalCrafter			VideoPhy-2			VideoFeedback2		
		ACC	AP	AUC	ACC	AP	AUC	ACC	AP	AUC	ACC	AP	AUC
Image	UnivFD	53.60$^{-19.25}$	71.49$^{-6.91}$	73.67$^{-8.38}$	57.91$^{-14.66}$	78.19$^{-1.62}$	79.94$^{-5.20}$	52.73$^{-8.61}$	61.25$^{-2.80}$	61.29$^{-7.36}$	55.72$^{-7.39}$	71.32$^{-9.96}$	73.10$^{-5.48}$
	NPR	52.43$^{-20.42}$	64.96$^{-13.44}$	60.09$^{-21.96}$	63.01$^{-9.56}$	79.81	76.56$^{-8.58}$	54.06$^{-7.28}$	63.45$^{-0.60}$	60.26$^{-8.39}$	59.94$^{-3.17}$	79.77$^{-1.51}$	78.58
	RIGID	48.74$^{-24.11}$	44.79$^{-33.61}$	42.64$^{-39.41}$	48.32$^{-24.25}$	46.66$^{-33.15}$	44.92$^{-40.22}$	48.10$^{-13.24}$	42.96$^{-21.09}$	39.36$^{-29.29}$	48.86$^{-14.25}$	49.83$^{-31.45}$	50.28$^{-28.30}$
Video	D3	–	74.95$^{-3.45}$	72.11$^{-9.94}$	–	76.55$^{-3.26}$	74.69$^{-10.45}$	–	63.54$^{-0.51}$	58.87$^{-9.78}$	–	81.28	77.05$^{-1.53}$
	STALL	–	78.40	78.10$^{-3.95}$	–	79.70$^{-0.11}$	79.20$^{-5.94}$	–	55.56$^{-8.49}$	50.97$^{-17.68}$	–	71.86$^{-9.42}$	73.20$^{-5.38}$
	CoCoDetect	56.11$^{-16.74}$	58.46$^{-19.94}$	59.31$^{-22.74}$	57.85$^{-14.72}$	59.56$^{-20.25}$	61.78$^{-23.36}$	47.46$^{-13.88}$	45.62$^{-18.43}$	45.87$^{-22.78}$	52.62$^{-10.49}$	52.46$^{-28.82}$	53.37$^{-25.21}$
	VideoMAEv2-L	62.99$^{-9.86}$	64.45$^{-13.95}$	67.27$^{-14.78}$	64.13$^{-8.44}$	64.46$^{-15.35}$	67.65$^{-17.49}$	54.17$^{-7.17}$	54.20$^{-9.85}$	54.78$^{-13.87}$	61.90$^{-1.21}$	62.13$^{-19.15}$	64.63$^{-13.95}$
	DeMamba	62.32$^{-10.53}$	68.87$^{-9.53}$	70.65$^{-11.40}$	64.37$^{-8.20}$	68.62$^{-11.19}$	70.55$^{-14.59}$	51.15$^{-10.19}$	50.17$^{-13.88}$	53.89$^{-14.76}$	60.57$^{-2.54}$	66.25$^{-15.03}$	65.98$^{-12.60}$
	NSG-VD	62.69$^{-10.16}$	67.50$^{-10.90}$	65.38$^{-16.67}$	64.84$^{-7.73}$	70.08$^{-9.73}$	68.43$^{-16.71}$	59.42$^{-1.92}$	62.06$^{-1.99}$	62.05$^{-6.60}$	60.61$^{-2.50}$	65.09$^{-16.19}$	64.02$^{-14.56}$
	DeCoF	70.78$^{-2.07}$	73.70$^{-4.70}$	77.25$^{-4.80}$	71.14$^{-1.43}$	75.50$^{-4.31}$	78.34$^{-6.80}$	61.34	64.05	68.65	53.30$^{-9.81}$	74.48$^{-6.80}$	75.11$^{-3.47}$
	ReStraV	72.85	77.36$^{-1.04}$	82.05	72.57	78.50$^{-1.31}$	85.14	53.93$^{-7.41}$	54.65$^{-9.40}$	57.62$^{-11.03}$	63.11	70.00$^{-11.28}$	76.74$^{-1.84}$
	Ours	**84.57**$^{+11.72}$	**93.27**$^{+14.87}$	**92.47**$^{+10.42}$	**88.83**$^{+16.26}$	**96.22**$^{+16.41}$	**96.17**$^{+11.03}$	**75.08**$^{+13.74}$	**87.69**$^{+23.64}$	**89.29**$^{+20.64}$	**82.01**$^{+18.90}$	**91.75**$^{+10.47}$	**92.20**$^{+13.62}$

Here, $\widehat{y}(x)$ is the AI-generated-content probability; $\mathbf{w}_f$ and b_f are the final classifier's coefficients and intercept (f: final readout). They are fitted on source-training videos with the encoder, layer indices, and neuron indices fixed. The training objective $\mathcal{L}_{\text{det}}$ is

$$\mathcal{L}_{\text{det}} = -\frac{1}{|\mathcal{D}_s^{\text{tr}}|} \sum_{(x,y)\in\mathcal{D}_s^{\text{tr}}} \left[y \log \widehat{y}(x) + (1-y) \log(1-\widehat{y}(x)) \right]. \tag{10}$$

Layer probes guide neuron selection, while the final classifier learns to combine the selected responses. The resulting detector remains fixed across target benchmarks. For another encoder, the same source-based procedure fits new probes, selects neurons, and trains a readout. Sparse selection reduces the readout dimension while preserving the appearance and temporal context already encoded in the pooled responses. The compact readout still draws on multiple depths of the encoder.

4 Experiment and Analysis

We compare detection performance across four benchmarks and 57 generator subsets. We then analyze layer selection, coupled scoring, and adaptive neuron selection, examine selected responses and authenticity decisions, and evaluate encoder choice and robustness to input degradation.

4.1 Experimental Setup

Datasets. We evaluate 57 generator subsets from four benchmarks: ten from GenVideo (Chen et al., 2026), fourteen from EvalCrafter (Liu et al., 2024), seven from VideoPhy-2 (Bansal et al., 2026), and twenty-six from VideoFeedback2 (He et al., 2025), spanning diverse video generators.

Metrics. We report video-level accuracy (Sokolova & Lapalme, 2009) (ACC), average precision (Su et al., 2015) (AP), and area under the receiver operating characteristic curve (Fawcett, 2006) (AUC). Scores (%) are unweighted macro averages over the same generator subsets; dashes identify unreported values and are excluded from numerical comparisons.

Implementation Details. The public source split (Ma et al., 2026) contains 200 Kinetics-400 (Kay et al., 2017) and 200 Open-Sora (Zheng et al., 2024) videos; a disjoint, class-balanced 400-video validation split ranks layers. We use frozen VideoMAE-Large (Tong et al., 2022); Appendix B gives the video inputs, optimization protocol, and adaptive neuron-selection settings.

Compared Methods. Table 1 compares image-level UnivFD (Ojha et al., 2023), NPR (Tan et al., 2024), and RIGID (He et al., 2024) with video-level D3 (Zheng et al., 2025) (cosine scoring), STALL (Ben Hayun et al., 2026), VideoMAE V2-Large (Wang et al., 2023a), DeMamba (Chen et al., 2026), DeCoF (Ma et al., 2025), CoCoDetect (Feng et al., 2026), ReStraV (Internò et al., 2025), and NSG-VD (Zhang et al., 2025). INSIDER builds its readout from selected forensic neurons across frozen layers, using limited labeled source videos for video authenticity detection.

4.2 Comparison with State-of-the-art Methods

Table 1 compares four AI-generated-video benchmarks. INSIDER achieves the highest listed ACC, AP, and AUC on each benchmark, with cross-benchmark means of 82.62%, 92.23%, and 92.53%, respectively. On GenVideo, its margins exceed ten percentage points on all three metrics. On EvalCrafter, AP and AUC exceed 96%, with double-digit margins over the strongest listed baselines. VideoPhy-2 shows the largest ranking advantage: AP and AUC improve by more than 20 percentage points. INSIDER also leads the macro comparison over the twenty-six VideoFeedback2 subsets, where different baselines rank second across metrics. The same source-trained detector and forensic neuron set are used across all 57 generator subsets, showing that the selected evidence supports authenticity decisions and score ranking across diverse generators and benchmarks.

4.3 Contributions of the Selection Stages

We analyze layer selection, forensic neuron mining, and temporal-input controls on GenVideo. These comparisons examine which encoder depths and neurons support detection and how the detector uses multi-frame input. The single-layer baseline uses layer 13 (zero-based), which has the highest source-validation probe accuracy among the individual layers.

Table 2: **Component Ablation on GenVideo.** Bold marks the best scores.

Neuron Set	Count	ACC	AP	AUC
All 24 Layers	No Layer Selection			
All Neurons	24,576	80.01	89.14	89.33
Forensic Neurons	1,005	78.94	89.45	89.34
Top-9 Layers	Candidate Layers			
All Neurons	9,216	83.01	90.60	90.77
Forensic Neurons	**402**	**84.57**	**93.27**	**92.47**

Contributions of Layer Selection and Forensic Neuron Mining. Table 2 separates the effects of layer selection and forensic neuron mining. Selecting layers improves the dense all-layer readout; mining neurons within those layers yields a further gain. The joint configuration achieves the highest ACC, AP, and AUC, while neither component alone matches it. Within the selected layers, reducing 9,216 neurons to 402 raises the three metrics by 1.56, 2.67, and 1.70 percentage points, respectively. INSIDER therefore improves detection with only 4.36% of the candidate neurons. Layer selection identifies useful depths, and neuron mining concentrates the frozen encoder's evidence into a compact readout that improves detection without updating the pretrained encoder. The sparse readout leads every metric.

Role of Layer Selection. Table 3 compares probe-selected layers with fixed depth ranges, random selection, the best single layer, and all 24 layers. Fixed ranges cover early through late regions as percentages of encoder depth. INSIDER and the single-layer baseline select layers using source-validation probe accuracy before target evaluation. INSIDER achieves the highest ACC, AP, and AUC. Both INSIDER and random selection use nine layers, so their difference reflects which depths are selected. INSIDER's ACC exceeds the best single-layer baseline by 10.47 percentage points and the all-layer configuration by 5.63 points. Combining informative layers thus improves on the strongest individual layer, whereas using every layer gives lower scores. The middle-to-late range is the strongest fixed region but remains below INSIDER on all three metrics. This supports source-guided selection across encoder depths to establish the candidate pool for forensic neuron mining.

Table 3: **Layer Selection Ablation on GenVideo.** Relative depth ranges may overlap; Random and INSIDER each use nine layers. Bold and dashed underlining mark the best and second-best scores.

	Alternative Layer Subsets					Reference Settings		
Setting	Single	Early	Middle	Middle–Late	Late	Random	All 24	**Ours**
Depth Range	–	0–40%	40–75%	50–85%	75–100%	–	0–100%	–
ACC	74.10	58.97	71.05	81.07	79.89	70.19	78.94	**84.57**
AP	86.96	62.01	79.28	90.06	88.82	80.25	89.45	**93.27**
AUC	84.91	61.71	76.12	89.21	88.29	80.68	89.34	**92.47**

4.4 Coupled Scoring and Adaptive Neuron Selection

Role of DL and SS Coupling. Table 4 compares removing DL, SS, or both with the default layers and per-layer neuron counts fixed. The control without either factor ranks neurons by response magnitude. DL measures the average absolute logit change after zeroing a pooled response in the fixed layer probe. SS measures supervised loss sensitivity to the response under the same probe, using separate class means. Their product couples this sensitivity with logit influence to rank neurons. INSIDER achieves the highest ACC, AP, and AUC, whereas removing either factor lowers all three metrics. Coupled scoring improves ACC, AP, and AUC by 8.73, 8.78, and 8.02 percentage points over the strongest control for each metric. These gains support coupled scoring for selecting forensic responses from the same layers at a fixed representation size.

Table 4: **Factor Ablation on GenVideo.** Fixed counts; superscripts give percentage-point differences from the best controls (underlined).

