arXiv:2608.03008v1 [cs.CV] 4 Aug 2026

V-FIND: Revealing the Intrinsic Forgery Knowledge Encoded in Video Forgery Detectors

Shichao Kan[1], Chengpeng Hong[1], Jingtong Dou[2], Chuancheng Shi[2], Yuhan Liu[1], Linrui Xu[1], Yixiong Liang[1], Yigang Cen[1], Yanpeng Sun[3], Fei Shen[3*], Tat-Seng Chua[3]

[1] Central South University
[2] The University of Sydney
[3] NExT++ Research Centre, National University of Singapore
[*] Corresponding author

Abstract

As generated videos become increasingly realistic, reliable video forgery detection is increasingly important. Existing studies typically optimize and use video forgery detectors as black boxes, while the latent forgery-discriminative knowledge inside them remains largely unexplored. Instead of continuing to rely on resource-intensive full-model retraining to steadily improve detection performance, we ask whether video forgery detection can also be achieved by uncovering and activating sparse forensic knowledge within the detector. We find that forgery-discriminative knowledge is not uniformly distributed across the full representation space, but is concentrated in a sparse set of functionally specialized neurons. Based on this insight, we propose a video forgery-intrinsic neuron discovery (V-FIND) framework. V-FIND first localizes critical layers that exhibit pronounced discrepancies between real and forged videos, and then identifies latent anchor neurons that consistently carry forgery-discriminative signals, organizing them into a compact forensic subspace. With the original backbone frozen and only a lightweight linear classifier trained, this subspace still delivers strong detection performance across multiple external benchmarks for generated videos. Further neuron intervention experiments provide direct evidence for the functional specificity of the discovered neurons. Overall, these results suggest that video forgery detectors contain sparse, extractable, and reusable forgery-discriminative knowledge, offering a new perspective on understanding and exploiting their intrinsic forensic capability.

1. Introduction

Artificial intelligence-generated content (AIGC) (Cao et al. 2026; Bie et al. 2025; Shi et al. 2026a) has advanced rapidly, driving the creation and deployment of high-quality synthetic media. This evolution is particularly pronounced in the field of video generation. Driven by breakthrough technologies such as latent and video diffusion models (Ho et al. 2022; Singer et al. 2022), text-to-video generation pipelines, diffusion transformers (DiTs) (Peebles and Xie 2023), and large-scale world simulators (Brooks et al. 2024; Kling 2024), contemporary commercial engines are now capable of synthesizing visually flawless videos. As synthetic videos become increasingly photorealistic, they pose more immediate threats to media authenticity, information security, and public trust. Reliable and effective video forgery detection has therefore become an important and urgent problem.

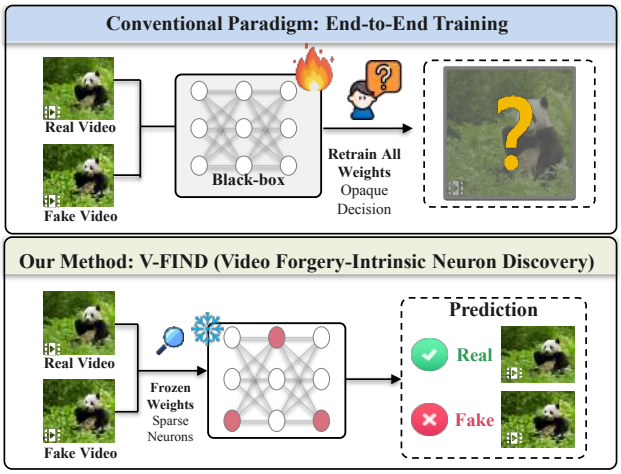

Figure 1: **Comparison of conventional end-to-end training and our V-FIND paradigm.** Existing methods retrain all backbone weights as an opaque black box. Conversely, V-FIND uncovers sparse intrinsic forgery neurons within a frozen backbone for parameter-efficient detection.