Factors		GenVideo		
DL	SS	ACC	AP	AUC
×	×	74.16$^{-1.68}$	84.49	84.45
×	✓	56.03$^{-19.81}$	58.12$^{-26.37}$	57.06$^{-27.39}$
✓	×	75.84	81.88$^{-2.61}$	82.96$^{-1.49}$
✓	✓	**84.57**$^{+8.73}$	**93.27**$^{+8.78}$	**92.47**$^{+8.02}$

Coupling Gain. Figure 3 expands the macro comparison in Table 4 to individual generators. Each point compares INSIDER with the strongest of response magnitude, DL only, and SS only for that generator and metric. All 30 comparisons favor INSIDER, so the positive macro gains are accompanied by improvements across all ten evaluated generators. Sora and ModelScope have the largest AP gains; MoonValley has the smallest, which remains positive. Selecting the reference separately for each comparison tests INSIDER against the strongest measured control rather than one control chosen for the whole benchmark. Positive ACC gains accompany positive AP and AUC gains, covering classification at the fixed threshold and ranking of authentic and AI-generated videos. This generator-level comparison complements Table 4: the table reports macro scores at fixed counts, whereas the figure checks each generator against its own strongest control. The gains extend across all evaluated generator subsets and all three metrics at the same fixed neuron budget.

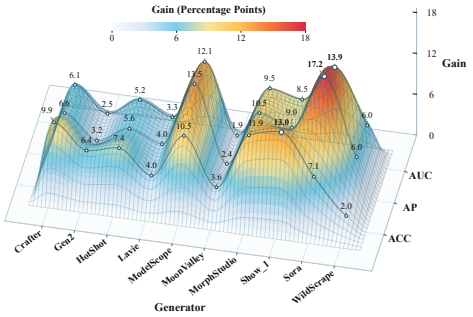

Figure 3: **Coupling Gains on GenVideo.** Gains over the best control per generator and metric. Points: observations; surface: interpolation.

Influence of Adaptive Neuron Selection. Figure 4 compares adaptive neuron selection with fixed per-layer neuron counts and an all-neuron readout on the default selected layers. Adaptive selection sets a separate count for each layer from its score curve. The adaptive setting achieves the highest ACC, AP, and AUC while retaining only 402 forensic neurons, or 4.36% of the candidate pool. Smaller fixed-count sets yield lower scores, and retaining more than 402 neurons provides no further improvement in the tested configurations. Even the strongest fixed-count setting, with 1,440 neurons, falls below adaptive selection on all three metrics. Using all 9,216 candidate neurons also underperforms adaptive selection. Relative to this dense readout, adaptive selection improves ACC, AP, and AUC by 1.56, 2.67, and 1.70 percentage points, respectively. With the candidate layers held fixed, these results support a compact sparse readout that improves detection without retaining the full candidate representation.

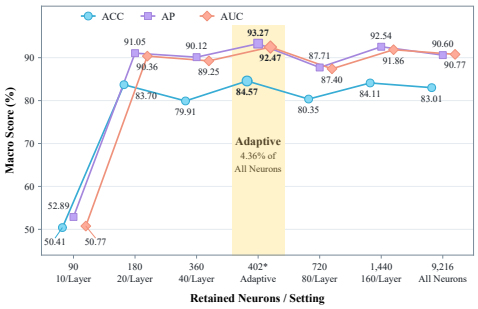

Figure 4: **Retained-Neuron Sensitivity on GenVideo.** Fixed counts, adaptive selection, and all neurons use the same candidate layers.

4.5 Selected Responses and Authenticity Decisions

Feature Distribution. Figure 5 compares No Selection and INSIDER on a balanced GenVideo subset. Each uses its own source-trained classifier: pooled final-layer features for No Selection and selected neurons for INSIDER. The axes show AI-generated probability and the leading source-fitted residual component orthogonal to the classifier direction. Target labels identify classes and centroids but fit neither axis. No Selection overlaps near the 0.5 threshold. INSIDER places authentic videos at lower probabilities and the dominant AI-generated density above the threshold, giving clearer class separation in both contours and probability marginals.

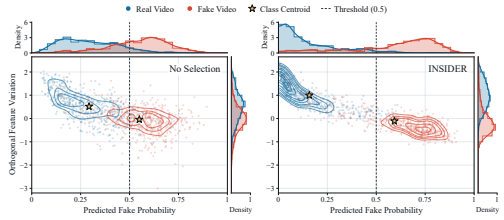

Figure 5: **Feature Distributions on GenVideo.** Stars mark class centroids; dashed lines mark the 0.5 threshold. Both methods share axes and density settings.

Temporal Responses and Decision Intervention. Figure 6 ranks neurons by their pooled contributions toward the correct class. The ten largest positive contributors form the intervention set, with R1–R6 showing its six strongest members. T1–T8 are consecutive two-frame tubelet bins; row-wise min-max scaling shows when each neuron responds most strongly. The adjoining panels compare predictions before and after zeroing the ten selected pooled responses through the fixed readout. The AI-generated probability rises from 0.340 to 0.981 for the authentic example and falls from 0.647 to 0.035 for the AI-generated example. Both cross the 0.5 boundary away from the correct class, connecting temporally varying responses to evidence used by the classifier.

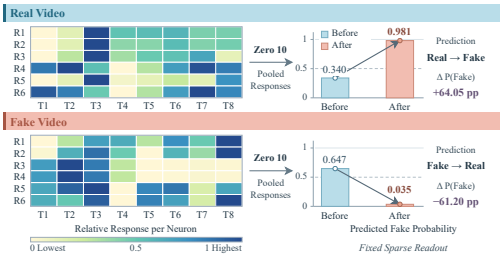

Figure 6: **Temporal Responses and Decision Shifts.** Six ranked profiles accompany shifts after zeroing ten selected pooled responses. Top: authentic; bottom: AI-generated.

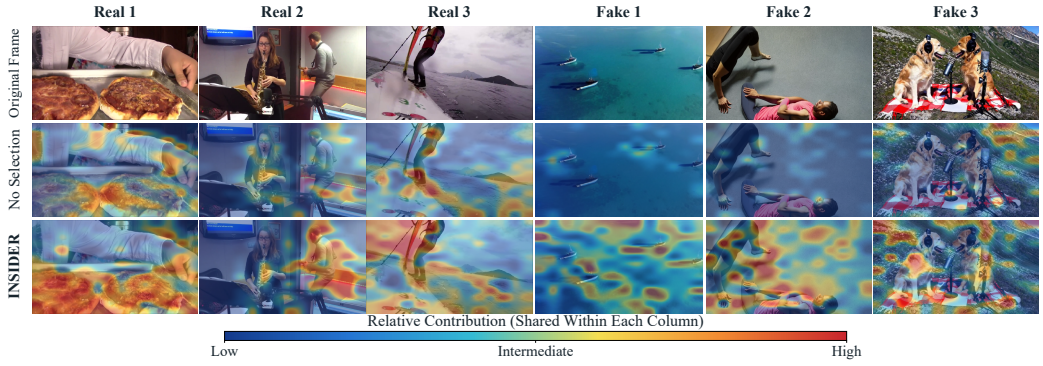

Figure 7: **Spatial Decision Contributions.** Matched color scales compare No Selection and INSIDER; warmer colors indicate stronger class support. See Appendix H.

Spatial Decision Contributions. Figure 7 compares positive spatial contributions to the indicated class under the two source-trained readouts. No Selection uses pooled final-layer features, whereas INSIDER uses selected forensic neurons. Both methods share the frozen encoder, input frames, within-column normalization, and overlay opacity. Warmer colors indicate stronger positive contributions to the class shown. Paired overlays reveal differences in class support at corresponding spatial locations. In the illustrated AI-generated examples, INSIDER provides stronger correct-class support in regions that receive weaker contributions from No Selection. With the encoder and inputs held fixed, the comparison shows how neuron selection and the learned readout change the spatial pattern of class support extracted from the same pretrained representation.

4.6 Robustness and Sensitivity Analysis

Generalization across Video Encoders. Figure 8 compares INSIDER with a direct final-layer head on VideoMAE-Large (Tong et al., 2022), VideoMAE V2-Large (Wang et al., 2023a), Video Swin-B (Liu et al., 2022), and Hiera-B+ (Ryali et al., 2023). The same source-based mining procedure fits separate probes, selects neurons, and trains readouts for each encoder. Each architecture supplies its own response coordinates and selected subset; the common element is the selection procedure applied to its pretrained representation. INSIDER improves all three metrics in every comparison. Video Swin-B shows the largest ACC, AP, and AUC gains of 30.84, 38.06, and 35.60 percentage points. It also achieves the highest ACC, while VideoMAE-Large has the highest AP and AUC. Each comparison keeps the pretrained encoder frozen for both readouts, showing that neuron mining improves detection across these four pretrained video encoders.

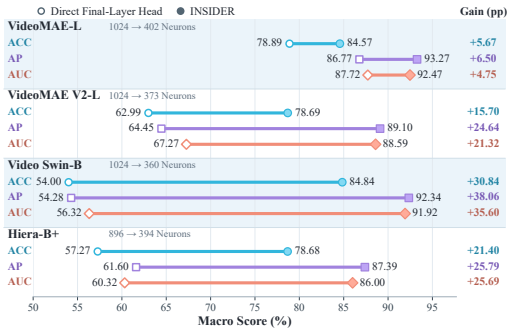

Figure 8: **Encoder Ablation on GenVideo.** Open/filled: direct head/INSIDER; gains in percentage points.

Robustness to Input Degradation. Figure 9 evaluates INSIDER under compression, resizing, and Gaussian blur on GenVideo. All transformed inputs are evaluated with the same selected neurons and classifier, so each condition tests the sensitivity of the same source-trained detector. The outer bars report absolute scores, while the inner colors show changes relative to the first setting of each transformation. These views show both the size of the performance change and the detection performance that remains. Compression has little effect on score ranking: AP and AUC vary by less than 0.4 percentage points within each tested JPEG or H.264 sweep. Spatial resizing also has a limited effect, with AP and AUC remaining above 92% even at 56×56 resolution. The detector continues to distinguish authentic from AI-generated videos effectively despite the reduction in spatial detail. Gaussian blur leads to a larger change, but AP and AUC remain close to 87% even at $\sigma = 3$. INSIDER therefore retains useful ranking performance under the strongest tested blur. The same source-trained detector maintains stable performance under compression and substantial resolution reduction and still distinguishes authentic from AI-generated videos after strong smoothing.

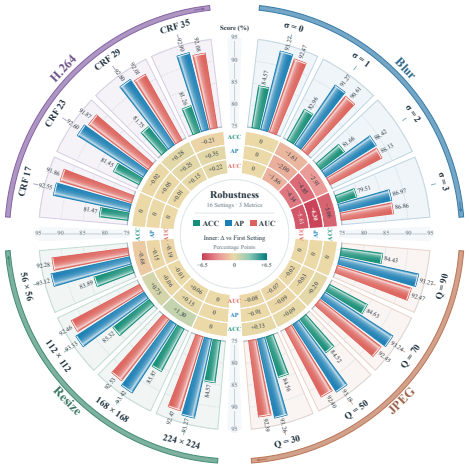

Figure 9: **Robustness on GenVideo.** Bars show ACC, AP and AUC from 75%; inner colors show percentage-point changes from the first setting of each degradation.

5 Conclusion

We introduced INSIDER to mine forensic knowledge from a frozen video encoder with limited source labels. Layer selection identified informative depths, and coupled DL and SS selected neurons for a sparse readout. With the detector fixed during target evaluation, INSIDER led macro ACC, AP, and AUC across four benchmarks and 57 generator subsets, reaching 92.53% benchmark-averaged AUC. Ablations, temporal-input controls, and encoder and degradation comparisons supported the selection stages, multi-frame evidence, and applicability across encoders. The study remained limited to supervised video-level binary detection with a pretrained encoder. Future work will examine label-efficient adaptation to new generators and spatiotemporal forgery localization.

AI USE STATEMENT

Generative AI tools were used to polish the grammar and wording of the manuscript. The authors reviewed the final text and take responsibility for its content.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *ICLR 2017 Workshop*, 2017. URL https://arxiv.org/abs/1610.01644.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A Challenging Action-Centric Physical Commonsense Evaluation in Video Generation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 118456–118470, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c02f6a1d5c55e16db50d339dad905b4d-Abstract-Conference.html.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network Dissection: Quantifying Interpretability of Deep Visual Representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6541–6549, July 2017. doi: 10.1109/CVPR.2017.354. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Bau_Network_Dissection_Quantifying_CVPR_2017_paper.html.

Omer Ben Hayun, Roy Betser, Meir Yossef Levi, Levi Kassel, and Guy Gilboa. Training-free Detection of Generated Videos via Spatial-Temporal Likelihoods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16299–16310, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Hayun_Training-free_Detection_of_Generated_Videos_via_Spatial-Temporal_Likelihoods_CVPR_2026_paper.html.

Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, and Huaxiong Li. DeMamba: AI-generated video detection on million-scale GenVideo benchmark. *Science China Information Sciences*, 69(6):162103, June 2026. doi: 10.1007/s11432-024-4894-0. URL https://link.springer.com/article/10.1007/s11432-024-4894-0.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge Neurons in Pretrained Transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502, 2022. doi: 10.18653/v1/2022.acl-long.581. URL https://aclanthology.org/2022.acl-long.581/.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. What Is One Grain of Sand in the Desert? Analyzing Individual Neurons in Deep NLP Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6309–6317, 2019. doi: 10.1609/aaai.v33i01.33016309. URL https://arxiv.org/abs/1812.09355.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. Analyzing Individual Neurons in Pre-trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4865–4880, 2020. doi: 10.18653/v1/2020.emnlp-main.395. URL https://aclanthology.org/2020.emnlp-main.395/.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. doi: 10.1016/j.patrec.2005.10.010. URL https://doi.org/10.1016/j.patrec.2005.10.010.

Huidong Feng, Wentao Chen, Jie Chen, Xinqi Cai, Ruolong Ma, Yinglin Zheng, Yuxin Lin, and Ming Zeng. CoCoVideo: The High-Quality Commercial-Model-Based Contrastive Benchmark for AI-Generated Video Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11346–11356, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Feng_CoCoVideo_The_High-Quality_Commercial-Model-Based_Contrastive_Benchmark_for_AI-Generated_Video_Detection_CVPR_2026_paper.html.

Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips Don't Lie: A Generalisable and Robust Approach to Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5039–5049, 2021. doi: 10.1109/CVPR46437.2021.00500. URL https://openaccess.thecvf.com/content/CVPR2021/html/Haliassos_Lips_Dont_Lie_A_Generalisable_and_Robust_Approach_To_Face_CVPR_2021_paper.html.

Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Leveraging Real Talking Faces via Self-Supervision for Robust Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14950–14962, 2022. doi: 10.1109/CVPR52688.2022.01453. URL https://openaccess.thecvf.com/content/CVPR2022/html/Haliassos_Leveraging_Real_Talking_Faces_via_Self-Supervision_for_Robust_Forgery_Detection_CVPR_2022_paper.html.

Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, Keming Wu, Benjamin Schneider, Quy Duc Do, Zhuofeng Li, Yiming Jia, Yuxuan Zhang, Guo Cheng, Haozhe Wang, Wangchunshu Zhou, Qunshu Lin, Yuanxing Zhang, Ge Zhang, Wenhao Huang, and Wenhu Chen. VideoScore2: Think before You Score in Generative Video Evaluation. *arXiv preprint arXiv:2509.22799*, 2025. URL https://arxiv.org/abs/2509.22799.

Zhiyuan He, Pin-Yu Chen, and Tsung-Yi Ho. RIGID: A Training-free and Model-Agnostic Framework for Robust AI-Generated Image Detection. *arXiv preprint arXiv:2405.20112*, 2024. doi: 10.48550/arXiv.2405.20112. URL https://arxiv.org/abs/2405.20112.

Christian Internò, Robert Geirhos, Markus Olhofer, Sunny Liu, Barbara Hammer, and David Klindt. AI-Generated Video Detection via Perceptual Straightening. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 20672–20705. Curran Associates, Inc., 2025. doi: 10.52202/085713-0697. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/1d9a43752c2819e03967c5c1b708169c-Abstract-Conference.html.

Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics Human Action Video Dataset. *arXiv preprint arXiv:1705.06950*, 2017. URL https://arxiv.org/abs/1705.06950.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677. PMLR, 2018. URL https://proceedings.mlr.press/v80/kim18d.html.

Rohit Kundu, Hao Xiong, Vishal Mohanty, Athula Balachandran, and Amit K. Roy-Chowdhury. Towards a Universal Synthetic Video Detector: From Face or Background Manipulations to Fully AI-Generated Content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28050–28060, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Kundu_Towards_a_Universal_Synthetic_Video_Detector_From_Face_or_Background_CVPR_2025_paper.html.

Yifei Li, Wenzhao Zheng, Yanran Zhang, Runze Sun, Yu Zheng, Lei Chen, Jie Zhou, and Jiwen Lu. Skyra: AI-Generated Video Detection via Grounded Artifact Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4482–4493, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_Skyra_AI-Generated_Video_Detection_via_Grounded_Artifact_Reasoning_CVPR_2026_paper.html.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: Benchmarking and Evaluating Large Video Generation Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22139–22149, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_EvalCrafter_Benchmarking_and_Evaluating_Large_Video_Generation_Models_CVPR_2024_paper.html.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video Swin Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3202–3211, 2022. doi: 10.1109/CVPR52688.2022.00320. URL https://openaccess.thecvf.com/content/CVPR2022/html/Liu_Video_Swin_Transformer_CVPR_2022_paper.html.

Long Ma, Zhiyuan Yan, Qinglang Guo, Yong Liao, Haiyang Yu, and Pengyuan Zhou. Detecting AI-Generated Video via Frame Consistency. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2025. doi: 10.1109/ICME59968.2025.11210049. URL https://arxiv.org/abs/2402.02085.

Long Ma, Zihao Xue, Yan Wang, Zhiyuan Yan, Jin Xu, Xiaorui Jiang, Haiyang Yu, Yong Liao, and Zhen Bi. Your One-Stop Solution for AI-Generated Video Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4458–4470, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Ma_Your_One-Stop_Solution_for_AI-Generated_Video_Detection_CVPR_2026_paper.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Dat Nguyen, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Vulnerability-Aware Spatio-Temporal Learning for Generalizable Deepfake Video Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10786–10796, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Nguyen_Vulnerability-Aware_Spatio-Temporal_Learning_for_Generalizable_Deepfake_Video_Detection_ICCV_2025_paper.html.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards Universal Fake Image Detectors That Generalize Across Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24480–24489, 2023. doi: 10.1109/CVPR52729.2023.02345. URL https://openaccess.thecvf.com/content/CVPR2023/html/Ojha_Towards_Universal_Fake_Image_Detectors_That_Generalize_Across_Generative_Models_CVPR_2023_paper.html.

Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29441–29454. PMLR, 2023. URL https://proceedings.mlr.press/v202/ryali23a.html.

Ville Satopää, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*, pp. 166–171, 2011. doi: 10.1109/ICDCSW.2011.20. URL https://www.cs.williams.edu/~jeannie/papers/kneedle-simplex11.pdf.

Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009. doi: 10.1016/j.ipm.2009.03.002. URL https://doi.org/10.1016/j.ipm.2009.03.002.

Wanhua Su, Yan Yuan, and Mu Zhu. A Relationship between the Average Precision and the Area Under the ROC Curve. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, pp. 349–352, 2015. doi: 10.1145/2808194.2809481. URL https://doi.org/10.1145/2808194.2809481.

Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the Up-Sampling Operations in CNN-based Generative Network for Generalizable Deepfake Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28130–28139, 2024. doi: 10.1109/CVPR52733.2024.02657. URL https://openaccess.thecvf.com/content/CVPR2024/html/Tan_Rethinking_the_Up-Sampling_Operations_in_CNN-based_Generative_Network_for_Generalizable_CVPR_2024_paper.html.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 10078–10093. Curran Associates, Inc., 2022. doi: 10.52202/068431-0732. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/416f9cb3276121c42eebb86352a4354a-Paper-Conference.pdf.

Danial Samadi Vahdati, Tai D. Nguyen, Aref Azizpour, and Matthew C. Stamm. Beyond Deepfake Images: Detecting AI-Generated Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4397–4408, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024W/WMF/html/Vahdati_Beyond_Deepfake_Images_Detecting_AI-Generated_Videos_CVPRW_2024_paper.html.

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 (CVPR)*, pp. 14549–14560, June 2023a. URL https://openaccess.thecvf.com/content/CVPR2023/html/Wang_VideoMAE_V2_Scaling_Video_Masked_Autoencoders_With_Dual_Masking_CVPR_2023_paper.html.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. AltFreezing for More General Video Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4129–4138, 2023b. doi: 10.1109/CVPR52729.2023.00402. URL https://openaccess.thecvf.com/content/CVPR2023/html/Wang_AltFreezing_for_More_General_Video_Face_Forgery_Detection_CVPR_2023_paper.html.

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection. In *Advances in Neural Information Processing Systems*, volume 36, pp. 4534–4565, 2023. doi: 10.52202/075280-0201. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0e735e4b4f07de483cbe250130992726-Abstract-Datasets_and_Benchmarks.html.

Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, and Li Yuan. DF40: Toward Next-Generation Deepfake Detection. In *Advances in Neural Information Processing Systems*, volume 37, pp. 29387–29434, 2024. doi: 10.52202/079017-0925. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/34239f60eca7ce9bee5280aaf81362d8-Abstract-Datasets_and_Benchmarks_Track.html.

Shuhai Zhang, Zihao Lian, Jiahao Yang, Daiyuan Li, Guoxuan Pang, Feng Liu, Bo Han, Shutao Li, and Mingkui Tan. Physics-Driven Spatiotemporal Modeling for AI-Generated Video Detection. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 174683–174730. Curran Associates, Inc., 2025. doi: 10.52202/085713-5810. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/ff7c9c90030f1eb3cbf5c81b6fbd9a05-Abstract-Conference.html.

Chende Zheng, Ruiqi Suo, Chenhao Lin, Zhengyu Zhao, Le Yang, Shuai Liu, Minghui Yang, Cong Wang, and Chao Shen. D3: Training-Free AI-Generated Video Detection Using Second-Order Features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12852–12862, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zheng_D3_Training-Free_AI-Generated_Video_Detection_Using_Second-Order_Features_ICCV_2025_paper.html.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-Sora: Democratizing Efficient Video Production for All. *arXiv preprint arXiv:2412.20404*, 2024. URL https://arxiv.org/abs/2412.20404.

APPENDIX

The appendix is organized as follows. Appendix A summarizes the notation. Appendix B describes the implementation and compact readout. Appendix C presents controls for layer count, temporal input, and source size. Appendix D examines dependence on selected neuron responses through coordinate interventions. Appendices E and F report macro and generator-level results across four random seeds. Appendix G provides complete comparisons across generator subsets, and Appendix H discusses the interpretation and design choices of INSIDER. The sequence moves from implementation and controlled comparisons to repeated runs and the full generator breakdown, allowing the summarized results to be traced to their underlying evaluations.

A CORE NOTATION

Table A.1 collects the symbols needed to follow the method and its analyses. Layer indices are one-based in equations and zero-based in experiments. The definitions distinguish quantities fitted on source-training videos from those evaluated on source-validation videos, as well as layer probes from the final sparse readout. This separation clarifies how each stage uses the frozen encoder and its pooled responses.

Table A.1: Notation used in INSIDER.

Symbol	Definition
x	Input video.
y	Class label: 0 for authentic and 1 for AI-generated content.
$\mathcal{D}_s^{\mathrm{tr}}$	Source-training split for probes, neuron scoring, and readout.
$\mathcal{D}_s^{\mathrm{val}}$	Disjoint source-validation split used for layer ranking.
$\mathcal{R} = \mathcal{D}_{s,0}^{\mathrm{tr}}$	Authentic source-training videos.
$\mathcal{F} = \mathcal{D}_{s,1}^{\mathrm{tr}}$	AI-generated source-training videos.
L	Number of encoder layers.
l	Encoder-layer index.
T	Number of temporal token positions.
P	Number of spatial token positions.
C	Number of channels at the layer under consideration.
$t,\ p,\ k$	Temporal-token, spatial-token, and channel indices.
$\mathbf{F}_l^{(x)}$	Layer-l response tensor of shape $T \times P \times C$.
$F_l^{(x)}(t,p,k)$	Scalar entry of the layer-l response tensor.
$h_{l,k}^{(x)}$	Temporally and spatially pooled response of channel k.
$\mathbf{h}_l^{(x)}$	Vector of the C pooled channel responses in layer l.
$\mathbf{w}_l,\ b_l$	Coefficient vector and intercept of the fixed layer-l probe.
q_l	Source-trained linear probe for layer l.
$p_l^{(x)}$	Probe probability that video x is AI-generated.
$z_l^{(x)},\ z_l(\cdot)$	AI-generated-content logit and its layer-l logit function.
$\sigma(\cdot)$	Sigmoid function mapping a logit to a probability.
D_l	Accuracy of probe q_l on source-validation videos.
K	Budget on the number of retained encoder layers.
$\mathcal{L}$	Candidate encoder-layer set.
$\mathcal{L}^\star$	Selected top-K layer set.
$\mathbf{h}_{l,k\leftarrow 0}^{(x)}$	Pooled response vector with only coordinate k set to zero.
$\mathrm{DL}_{l,k}$	Mean absolute probe-logit change after zeroing a response.
$\mathbb{E}_x$	Source-training average with equal-weight scoring-batch means.
$\mathcal{L}_{\mathrm{BCE}}(x,y)$	Per-video binary cross-entropy of the fixed layer probe.
$\partial\mathcal{L}_{\mathrm{BCE}}/\partial h_{l,k}^{(x)}$	Sensitivity of probe loss to a pooled channel response.
$\mathrm{SS}_{l,k}$	Half the absolute difference in class-mean loss sensitivities.
$\mathbb{E}_{x\in\mathcal{R}},\ \mathbb{E}_{x\in\mathcal{F}}$	Averages over all source-training videos in each class.