In response to these escalating threats, the defense landscape has dynamically evolved from deepfake detectors (Qian et al. 2020; Yan et al. 2023; Xu et al. 2023) focused on facial manipulations to more recent detectors for fully AI-generated videos (Ma et al. 2024; Ni et al. 2025; Ji et al. 2024; Song et al. 2024; Chen et al. 2024; Corvi et al. 2025; Kundu et al. 2025; Li et al. 2026a,b). While this steady progression has pushed benchmark performance to new heights, it has largely relied on a homogeneous recipe of scaling up visual backbones, expanding training corpora, and performing resource-intensive end-to-end optimization. Consequently, existing studies typically optimize and use video forgery detectors as indivisible black boxes, while the latent forgery-discriminative knowledge encoded within them has yet to be systematically uncovered and exploited. This gap prompts us to rethink video forgery detection from an intrinsic-knowledge perspective. Instead of continuing to rely on full-model retraining to steadily improve detection performance, we ask whether video forgery detection can also be achieved by uncovering and activating sparse forensic knowledge within the detector. Figure 1 contrasts this perspective

with the conventional end-to-end detector paradigm.

Under this rethinking, we shift the focus from optimizing the detector as an indivisible whole to examining how forgery-discriminative knowledge is organized within its internal representation space. Our central view is that such knowledge is not uniformly distributed across the full representation space, but is concentrated in a sparse set of functionally specialized neurons. If these neurons consistently carry forgery-discriminative signals, they should be explicitly localizable and could be organized into a compact forensic subspace. This possibility suggests a complementary route to video forgery detection: rather than continuing to update the full detector, we seek to uncover and activate its sparse forensic knowledge while keeping the original backbone frozen. If the resulting subspace can still deliver strong detection performance with only a lightweight linear classifier trained, it would indicate that the detector contains sparse, extractable, and reusable forgery-discriminative knowledge. This rethinking therefore shifts the objective from solely optimizing a black-box detector toward understanding and exploiting its intrinsic forensic capability.

To this end, we propose V-FIND, a video forgery-intrinsic neuron discovery framework that uncovers forgery-discriminative knowledge through a coarse-to-fine procedure. V-FIND first localizes critical layers that exhibit pronounced discrepancies between real and forged videos, and then identifies latent anchor neurons (LANs) that consistently carry forgery-discriminative signals, organizing them into a compact forensic subspace. With the original backbone frozen and only a lightweight linear classifier trained, this subspace still delivers strong detection performance across multiple external benchmarks for generated videos. Further neuron intervention experiments provide direct evidence for the functional specificity of the discovered neurons. Overall, these results suggest that video forgery detectors contain sparse, extractable, and reusable forgery-discriminative knowledge, offering a new perspective on understanding and exploiting their intrinsic forensic capability. Our contributions are threefold:

- We reveal that forgery-discriminative knowledge in video forgery detectors is concentrated in a sparse set of functionally specialized neurons, motivating an intrinsic-knowledge perspective for video forgery detection.
- We propose V-FIND, a video forgery-intrinsic neuron discovery framework that localizes critical layers and identifies latent anchor neurons, organizing them into a compact forensic subspace in a coarse-to-fine.
- We show that the forensic subspace achieves strong detection performance with a frozen backbone and a lightweight linear classifier, and neuron intervention verifies the functional specificity of identified neurons.

2. Related Work

AI-Generated Video Detection. AI-generated video detection has recently expanded from traditional deepfake detection (Yan et al. 2023; Xu et al. 2023) to the detection of fully synthetic videos (Ma et al. 2024; Kundu et al. 2025; Ni et al. 2025). Existing methods mainly follow two technical routes. One line focuses on single-frame spatial artifacts (Ji et al. 2024; Song et al. 2024; Li et al. 2026a), detecting synthetic content through anomalous textures, local visual cues, or multimodal representations extracted from individual frames. Another line emphasizes multi-frame temporal or spatiotemporal inconsistencies (Chen et al. 2024; Corvi et al. 2025; Li et al. 2026b), improving robustness across generators through stronger video backbones, video-level representation learning, temporal modeling, forensic-oriented augmentation, or native-scale artifact preservation. Despite these different designs, both lines mainly treat the detector as a black-box mapping from input videos to binary labels, focusing on how to build a stronger detector rather than on how forgery-discriminative knowledge is organized across layers and neurons inside the detector.