Table A.1: Notation used in INSIDER (continued).

Symbol	Definition
$S_{l,k}^{\text{DL-SS}}$	Coupled neuron score: $\text{DL}_{l,k}\,\text{SS}_{l,k}$.
K_l	Adaptive retained-neuron count in layer l; $K_l = \lvert\mathcal{A}_l\rvert$.
$\pi_l(j)$	Channel index at rank j in descending layer-l scores.
$\mathcal{A}_l$	Selected forensic-neuron indices in layer l.
$\mathbf{h}_l^{(x)}[\mathcal{A}_l]$	Subvector of selected pooled responses in layer l.
$\mathbf{w}_f,\ b_f$	Coefficient vector and intercept of the final linear classifier.
$\widehat{y}(x)$	Final probability that video x is AI-generated.
$\mathcal{L}_{\text{det}}$	Mean binary cross-entropy for final-readout training.
Δ	Change relative to the stated reference, in percentage points.
σ (blur)	Gaussian-blur standard deviation; distinct from $\sigma(\cdot)$.

B Implementation and Compact Readout

The default encoder was frozen VideoMAE-Large, with 24 blocks and 1,024 channels per block. Layer selection retained $\{7, 12, 13, 14, 17, 18, 19, 20, 21\}$. The source-training and source-validation splits each contained 400 class-balanced videos and were disjoint. Both DL and SS used source-training responses through the same fitted layer probe; source-validation accuracy supplied the layer-ranking scores. Each video supplied 16 RGB frames resized to 224×224 and normalized with ImageNet channel statistics. The disjoint splits assign different roles to probe fitting and layer ranking while keeping both within the source data. After the layers and neuron indices are selected, target videos pass through the same frozen encoder and fixed readout. Table 1 compares methods across the same four benchmarks and 57 generator subsets. Source-trained readouts are fitted on source data, while training-free and checkpoint-based baselines retain their prescribed training and preprocessing procedures. This provides a common evaluation and aggregation protocol without treating the methods' different training requirements as identical.

Independent linear probes were trained for 10 epochs using Adam with a learning rate of 10^{-3} and a batch size of 256. The final linear classifier was trained for 10 epochs using Adam with a learning rate of 10^{-4} and a batch size of 32. During sparse-readout training, only the classifier weights and bias were updated, and the final-epoch model was used for evaluation. An AI-generated probability greater than 0.5 was considered a positive prediction. We treated GenVideo as the probe set, and the reported GenVideo results were obtained on this probe set by averaging scores equally across its ten fixed generator subsets. We selected $K = 9$ based on the layer-count comparison on GenVideo (Figure C.1), balancing detection performance and readout size. This value was then fixed before evaluation on EvalCrafter, VideoPhy-2, and VideoFeedback2. Probe fitting and sparse-readout training were performed as separate optimization steps, and the final classifier received the selected coordinates without updating the encoder.

For decision leverage (DL), absolute logit changes were averaged within 256-video scoring batches, followed by an equal-weight average across batches. For supervised sensitivity (SS), each class mean used all source-training videos of that class. Kneedle operated on each normalized descending score curve with the convex/decreasing setting. Retained counts were constrained to $[40, 512]$. Layers 12 and 22 tied at the nine-layer cutoff with 97.25% probe accuracy; the default set retained layer 12. Normalizing and ranking scores within each layer lets the adaptive rule assign different counts at different depths without comparing raw scores across layers.

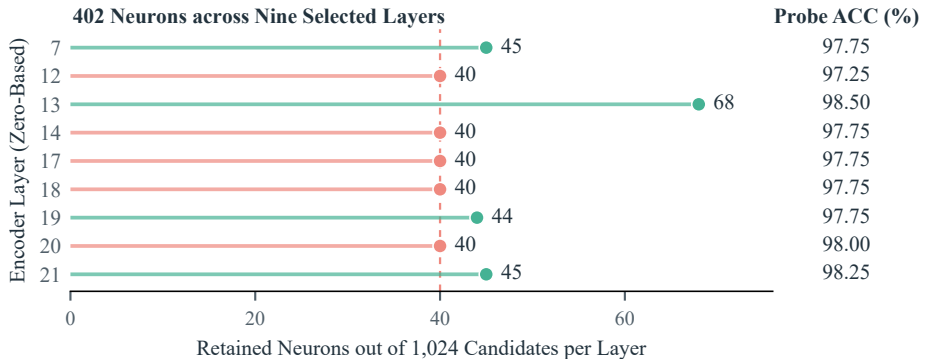

Figure B.1: **Default Layer Configuration.** Probe ACC is measured on source-validation videos; each retained layer has 1,024 candidate neurons. Coral markers identify the five layers at the lower bound of 40 neurons. The nine layers contribute 402 neurons in total.

The retained coordinates span several encoder depths, with five layers at the 40-neuron lower bound and the other four contributing 202 neurons. This distribution yields a 402-dimensional readout from the 9,216 candidate channels in the selected layers. Figure B.1 shows this allocation directly, with channels drawn from several noncontiguous depths. The final 402 features represent 4.36% of the 9,216 channels available in the selected layers, making the reduction in readout dimension explicit.

C Layer-Count and Temporal-Input Controls

Sensitivity to the Retained Layer Count. Figure C.1 compares seven to twelve probe-ranked layers on GenVideo. Seven and eight layers score below the nine-layer setting, while eleven and twelve layers also reduce ACC, AP, and AUC. Ten layers reach a marginally higher peak but increase the readout from 402 to 442 neurons for gains below 0.5 percentage points on each metric. We therefore chose nine layers as the default: it captures the improvement over smaller budgets and remains close to the peak with a more compact representation.

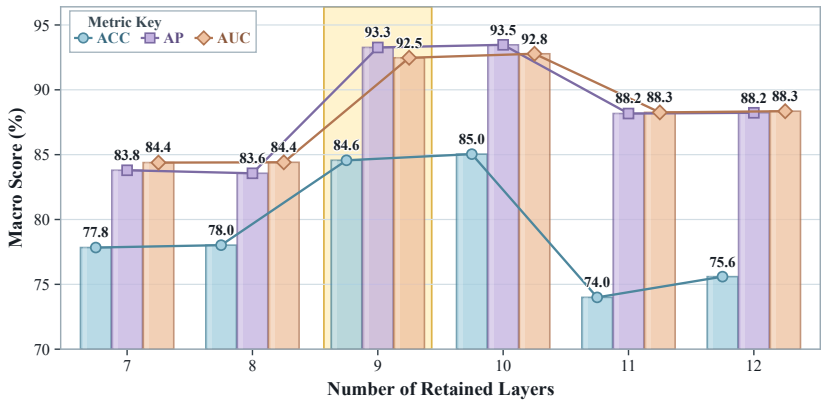

Figure C.1: **Sensitivity to the Number of Selected Layers on GenVideo.** Layers are ranked by linear-probe accuracy. The highlighted group marks the nine-layer configuration used by INSIDER.

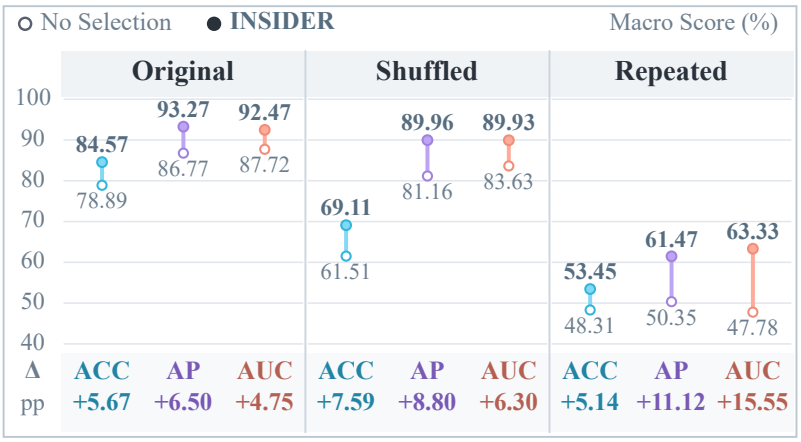

Figure C.2: **Temporal-Input Control on GenVideo.** Encoder and readouts are fixed. Shuffled permutes 16 sampled frames; Repeated uses the ninth frame 16 times. Open/filled circles show No Selection/INSIDER macro scores (%); Δ denotes gain in percentage points.

Role of Temporal Input. Figure C.2 compares Original, Shuffled, and Repeated inputs with the encoder and each method's source-trained classifier fixed. No Selection uses pooled final-layer features; INSIDER uses selected forensic neurons. Both perform best on Original. Shuffling preserves the sampled frames but changes their order; repetition removes frame variation and causes a substantially larger decline. This contrast supports the value of distinct frames for authenticity detection. INSIDER improves ACC, AP, and AUC in every condition, with ACC gains exceeding five percentage points. Under Shuffled input, INSIDER records 69.11% ACC, 89.96% AP, and 89.93% AUC; the ranking measures remain close to their Original values of 93.27% and 92.47%. Repeated input yields 61.47% AP and 63.33% AUC, a larger drop after distinct frames are replaced with copies of one frame. Even in this condition, INSIDER exceeds No Selection by 11.12 AP and 15.55 AUC percentage points. Thus, multi-frame input benefits the selected representation, and neuron selection improves detection even when frame order or diversity changes.

C.1 Sensitivity to Source Size

Sensitivity to Source Size. Figure C.3 compares six labeled source-set sizes from 100 to 1,600 videos on GenVideo. The 400-video setting has the highest ACC, AP, and AUC in this sweep, including 92.47% AUC. The 600-, 800-, and 1,600-video settings reach 86.92–87.90% AUC and retain 398, 371, and 382 neurons, respectively, rather than the 402 retained at 400 videos. Thus, source size also changes the selected representation. A shifted Kneedle score-curve knee may alter which responses enter the readout, offering a plausible explanation for the non-monotonic trend. The strong 400-video result supports a compact detector with modest source annotation; adapting the selection rule to larger sets merits further study.

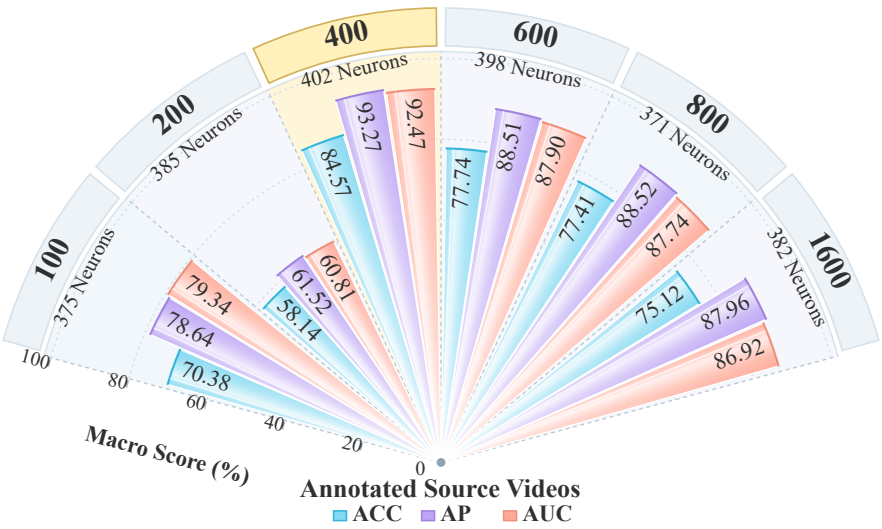

Figure C.3: **Source-Size Sensitivity of INSIDER on GenVideo.** Bars show macro ACC, AP, and AUC; inner labels report the number of forensic neurons. Gold highlights the 400-video setting.

D Dependence on Selected Neuron Responses

We tested paired coordinate interventions on all 400 source-validation videos, fixing the encoder, classifier, and 402 selected coordinates. For each video, the ten pooled coordinates with highest signed contributions to its labeled class were jointly zeroed. The control removed ten distinct coordinates sampled uniformly without replacement, using 200 masks per video. Both removals are evaluated on exactly the same videos, limiting variation caused by the composition of the evaluation set. Keeping the learned parameters fixed isolates the effect of changing pooled responses from any effect of retraining.

The correct-class margin was the AI-generated logit for AI-generated videos and its negative for authentic videos. We measured its decrease after removal, averaging random-mask effects within each video to obtain 400 paired observations. Confidence intervals used 20,000 video-level percentile bootstrap resamples with model and masks fixed, preserving class balance for the combined estimate. This signed definition puts both classes on the same scale: a positive decrease always means weaker support for the labeled class. Averaging random masks before pairing leaves one comparison per video instead of treating its 200 masks as separate observations.

Table D.1: **Accuracy (%) before and after Removing Ten Pooled Coordinates.** Random averages the 200 masks per video.

Group	n	Original	Targeted	Random
Authentic	200	87.00	7.00	86.3625
AI-Generated	200	98.50	0.00	98.0075
All	400	92.75	3.50	92.185

Targeted removal lowers accuracy sharply for both authentic and AI-generated videos, whereas random removal leaves each class close to its original accuracy. Authentic-video accuracy changes from 87.00% to 7.00% under targeted removal, while AI-generated-video accuracy changes from 98.50% to 0.00%. Random removal keeps the corresponding accuracies at 86.3625% and 98.0075%, showing the importance of which coordinates are removed.

Table D.2: **Decrease in Correct-Class Logit Margin.** Paired differences are targeted minus random removal; intervals are 95% bootstrap CIs.

Group	Targeted Drop	Random Drop	Paired Difference	95% CI
Authentic	3.396	0.037	3.359	[3.293, 3.424]
AI-Generated	3.018	0.032	2.986	[2.941, 3.030]
All	3.207	0.035	3.172	[3.133, 3.212]

The targeted-minus-random margin decrease is positive in both classes, and both class-specific 95% intervals remain above zero. The paired difference is 3.359 for authentic videos and 2.986 for AI-generated videos, so the effect appears in each class. Both bootstrap intervals stay above zero after resampling at the video level, matching the direction of the accuracy changes.

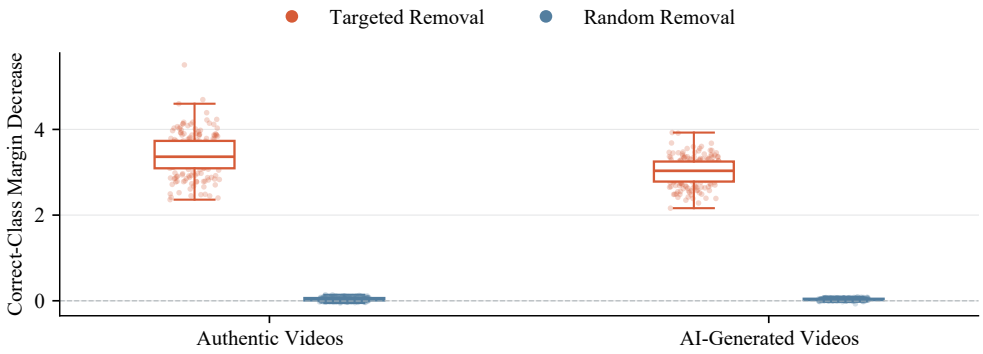

Figure D.1: **Effects of Removing Selected Coordinates.** Each point represents one video; random-removal points average its 200 masks. Boxes show medians and interquartile ranges; whiskers extend to observations within 1.5 interquartile ranges.

Targeted removal reduced the correct-class margin by 3.207 versus 0.035 for random removal. Accuracy fell from 92.75% to 3.50%; random removal retained 92.185%. This intervention tests the fixed classifier's dependence on selected responses, complementing the main-text ablation of scoring rules at matched retained counts. Figure D.1 displays the video-level changes behind these averages, with targeted and random removals shown separately. The population view complements the two tables by showing how the selected coordinates affect individual predictions under one fixed detector.

D.1 Fixed-Probe and Readout Identities

The fitted layer probe in Eq. 2 stays fixed while its pooled coordinates are scored. Let $w_{l,k}$ be entry k of $\mathbf{w}_l$. Zeroing $h_{l,k}^{(x)}$ changes the linear logit by $w_{l,k}h_{l,k}^{(x)}$, so Eq. 5 becomes

$$\mathrm{DL}_{l,k} = \mathbb{E}_x \left| w_{l,k} h_{l,k}^{(x)} \right| = |w_{l,k}| \, \mathbb{E}_x \left| h_{l,k}^{(x)} \right| . \tag{11}$$

The expectation retains the equal-weight scoring-batch average used in the main-text definition. For binary cross-entropy, $\partial \mathcal{L}_{\mathrm{BCE}}(x, y)/\partial h_{l,k}^{(x)} = w_{l,k}(p_l^{(x)} - y)$. Since authentic videos have $y = 0$ and AI-generated videos have $y = 1$, Eq. 6 simplifies to

$$\mathrm{SS}_{l,k} = \frac{|w_{l,k}|}{2} \left(1 + \mathbb{E}_{x \in \mathcal{R}} p_l^{(x)} - \mathbb{E}_{x \in \mathcal{F}} p_l^{(x)} \right) . \tag{12}$$

Substituting these identities into Eq. 7 gives

$$S_{l,k}^{\text{DL-SS}} = \frac{|w_{l,k}|^2}{2} \, \mathbb{E}_x \left| h_{l,k}^{(x)} \right| \left(1 + \mathbb{E}_{x \in \mathcal{R}} p_l^{(x)} - \mathbb{E}_{x \in \mathcal{F}} p_l^{(x)} \right) . \tag{13}$$

The parenthesized factor is shared by all channels in layer l. Thus, SS alone ranks channels by $|w_{l,k}|$ within that layer, while the coupled score also accounts for the average pooled response magnitude. These identities describe the layer probes used for selection; the final readout is trained separately on the retained coordinates.

Margin Change Under Coordinate Removal. Let $\mathbf{v}^{(x)} = \mathrm{Concat}_{l\in\mathcal{L}^\star}\big(\mathbf{h}_l^{(x)}[\mathcal{A}_l]\big)$ be the selected vector in Eq. 9. For $y \in \{0,1\}$, its correct-class logit margin is $m_y(x) = (2y-1)\big(\mathbf{w}_f^\top \mathbf{v}^{(x)} + b_f\big)$. If $\mathcal{I}$ indexes pooled coordinates set to zero and $m_y(x;\mathcal{I}\leftarrow 0)$ is the resulting margin, the fixed linear readout gives

$$m_y(x) - m_y(x;\mathcal{I}\leftarrow 0) = (2y-1)\sum_{i\in\mathcal{I}} w_{f,i} v_i^{(x)}. \tag{14}$$

Each summand is the signed contribution used to rank the targeted coordinates. The same identity applies to a random mask; only the removed index set changes.

E Results across Random Seeds

We repeated neuron mining and sparse readout training with four random seeds, keeping the encoder, source splits, nine-layer configuration, and training settings fixed. Each run was evaluated on all ten GenVideo generators, with equal weight assigned to each generator. Because neuron mining is rerun, the four results capture variation in the selected feature sets as well as the final classifier weights. Keeping the source partitions and ten evaluation subsets unchanged makes the macro scores directly comparable.

Table E.1 reports all four runs and the mean and sample standard deviation of each macro metric. Figure E.1 shows every macro score, and Appendix F presents the corresponding generator-level results. The table places selected-neuron count beside each run's three metrics rather than reducing the experiment to one average. The figure adds a visual comparison of the four macro results while retaining the same generator-balanced aggregation.

Table E.1: **Macro Results across Four Random Seeds on GenVideo.** Scores are percentages; SD is the sample standard deviation across runs.

Neurons	ACC	AP	AUC
402	84.57	93.27	92.47
394	83.23	91.09	90.04
371	84.85	92.83	92.42
360	84.79	91.91	91.78
Mean $\pm$ SD	84.36 ± 0.76	92.27 ± 0.97	91.68 ± 1.13

The 371-neuron readout records the highest ACC at 84.85%, while the 402-neuron readout gives the highest AP and AUC at 93.27% and 92.47%. Every run retains fewer than 5% of the 9,216 candidate channels, and the ranking scores remain high across these compact readouts.

Across the four runs, mean ACC, AP, and AUC were 84.36%, 92.27%, and 91.68%, with sample standard deviations of 0.76, 0.97, and 1.13 percentage points, respectively. ACC remained between 83.23% and 84.85%, while AP and AUC stayed above 91% and 90%. Detection performance remained consistent when neuron mining and readout training were repeated under the same configuration. The selected sets contain 360 to 402 neurons, yet every run maintains AP above 91% and AUC above 90%. High ranking scores across these different selected-set sizes show that the observed result is not tied to one exact neuron count.

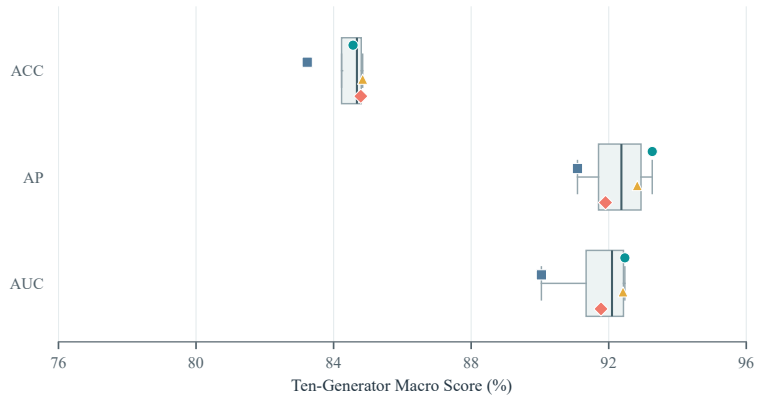

Figure E.1: **Macro Results across Four Random Seeds.** Each point is one run's ten-generator macro score. Boxes show medians and interquartile ranges; whiskers follow the 1.5-interquartile-range rule. Colors and symbols distinguish the four runs; offsets separate overlapping observations.

F Generator-Level Results across Random Seeds

Figure F.1 shows ACC, AP, and AUC across four random seeds for all ten GenVideo generator subsets. Each panel contains forty observed scores, with four results for each generator. AI-generated content is the positive class. Showing the generators individually checks whether the consistency in Table E.1 remains visible before averaging across subsets. Matching symbols let readers follow the same run across ACC, AP, and AUC for each generator.

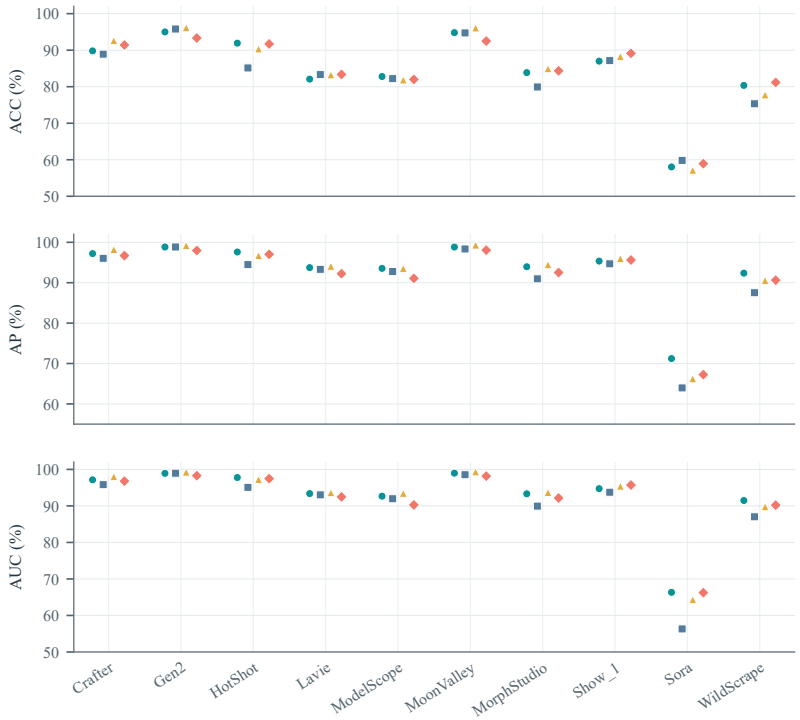

Figure F.1: **Generator-Level Results across Four Random Seeds.** Panels show ACC, AP, and AUC (%) for all ten generators. Each point represents one run; matching colors and symbols denote the same run across panels. Horizontal offsets separate observations within each generator.

For most generators, the four scores within each metric occupy a narrow band. The panels retain the same broad pattern across runs: stronger and weaker generator subsets remain distinguishable without a change in the evaluation protocol. This separated view supports the macro summary without hiding differences among generator subsets. Repeated selection moves individual points, but the broader pattern of comparatively stronger and weaker generators remains recognizable across the three panels.