Neuron Interpretability. Neuron interpretability studies suggest that model capabilities are often not uniformly distributed across the full representation space, but can be localized to sparse, functionally specialized neurons, concepts, circuits, or subspaces (Bau et al. 2017; Kim et al. 2018; Olah, Mordvintsev, and Schubert 2017; Alain and Bengio 2016; Olah et al. 2020; Bau et al. 2020; Schwettmann et al. 2023; Huang et al. 2025; Tang et al. 2024). Recent work further uses domain-specific neuron discovery and sparse feature decomposition (Huo et al. 2024; Zaigrajew, Baniecki, and Biecek 2025; Huang et al. 2026; Lim et al. 2024) to reveal specialized internal units. Complementary lines on concept ablation, neuron-level intervention, and circuit discovery (Kumari et al. 2023; He et al. 2025; Kwon, Lee, and Choi 2025) indicate that manipulating a small subset of internal units can amplify, suppress, or erase specific behaviors, while neuron importance scoring (Xie et al. 2021) provides practical tools for ranking and selecting task-relevant units. The convergence of these perspectives lays the groundwork for our V-FIND framework, marking a paradigm shift from acquiring external skills through fine-tuning to explicitly awakening intrinsic forgery-discriminative knowledge already encoded within internal representations.

3. Method

As shown in Figure 2, V-FIND follows a coarse-to-fine pipeline. First we localize critical layers by measuring complementary discrepancies between real and forged representations, and then identifies latent anchor neurons (LANs) within these layers using layer-wise probe-based neuron scoring. The selected LAN activations form a compact forensic subspace, which is fed into a lightweight linear classifier while the backbone remains frozen.

Formally, given an input video clip

$$x = \{I_t\}_{t=1}^{T} \in \mathrm{R}^{T\times H\times W\times C}, \tag{1}$$

where T is the number of sampled frames and H, W, and C denote the frame height, width, and channels, respectively. The hidden representation at layer ℓ is

$$H^{(\ell)}(x) \in \mathrm{R}^{N_t\times N_s\times M}, \tag{2}$$

where N_t, N_s and M denote the temporal tokens, spatial tokens, and latent dimension, respectively. We obtain the