G Complete Comparison across Generator Subsets

Tables G.1–G.9 expand Table 1 to every generator subset for all twelve detectors. The four benchmarks contain 10, 14, 7, and 26 generator subsets, respectively. D3 uses cosine scoring. ACC is unavailable for D3 and STALL and is marked by a dash. Each reported metric is a percentage; bold marks the largest reported value within a generator and metric. The per-generator presentation preserves variation hidden by a benchmark mean and uses the same metric headings for image- and video-level methods. It also shows how broadly INSIDER's macro leads are distributed across the evaluated generators.

Table G.1: **Complete Comparison on GenVideo,** part 1 of 2. Bold marks the largest reported score for each generator and metric. A dash denotes an unavailable ACC value.

Detection Level	Method	Metric	Crafter	Gen2	HotShot	Lavie	Model Scope
Image	UnivFD	ACC	51.83	51.75	53.17	53.42	61.17
		AP	70.14	65.54	68.92	71.58	86.72
		AUC	73.37	67.99	70.27	73.98	87.72
	NPR	ACC	53.08	50.08	50.00	50.00	54.42
		AP	82.51	46.92	43.58	48.16	84.99
		AUC	83.33	43.72	29.77	42.21	82.11
	RIGID	ACC	48.83	48.50	48.92	48.58	49.92
		AP	39.99	46.90	50.16	46.75	49.62
		AUC	35.43	47.98	50.85	46.34	48.98
Video	D3	ACC	–	–	–	–	–
		AP	77.18	87.93	72.24	76.94	65.24
		AUC	76.36	85.76	67.04	77.18	62.83
	STALL	ACC	–	–	–	–	–
		AP	83.80	87.30	77.80	82.80	75.70
		AUC	83.00	85.60	78.80	83.30	75.60
	CoCoDetect	ACC	55.90	63.55	47.86	52.68	62.14
		AP	57.78	66.35	47.65	49.80	72.24
		AUC	59.31	70.16	47.01	52.24	71.19
	VideoMAEv2-L	ACC	66.81	66.27	63.07	64.82	56.86
		AP	67.23	67.73	65.90	66.40	55.54
		AUC	71.08	70.88	68.51	68.87	59.61
	DeMamba	ACC	71.78	72.14	41.57	69.68	55.71
		AP	81.02	81.33	45.49	77.25	59.72
		AUC	80.59	81.86	50.68	79.16	61.59
	NSG-VD	ACC	64.50	55.70	68.90	56.00	61.60
		AP	70.16	57.72	74.91	58.78	65.77
		AUC	67.26	56.03	72.45	57.10	63.99
	DeCoF	ACC	67.88	80.14	72.29	67.11	63.64
		AP	71.43	87.16	76.26	68.94	67.61
		AUC	74.39	88.92	78.81	73.03	71.14

Table G.1: **Complete Comparison on GenVideo,** part 1 of 2 (continued).

Detection Level	Method	Metric	Crafter	Gen2	HotShot	Lavie	Model Scope
Video	ReStraV	ACC	81.12	66.12	**94.00**	56.86	76.36
		AP	84.44	74.60	93.74	62.42	84.39
		AUC	91.75	83.33	96.32	65.74	89.89
	INSIDER	ACC	**89.81**	**95.00**	91.93	**82.07**	**82.79**
		AP	**97.20**	**98.84**	**97.61**	**93.74**	**93.55**
		AUC	**97.14**	**98.92**	**97.77**	**93.38**	**92.67**

INSIDER has the highest AP and AUC on all five GenVideo subsets in this group and the highest ACC on four. On Crafter and Gen2, its AP and AUC all exceed 97%, and Gen2 reaches 95.00% ACC. Across all five columns, AP remains above 93% and AUC above 92%, pairing four ACC leads with consistently strong score ranking.

Table G.2: **Complete Comparison on GenVideo,** part 2 of 2. Bold marks the largest reported score for each generator and metric. A dash denotes an unavailable ACC value.

Detection Level	Method	Metric	Moon Valley	Morph Studio	Show_1	Sora	Wild Scrape
Image	UnivFD	ACC	51.08	54.33	52.25	54.46	52.50
		AP	65.73	74.87	68.64	74.65	68.15
		AUC	67.96	77.78	71.44	77.10	69.13
	NPR	ACC	57.83	52.25	50.00	52.68	53.92
		AP	91.33	66.18	41.99	**80.40**	63.50
		AUC	91.68	65.45	25.89	80.13	56.62
	RIGID	ACC	48.58	48.75	48.83	47.32	49.17
		AP	40.93	43.36	49.49	36.36	44.39
		AUC	37.18	40.59	51.43	24.30	43.31
Video	D3	ACC	–	–	–	–	–
		AP	84.44	69.45	79.01	67.73	69.34
		AUC	82.30	67.10	76.58	63.36	62.62
	STALL	ACC	–	–	–	–	–
		AP	75.50	81.40	79.50	73.90	66.20
		AUC	72.00	80.30	80.80	71.60	70.30
	CoCoDetect	ACC	66.05	52.93	58.07	53.57	48.36
		AP	67.99	55.19	61.32	57.75	48.53
		AUC	73.59	52.96	62.90	56.51	47.24
	VideoMAEv2-L	ACC	66.61	63.86	66.50	50.89	64.23
		AP	69.19	64.00	69.66	54.78	64.10
		AUC	72.49	67.83	71.96	52.87	68.59
	DeMamba	ACC	74.92	65.14	62.29	56.25	53.73
		AP	82.40	72.07	70.13	60.41	58.85
		AUC	83.83	75.44	71.29	61.13	60.97
	NSG-VD	ACC	65.90	60.30	70.70	59.82	63.48
		AP	70.70	64.22	77.03	67.49	68.27
		AUC	70.23	61.88	75.06	62.24	67.59
	DeCoF	ACC	74.76	68.86	68.64	**78.57**	65.91
		AP	76.83	70.93	71.30	76.71	69.80
		AUC	81.12	75.25	75.60	**82.24**	72.06

Table G.2: **Complete Comparison on GenVideo,** part 2 of 2 (continued).

Detection Level	Method	Metric	Moon Valley	Morph Studio	Show_1	Sora	Wild Scrape
Video	ReStraV	ACC	79.71	82.71	66.71	52.68	72.27
		AP	83.46	89.28	76.69	47.27	77.36
		AUC	90.84	92.64	88.34	41.61	80.05
	INSIDER	ACC	**94.81**	**83.86**	**87.00**	58.04	**80.36**
		AP	**98.83**	**93.96**	**95.35**	71.22	**92.37**
		AUC	**98.98**	**93.31**	**94.73**	66.33	**91.47**

INSIDER leads ACC, AP, and AUC on MoonValley, MorphStudio, Show_1, and WildScrape. MoonValley reaches 94.81% ACC and 98.98% AUC, while the other three winning columns range from 80.36% to 87.00% ACC. AP and AUC exceed 91% on each of these four generators, pairing classification gains with strong ranking.

Across the two GenVideo blocks, INSIDER leads AP and AUC on nine of ten generator subsets and ACC on eight. The benchmark scores in Table 1 therefore aggregate gains that recur across generator columns, with both score ranking and thresholded classification represented in the detailed results.

Table G.3: **Complete Comparison on EvalCrafter,** part 1 of 2. Bold marks the largest reported score for each generator and metric. A dash denotes an unavailable ACC value.

Detection Level	Method	Metric	Moon Valley	floor33	gen2	gen2_d ecemb er	hotshot	lavie-b ase	lavie-i nterpol ation
Image	UnivFD	ACC	58.07	57.00	54.26	53.86	54.43	60.93	55.00
		AP	79.96	78.40	75.04	69.19	72.82	82.21	74.94
		AUC	81.55	80.33	78.42	71.49	74.54	82.49	77.74
	NPR	ACC	79.47	66.64	57.86	62.30	50.64	60.93	51.64
		AP	95.91	88.36	77.79	86.28	52.73	86.18	63.71
		AUC	94.88	85.44	74.72	84.04	44.14	84.00	62.47
	RIGID	ACC	47.44	47.71	48.31	47.71	48.71	49.21	48.50
		AP	41.54	43.34	50.17	36.00	52.35	49.13	46.33
		AUC	37.70	39.57	53.77	25.43	53.26	46.54	44.18
Video	D3	ACC	–	–	–	–	–	–	–
		AP	83.07	68.10	86.98	89.90	68.79	70.23	73.90
		AUC	81.51	66.08	83.83	87.34	63.82	71.52	74.03
	STALL	ACC	–	–	–	–	–	–	–
		AP	73.70	79.80	85.50	87.10	75.90	83.20	80.00
		AUC	70.20	79.00	83.50	86.10	77.40	82.90	82.00
	CoCoDetect	ACC	65.42	52.21	63.07	64.81	47.14	48.57	57.43
		AP	65.36	53.36	63.60	67.47	46.51	45.63	53.32
		AUC	72.22	52.19	67.65	71.80	46.14	46.07	57.78
	VideoMAEv2-L	ACC	66.93	63.93	66.67	64.66	63.14	64.07	63.93
		AP	67.35	62.66	66.85	65.07	64.74	65.00	64.58
		AUC	71.43	66.86	71.13	67.98	67.57	67.58	67.32
	DeMamba	ACC	72.92	68.14	73.72	69.89	36.50	69.29	64.93
		AP	79.68	68.96	78.80	75.41	42.25	73.78	68.85
		AUC	81.10	73.20	80.63	77.07	43.87	76.51	72.15

Table G.3: **Complete Comparison on EvalCrafter,** part 1 of 2 (continued).

Detection Level	Method	Metric	Moon Valley	floor33	gen2	gen2_december	hotshot	lavie-base	lavie-interpolation
Video	NSG-VD	ACC	65.90	59.60	55.40	65.20	69.60	55.80	61.00
		AP	70.70	63.49	55.59	71.67	76.17	58.52	64.73
		AUC	70.23	61.52	55.38	70.06	74.01	56.60	63.05
	DeCoF	ACC	75.96	68.36	83.11	80.83	72.00	64.21	71.93
		AP	80.31	72.86	91.18	87.98	79.10	64.27	76.79
		AUC	83.53	75.42	92.11	89.68	79.75	69.40	79.06
	ReStraV	ACC	79.71	82.71	61.45	71.03	**94.00**	68.43	45.79
		AP	83.46	89.28	71.77	79.44	93.74	79.86	44.54
		AUC	90.84	92.64	81.38	86.77	96.32	86.26	47.23
	INSIDER	ACC	**95.05**	**84.00**	**93.47**	**96.42**	92.07	**82.64**	**81.43**
		AP	**98.69**	**93.99**	**98.25**	**98.97**	**97.38**	**93.76**	**93.05**
		AUC	**98.91**	**93.55**	**98.39**	**99.15**	**97.72**	**93.52**	**92.91**

INSIDER leads AP and AUC on all seven subsets here and ACC on six, with AP and AUC above 92% throughout. Across these seven EvalCrafter generators, its AP ranges from 93.05% to 98.97% and AUC from 92.91% to 99.15%. Every column therefore contributes a high AP and AUC value to the group-level result.

Table G.4: **Complete Comparison on EvalCrafter**, part 2 of 2. Bold marks the largest reported score for each generator and metric. A dash denotes an unavailable ACC value.

Detection Level	Method	Metric	mix-sr	models cope	pika	pika_v1_december_1	show_1	videocrafter-v1.0	zeroscope
Image	UnivFD	ACC	57.86	66.64	57.75	57.14	53.43	59.36	64.97
		AP	79.51	88.02	78.47	78.87	71.10	79.34	86.79
		AUC	80.94	88.63	79.81	81.22	74.13	79.90	87.96
	NPR	ACC	88.93	59.64	58.32	64.79	50.64	75.07	55.27
		AP	**98.74**	84.54	78.50	88.59	52.58	95.20	68.28
		AUC	**98.39**	80.00	75.32	86.18	47.28	94.19	60.78
	RIGID	ACC	48.21	48.71	48.06	48.21	48.29	47.86	49.50
		AP	37.53	47.32	49.03	47.55	51.49	41.12	60.38
		AUC	29.08	45.78	49.63	47.96	53.76	36.40	65.89
Video	D3	ACC	–	–	–	–	–	–	–
		AP	82.21	62.58	84.81	84.48	77.30	74.19	65.15
		AUC	79.84	62.89	83.22	82.63	75.29	74.80	58.91
	STALL	ACC	–	–	–	–	–	–	–
		AP	84.10	73.60	79.30	85.00	77.80	81.20	68.90
		AUC	82.50	74.10	78.40	83.90	79.60	81.70	68.00
	CoCoDetect	ACC	57.36	61.43	63.70	66.29	57.36	55.07	50.00
		AP	57.90	69.77	65.53	71.04	59.58	56.33	58.38
		AUC	60.43	70.27	70.10	76.13	62.00	57.48	54.73
	VideoMAEv2-L	ACC	67.36	56.93	68.58	69.86	66.57	64.71	50.50
		AP	67.58	54.37	71.54	73.15	68.40	63.31	47.88
		AUC	71.49	58.70	73.78	75.65	71.06	67.79	48.82

Table G.4: **Complete Comparison on EvalCrafter**, part 2 of 2 (continued).

Detection Level	Method	Metric	mix-sr	models cope	pika	pika_v 1_dece mber_ 1	show_ 1	videocr after-v 1.0	zerosc ope
Video	DeMamba	ACC	77.29	56.93	68.08	72.43	63.57	65.14	42.39
		AP	85.94	58.41	72.02	78.23	67.92	68.30	42.10
		AUC	86.01	61.13	74.88	79.50	70.25	70.95	40.46
	NSG-VD	ACC	65.70	61.10	70.20	69.90	70.20	66.20	72.00
		AP	72.46	65.80	76.29	75.56	76.32	73.59	80.21
		AUC	70.67	63.64	75.62	74.25	74.46	71.55	77.04
	DeCoF	ACC	77.79	64.07	74.39	73.07	69.50	65.50	55.27
		AP	82.07	67.05	82.34	76.63	73.18	66.71	56.49
		AUC	84.86	70.44	83.41	80.97	76.87	70.64	60.58
	ReStraV	ACC	81.36	76.36	61.55	59.36	66.71	81.43	**86.12**
		AP	84.57	84.39	72.92	69.38	76.69	86.70	82.30
		AUC	91.95	89.89	79.99	75.66	88.34	92.53	92.10
	INSIDER	ACC	**91.93**	**82.93**	**96.13**	**93.57**	**87.14**	**87.57**	79.26
		AP	97.77	**93.65**	**99.28**	**98.35**	**95.30**	**96.06**	**92.57**
		AUC	98.09	**92.96**	**99.32**	**98.26**	**94.85**	**95.88**	**92.86**

Across the seven EvalCrafter generators in this table, INSIDER leads all three metrics on five, including ModelScope, both Pika variants, Show_1, and VideoCrafter-v1.0. It ranks first on six of seven generators for each metric. On Pika, it reaches 96.13% ACC, 99.28% AP, and 99.32% AUC; the other Pika variant also exceeds 93% ACC and 98% AP and AUC.

Table G.5: **Complete Comparison on VideoPhy-2.** Bold marks the largest reported score for each generator and metric. A dash denotes an unavailable ACC value.

Detection Level	Method	Metric	cogvid eo	cosmos	hunyua n	ray2	sora	videocr after	wan
Image	UnivFD	ACC	55.35	54.27	50.51	54.31	50.83	50.85	52.96
		AP	71.12	70.88	54.50	72.49	42.31	56.42	61.06
		AUC	71.80	74.11	56.30	73.09	34.58	59.15	60.01
	NPR	ACC	51.95	52.82	50.68	52.79	51.67	50.85	67.68
		AP	62.49	63.41	49.89	63.09	72.30	45.27	87.71
		AUC	61.02	58.65	46.28	56.66	77.62	36.85	84.76
	RIGID	ACC	48.30	47.95	48.38	48.22	48.33	47.72	47.80
		AP	50.37	39.52	48.99	40.89	36.70	42.15	42.07
		AUC	52.27	33.93	50.30	36.32	26.58	38.45	37.66
Video	D3	ACC	–	–	–	–	–	–	–
		AP	62.12	68.67	63.26	64.40	64.32	59.41	62.63
		AUC	60.49	63.87	60.23	59.89	57.14	52.52	57.95
	STALL	ACC	–	–	–	–	–	–	–
		AP	56.16	52.12	52.14	51.54	54.97	66.57	55.41
		AUC	53.73	46.49	47.11	46.53	51.14	60.77	51.02
	CoCoDetect	ACC	38.46	42.74	41.74	47.72	52.50	59.56	49.49
		AP	40.09	41.21	41.86	43.84	49.89	57.83	44.65
		AUC	37.05	40.02	39.77	44.00	52.56	61.29	46.42

Table G.5: **Complete Comparison on VideoPhy-2,** part 1 of 1 (continued).

Detection Level	Method	Metric	cogvideo	cosmos	hunyuan	ray2	sora	videocrafter	wan
Video	VideoMAEv2-L	ACC	56.37	51.97	54.09	64.72	47.50	48.48	56.09
		AP	56.15	50.33	55.99	65.72	50.40	46.25	54.53
		AUC	56.95	51.95	55.38	69.06	48.03	45.76	56.33
	DeMamba	ACC	52.21	51.97	47.36	46.45	50.00	57.53	52.54
		AP	51.46	50.16	47.67	46.38	46.97	57.85	50.73
		AUC	55.32	54.39	50.26	47.65	49.08	64.74	55.76
	NSG-VD	ACC	68.97	52.17	57.39	60.00	57.98	**62.73**	56.67
		AP	73.22	52.52	56.78	65.56	65.49	59.34	61.50
		AUC	70.86	53.91	57.24	63.69	65.00	63.00	60.67
	DeCoF	ACC	65.62	62.39	64.40	58.76	63.33	58.04	56.85
		AP	65.58	63.77	70.11	61.73	68.01	60.30	58.85
		AUC	72.14	68.85	70.62	63.83	73.11	68.45	63.52
	ReStraV	ACC	**76.66**	54.62	47.44	50.51	49.17	47.55	51.61
		AP	79.76	57.38	48.33	48.14	43.70	53.24	52.01
		AUC	**89.46**	58.14	52.35	47.56	41.58	61.20	53.06
	INSIDER	ACC	72.41	**74.87**	**81.94**	**97.59**	**66.67**	53.30	**78.76**
		AP	**87.71**	**90.26**	**94.27**	**98.99**	**82.64**	**67.43**	**92.50**
		AUC	88.07	**92.19**	**94.67**	**99.32**	**84.47**	**73.58**	**92.72**

INSIDER leads AP on all seven VideoPhy-2 subsets and AUC on six, extending its macro ranking advantage across individual generators. Ray2 provides a clear example, with 97.59% ACC, 98.99% AP, and 99.32% AUC. Hunyuan and Wan also favor INSIDER on all three metrics, so the strong macro result covers several VideoPhy-2 generators.

Table G.6: **Complete Comparison on VideoFeedback2,** part 1 of 4. Bold marks the largest reported score for each generator and metric. A dash denotes an unavailable ACC value.

Detection Level	Method	Metric	anidiff	cogvideox15_5b	cogvideox_2b	cogvideox_5b	hotshot_xl	kling	latte
Image	UnivFD	ACC	52.80	55.69	53.90	51.50	52.60	51.70	52.90
		AP	68.60	76.03	68.59	59.73	68.85	58.39	69.26
		AUC	72.50	76.96	69.72	62.13	72.00	59.15	72.47
	NPR	ACC	51.00	59.39	56.50	55.00	50.70	60.30	55.20
		AP	64.34	83.15	80.67	77.92	49.69	79.95	81.02
		AUC	70.23	80.73	79.35	76.75	42.26	74.96	81.16
	RIGID	ACC	48.30	49.47	50.80	49.40	48.30	47.90	48.30
		AP	54.01	49.68	54.95	48.97	54.93	38.41	48.27
		AUC	59.10	47.42	55.62	48.04	58.92	31.21	48.97
Video	D3	ACC	–	–	–	–	–	–	–
		AP	86.25	88.18	79.99	79.72	86.31	87.36	81.87
		AUC	82.63	85.19	76.26	75.00	81.89	84.23	78.41
	STALL	ACC	–	–	–	–	–	–	–
		AP	79.04	75.90	74.59	76.98	72.11	70.16	71.59
		AUC	79.80	77.28	76.18	79.77	74.63	71.48	72.91

Table G.6: **Complete Comparison on VideoFeedback2,** part 1 of 4 (continued).

Detection Level	Method	Metric	anidiff	cogvideox15_5b	cogvideox_2b	cogvideox_5b	hotshot_xl	kling	latte
Video	CoCoDetect	ACC	48.40	53.57	44.40	57.10	50.30	54.80	47.50
		AP	49.21	50.48	43.33	50.61	48.10	53.93	48.41
		AUC	49.52	53.33	41.30	55.25	49.91	57.22	46.91
	VideoMAEv2-L	ACC	68.90	63.62	59.80	65.60	65.10	57.80	66.30
		AP	67.57	64.64	60.52	66.05	63.97	57.96	67.40
		AUC	72.29	66.22	62.42	69.12	67.69	60.66	70.57
	DeMamba	ACC	66.10	56.88	60.10	59.80	55.40	62.50	61.40
		AP	72.10	59.33	65.94	65.53	58.86	69.47	68.76
		AUC	72.39	58.82	65.18	66.03	58.60	69.16	68.20
	NSG-VD	ACC	72.90	68.39	60.30	62.30	69.10	57.00	57.20
		AP	78.86	75.72	64.62	67.24	75.38	60.38	60.00
		AUC	79.56	75.73	62.08	65.45	74.16	59.15	57.91
	DeCoF	ACC	53.50	51.06	52.10	53.20	53.00	53.80	53.60
		AP	77.70	67.75	72.13	75.15	73.95	88.30	77.22
		AUC	78.43	66.80	70.11	74.89	73.82	87.73	77.93
	ReStraV	ACC	76.60	59.92	57.50	63.20	87.50	53.40	76.90
		AP	81.23	72.40	69.15	73.42	84.57	59.95	81.32
		AUC	90.07	83.78	80.09	84.17	92.99	65.33	89.61
	INSIDER	ACC	**88.50**	**86.64**	**69.90**	**82.00**	**92.50**	**87.20**	**86.30**
		AP	**96.10**	**95.26**	**85.88**	**93.52**	**97.57**	**96.00**	**95.18**
		AUC	**97.11**	**94.93**	**85.40**	**93.19**	**98.15**	**96.17**	**94.87**

INSIDER leads ACC, AP, and AUC on every generator in this first VideoFeedback2 group. These seven columns include multiple CogVideoX variants alongside other generators. INSIDER's AP is at least 85.88% and its AUC at least 85.40% throughout this block; every reported generator-metric maximum belongs to its row.

Table G.7: **Complete Comparison on VideoFeedback2,** part 2 of 4. Bold marks the largest reported score for each generator and metric. A dash denotes an unavailable ACC value.

Detection Level	Method	Metric	lavie_base	ltx_video_091	ltx_video_095	magictime	mochi1_preview	modelscope	openosora_plan_v1_3
Image	UnivFD	ACC	57.80	63.10	52.90	52.51	57.80	55.40	56.53
		AP	79.12	86.39	68.76	67.38	80.52	75.06	74.17
		AUC	79.53	87.28	72.07	69.74	82.47	77.72	74.70
	NPR	ACC	61.50	61.00	55.50	53.11	51.90	50.00	74.40
		AP	84.99	78.87	81.22	77.40	62.36	**88.96**	**96.33**
		AUC	82.69	72.46	80.48	79.20	58.79	**92.07**	**96.07**
	RIGID	ACC	48.90	49.60	48.70	47.70	48.90	48.90	48.90
		AP	49.83	57.42	49.81	50.76	52.85	56.28	53.88
		AUC	50.06	61.25	53.41	54.14	55.44	61.29	59.03
Video	D3	ACC	–	–	–	–	–	–	–
		AP	86.22	78.31	86.54	86.22	78.52	71.41	89.66
		AUC	84.34	73.91	83.49	82.88	74.22	61.19	87.72

Table G.7: **Complete Comparison on VideoFeedback2,** part 2 of 4 (continued).