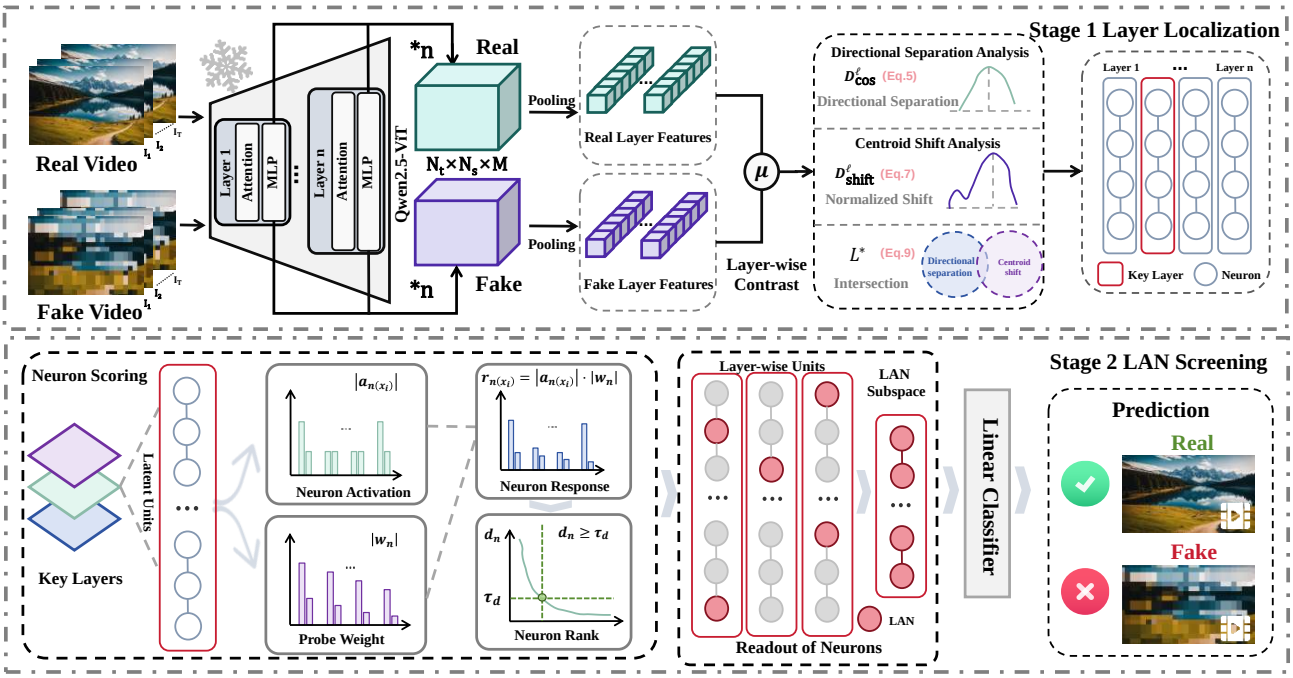

Figure 2: **Overview of the proposed V-FIND framework.** We first localize critical layers by complementary layer-wise discrepancy signals, then identify sparse latent anchor neurons within these layers, and finally train a lightweight detector on the resulting compact forensic subspace.

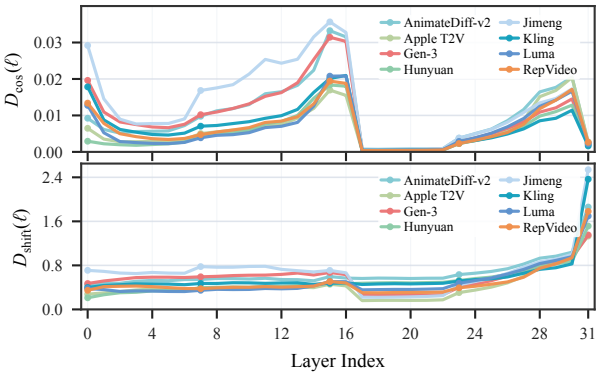

Figure 3: **Layer-localization signals across 8 fake-source subsets** from the detector training distribution.

unit-level representation by spatiotemporal average pooling:

$$\mathbf{a}^{(\ell)}(x) = \frac{1}{N_t N_s} \sum_{i=1}^{N_t} \sum_{j=1}^{N_s} H_{i,j}^{(\ell)}(x). \tag{3}$$

The scalar $a_k^{(\ell)}(x)$ denotes the response of the k-th latent unit at layer ℓ for input video x. We treat each unit index pair $n = (\ell, k)$ as a candidate latent anchor neuron. This unit-level representation serves as the common internal object for layer localization, neuron discovery, and compact forensic subspace construction.

3.1 Layer Localization

We first localize layers that exhibit strong real–fake discrepancies, thereby narrowing the search space for subsequent LAN discovery. Specifically, we compute two complementary layer-wise discrepancy signals on the discovery split and retain the layers highlighted by both.