Detection Level	Method	Metric	lavie_base	ltx_video_091	ltx_video_095	magicitime	mochi1_preview	modelscope	openso ra_plan _v1_3
Video	STALL	ACC	–	–	–	–	–	–	–
		AP	78.80	67.54	65.61	69.52	64.86	68.42	79.12
		AUC	79.56	69.40	67.77	67.65	68.55	70.87	80.12
	CoCoDetect	ACC	50.90	59.60	57.30	50.50	38.50	19.30	62.85
		AP	47.62	67.71	57.94	49.51	41.00	32.04	57.10
		AUC	49.41	68.32	60.91	50.86	36.53	8.67	64.25
	VideoMAEv2-L	ACC	65.50	62.50	59.00	62.63	66.80	46.00	68.27
		AP	69.42	62.98	60.05	62.12	66.47	44.08	71.38
		AUC	71.18	66.88	61.36	65.70	70.10	43.20	73.90
	DeMamba	ACC	68.40	53.50	61.80	57.62	52.40	44.10	68.57
		AP	77.17	57.07	69.81	62.10	54.93	47.28	79.39
		AUC	75.84	55.12	67.95	63.51	53.13	48.01	77.65
	NSG-VD	ACC	57.40	60.60	61.30	66.57	61.50	**62.70**	57.52
		AP	60.14	66.69	65.50	72.43	65.18	68.10	60.45
		AUC	59.53	64.08	63.07	72.54	63.88	67.83	58.90
	DeCoF	ACC	52.70	52.60	53.90	53.71	52.50	52.00	53.92
		AP	65.66	60.92	79.87	74.07	72.91	49.30	79.97
		AUC	67.35	61.37	80.77	75.32	71.21	51.84	79.57
	ReStraV	ACC	59.20	76.30	49.60	62.02	69.50	46.00	68.78
		AP	71.77	83.75	52.52	73.48	78.17	52.76	78.16
		AUC	83.61	88.74	56.38	85.13	87.01	60.76	82.38
	INSIDER	ACC	**87.50**	**85.00**	**78.20**	**95.39**	**87.60**	51.40	**87.75**
		AP	**95.52**	**94.76**	**91.35**	**98.13**	**96.38**	58.95	95.80
		AUC	**94.95**	**94.30**	**91.72**	**98.77**	**96.61**	61.85	95.22

INSIDER leads ACC on six subsets and AP and AUC on five, including all three metrics for the first five generators. The first five columns, from LaVie-Base through Mochi1-Preview, show simultaneous leads in classification and score ranking. Their AP and AUC values stay above 91%, while ACC ranges from 78.20% to 95.39%, showing the pattern across varied score levels.

Table G.8: **Complete Comparison on VideoFeedback2,** part 3 of 4. Bold marks the largest reported score for each generator and metric. A dash denotes an unavailable ACC value.

Detection Level	Method	Metric	opensor a_v1_2	pika_v2 _2	ruyi	sora	stepvide o_t2v	stepvide o_t2v_l ow_vra m
Image	UnivFD	ACC	53.20	49.84	54.89	53.20	50.80	52.45
		AP	72.66	55.03	69.62	68.72	56.76	59.58
		AUC	77.22	59.43	70.80	71.35	58.03	58.94
	NPR	ACC	50.40	59.09	58.97	62.46	82.60	81.53
		AP	49.20	80.80	83.80	**87.41**	97.01	97.37
		AUC	43.52	77.15	81.75	84.64	96.60	96.85
	RIGID	ACC	50.10	48.28	47.55	47.98	48.40	48.37
		AP	57.10	40.07	47.52	39.60	43.15	44.13
		AUC	61.26	35.57	48.55	33.57	39.58	41.66

Table G.8: **Complete Comparison on VideoFeedback2,** part 3 of 4 (continued).

Detection Level	Method	Metric	opensora_v1_2	pika_v2_2	ruyi	sora	stepvideo_t2v	stepvideo_t2v_low_vram
Video	D3	ACC	–	–	–	–	–	–
		AP	76.28	80.22	86.52	79.75	84.04	86.24
		AUC	71.94	77.04	82.67	74.27	81.81	83.87
	STALL	ACC	–	–	–	–	–	–
		AP	61.15	73.61	70.61	70.79	70.15	68.43
		AUC	63.65	75.02	73.24	70.72	71.42	71.25
	CoCoDetect	ACC	60.10	52.82	60.33	47.64	56.00	57.86
		AP	62.62	50.75	61.69	48.93	52.74	53.82
		AUC	65.55	52.88	65.39	48.89	56.85	58.02
	VideoMAEv2-L	ACC	64.80	55.17	64.40	50.51	61.40	61.73
		AP	67.73	54.13	68.01	49.17	63.83	63.19
		AUC	69.24	56.01	67.34	50.54	63.14	65.64
	DeMamba	ACC	58.50	**66.30**	57.88	55.22	63.40	60.92
		AP	64.60	72.32	63.32	57.26	69.19	67.54
		AUC	62.76	73.24	63.16	58.60	68.96	66.93
	NSG-VD	ACC	65.40	56.58	55.71	62.18	59.80	59.53
		AP	69.73	58.60	64.24	67.43	64.05	63.77
		AUC	69.44	57.66	62.10	65.71	62.13	63.12
	DeCoF	ACC	52.20	54.08	53.80	54.04	54.00	53.98
		AP	62.35	**89.39**	78.71	85.30	83.11	84.77
		AUC	62.43	**89.47**	80.12	**87.02**	82.76	84.43
	ReStraV	ACC	**87.20**	48.59	49.46	45.29	61.40	62.55
		AP	88.29	49.80	53.36	43.89	72.27	70.94
		AUC	**93.60**	50.34	58.29	44.56	74.31	72.24
	INSIDER	ACC	79.20	65.83	**87.23**	**67.85**	**91.40**	**91.53**
		AP	**91.55**	81.67	**94.34**	83.31	**97.21**	**97.41**
		AUC	90.78	82.64	**94.73**	84.64	**97.31**	**97.24**

INSIDER leads all three metrics on Ruyi and both StepVideo subsets and also has the highest ACC on Sora. On Ruyi it records 87.23% ACC, 94.34% AP, and 94.73% AUC. Both StepVideo variants exceed 91% ACC and 97% on AP and AUC, extending the three-metric lead across distinct generator columns.

Table G.9: **Complete Comparison on VideoFeedback2,** part 4 of 4. Bold marks the largest reported score for each generator and metric. A dash denotes an unavailable ACC value.

Detection Level	Method	Metric	text2video_zero	vchitect 2	videocrafter2	wanx21_14b	wanx21_1_3b	zeroscope
Image	UnivFD	ACC	65.80	54.21	50.00	70.82	56.30	70.00
		AP	90.95	74.43	49.97	90.00	74.64	91.08
		AUC	91.94	77.83	51.06	89.52	74.94	91.06
	NPR	ACC	50.40	57.92	55.90	72.95	80.80	50.00
		AP	71.95	78.96	72.94	95.29	**97.65**	74.84
		AUC	74.43	76.28	69.05	94.91	**97.29**	83.53
	RIGID	ACC	55.10	47.39	47.60	48.75	48.70	48.10
		AP	66.32	47.64	36.18	51.25	48.84	53.79
		AUC	65.21	48.99	26.01	53.74	49.21	60.06
Video	D3	ACC	–	–	–	–	–	–
		AP	42.97	85.21	90.13	79.68	84.93	70.70
		AUC	38.76	81.17	87.09	72.58	80.66	60.08
	STALL	ACC	–	–	–	–	–	–
		AP	74.49	79.45	77.35	67.96	69.42	70.76
		AUC	73.44	80.66	75.65	70.70	71.08	70.50
	CoCoDetect	ACC	50.30	65.93	57.60	66.73	58.00	39.75
		AP	53.17	69.14	58.96	56.65	55.99	42.62
		AUC	50.53	74.58	61.09	66.04	59.08	36.31
	VideoMAEv2-L	ACC	58.00	63.83	65.40	62.81	63.20	60.25
		AP	54.25	62.18	67.08	62.29	62.86	56.03
		AUC	58.14	66.89	69.38	66.74	65.65	60.40
	DeMamba	ACC	52.30	69.34	77.40	66.37	62.80	55.95
		AP	54.89	77.51	88.65	72.66	68.09	58.68
		AUC	57.65	76.64	87.77	72.87	67.66	59.75
	NSG-VD	ACC	56.90	51.95	60.90	58.90	56.70	56.46
		AP	66.76	53.21	63.60	61.75	59.22	59.29
		AUC	63.38	53.02	63.48	62.13	58.35	60.02
	DeCoF	ACC	53.30	53.91	53.90	54.09	53.60	53.29
		AP	49.70	83.14	86.09	82.86	79.18	56.85
		AUC	53.65	84.27	86.19	84.36	79.54	61.57
	ReStraV	ACC	**85.90**	64.33	75.20	52.85	52.10	49.62
		AP	**91.01**	74.54	80.79	62.59	56.35	63.40
		AUC	**92.07**	87.01	89.91	68.00	59.17	75.72
	INSIDER	ACC	55.30	**81.46**	**94.40**	**93.42**	**81.40**	**77.47**
		AP	72.42	**94.21**	**98.37**	**97.63**	94.03	**92.95**
		AUC	75.39	**95.29**	**98.73**	**98.21**	94.05	**94.88**

INSIDER leads ACC on five of these six subsets and AP and AUC on four, with four columns led across all metrics. The simultaneous leads occur on Vchitect2, VideoCrafter2, Wanx21-14B, and ZeroScope. VideoCrafter2 reaches 94.40% ACC, 98.37% AP, and 98.73% AUC; Wanx21-14B also exceeds 93% ACC and 97% in both ranking metrics.

H Further Discussion

▷ **Q1. What does a forensic neuron represent in INSIDER?** A forensic neuron is a selected channel coordinate in the frozen video encoder. Its spatial and temporal responses are pooled into one video-level value before scoring and classification. The term identifies a coordinate useful for

authenticity detection under the fitted probes and final readout. INSIDER selects evidence already present in pretrained representations, while the learned linear readout determines how that evidence contributes to the final decision. Layer-channel indexing lets the selected set combine responses from several encoder depths, and the final classifier weights those pooled values jointly. The resulting representation connects explicit coordinate selection with a compact video-level prediction.

▷ **Q2. Why are both decision leverage and supervised sensitivity needed?** DL measures the mean absolute change in a fixed layer probe's logit after a pooled response is zeroed. SS measures class-averaged supervised loss sensitivity to that response under the same probe, using the observed labels. Their product combines this supervised signal with the response's influence on the probe score. The matched-count factor ablation in Table 4 supports using both: removing either factor lowers ACC, AP, and AUC while keeping the retained-neuron budget fixed. The ranking uses source-training responses and no target labels. Both factors are computed for the same pooled coordinates through the same fixed probe, making their product a consistent within-layer ranking rule. In the matched-count GenVideo comparison, that coupled ranking produces the highest reported scores on all three metrics.

▷ **Q3. What do the coordinate interventions establish?** The interventions measure the fixed classifier's dependence on selected pooled responses. Targeted removal lowers the correct-class margin much more than matched random removal, and the illustrated predictions cross the decision boundary away from their labeled classes. The encoder and readout remain fixed during removal, so the observed change is directly attributable to the intervened representation. Table D.1 shows the accuracy effect separately for the two classes, and Table D.2 shows positive paired margin differences with intervals above zero. Figure D.1 then presents the distribution of changes across source-validation videos.

▷ **Q4. How should the spatial contribution maps be read?** Figure 7 shows positive spatial contributions to the indicated class, computed using each method's source-trained linear classifier. No Selection uses pooled final-layer features, whereas INSIDER uses selected forensic neuron responses. Both methods share the frozen encoder, displayed frames, and overlay opacity. Within each column, the paired maps share a normalization scale, allowing contributions at corresponding spatial locations to be compared. Warmer colors indicate stronger positive support for the indicated class. The scales support comparisons within each video. In the illustrated AI-generated cases, INSIDER shows stronger correct-class support in regions that contribute weakly under No Selection. These differences illustrate how neuron selection and readout change the use of existing video features without updating the encoder.

▷ **Q5. Why retain nine layers?** Figure C.1 compares seven to twelve probe-ranked layers on GenVideo. Nine layers improve substantially over seven or eight, while eleven and twelve yield lower scores. Although ten layers score slightly higher, the nine-layer setting remains within 0.5 percentage points on every metric with 402 rather than 442 selected neurons. We chose nine on GenVideo as a compact default and kept this configuration fixed for the other benchmarks.

▷ **Q6. How is INSIDER applied to another video encoder?** INSIDER repeats source-based probe fitting, layer selection, neuron mining, and readout training for each encoder. The selected coordinates and classifier are encoder-specific; Figure 8 reports the four resulting comparisons. Reapplying the same stages allows layers and coordinates to reflect each encoder's own representation while preserving a fixed detector at target evaluation. The cross-encoder figure shows the resulting performance comparisons under this repeated source-based procedure.

Limitations and Future Work. First, the experiments use 16 sampled frames per video, leaving shorter input budgets untested. Future work could compare smaller frame sets under the same evaluation protocol to measure the tradeoff between inference cost and detection performance. Second, predictions use a fixed 0.5 threshold. Calibrating this threshold on held-out source videos and then applying it unchanged across target benchmarks could offer more control over the operating point without using target labels. These practical studies would complement label-efficient updates for new generators and spatiotemporal forgery localization.