The first signal measures directional separation between real and fake representations. For each layer ℓ and class $c \in \{\text{real}, \text{fake}\}$, we define the class centroid

$$\mu_c^{(\ell)} = \frac{1}{|\mathcal{D}_c|} \sum_{x \in \mathcal{D}_c} \mathbf{a}^{(\ell)}(x). \tag{4}$$

Based on these centroids, we define the directional-separation signal

$$D_{\cos}(\ell) = 1 - \frac{\mu_{\text{real}}^{(\ell)\top} \mu_{\text{fake}}^{(\ell)}}{\|\mu_{\text{real}}^{(\ell)}\|_2 \|\mu_{\text{fake}}^{(\ell)}\|_2}, \tag{5}$$

which measures how strongly the real and fake centroids diverge in feature orientation. This gives the first candidate layer set

$$\mathcal{L}_{\text{sep}} = \{\ell \mid D_{\cos}(\ell) > \tau_{\text{sep}}\}, \tag{6}$$

where τ_{sep} is determined from the full-layer statistics. This set retains layers whose directional separation is unusually strong relative to the layer distribution of the whole backbone. As shown in Figure 3, the score is concentrated in specific middle and late layers rather than being uniformly distributed across the backbone.

To complement directional separation to fully characterize layer-wise discrepancies, we also measure the normalized centroid shift. Specifically, let $\bar{\sigma}_\ell^2$ denote the mean within-class channel variance at layer ℓ. It is used to normalize the centroid shift by the within-class variation, and its exact definition is deferred to the appendix. We then define the

Model	MovieGen		Wan 2.1		Wan 1.3B		Hailuo		Seaweed		Seedance		StepVideo		Mean	
	ACC	AP	ACC	AP	ACC	AP	ACC	AP	ACC	AP	ACC	AP	ACC	AP	mACC	mAP
RINE	52.97	71.11	45.35	40.41	33.56	32.69	51.40	47.46	51.63	46.54	44.65	38.98	70.23	69.93	49.47	46.00
FatFormer	50.02	58.84	50.00	39.06	50.17	54.96	50.00	43.80	50.00	44.55	50.00	42.16	50.23	52.36	50.07	46.15
B-Free	64.30	70.38	53.72	59.80	68.32	73.81	55.81	62.62	31.63	36.08	36.05	38.93	48.60	52.37	49.02	53.94
Effort	70.74	80.60	73.26	76.61	66.95	73.75	67.44	70.67	86.05	95.48	83.26	91.65	62.79	63.41	73.29	78.60
WaveRep	65.30	92.97	58.84	83.50	53.60	85.31	59.53	95.64	59.53	99.12	57.21	78.64	59.53	94.04	58.04	89.38
VINA	98.50	99.80	75.35	95.44	89.90	98.08	74.19	94.47	73.02	90.51	74.42	95.00	75.58	**97.09**	77.08	95.10
F3Net	92.51	96.20	71.86	85.79	69.86	78.05	66.98	80.31	74.88	92.13	73.26	87.32	66.98	81.26	70.64	84.14
TALL	91.71	96.07	64.65	83.59	67.81	80.78	61.86	78.58	64.42	88.81	63.49	82.81	63.49	83.39	64.29	82.99
NPR	92.66	97.10	71.63	86.93	66.27	82.70	72.33	87.51	74.88	92.71	73.26	92.08	72.09	90.96	71.74	88.82
TimeSformer	91.41	96.91	68.84	82.44	72.09	83.04	66.28	75.42	69.77	83.49	67.91	80.29	67.21	78.55	68.68	80.54
CLIP ViT-L/14	**99.20**	**99.95**	78.14	98.78	77.74	93.14	77.21	87.43	77.91	97.58	77.21	92.05	75.58	86.85	77.30	92.64
X-CLIP-B/16	98.55	99.87	76.28	94.94	72.43	94.07	75.12	**95.76**	75.12	87.65	75.35	**97.78**	72.33	81.65	74.44	91.98
X-CLIP-L/14	98.85	99.94	80.00	**99.62**	86.13	82.73	79.53	92.91	78.60	**99.16**	80.00	96.18	79.53	95.71	80.63	94.39
Moon-ViT	98.25	99.24	76.74	93.84	79.62	90.39	75.81	87.24	76.74	94.11	75.81	90.68	74.88	81.69	76.60	89.66
Qwen2.5-ViT	97.20	99.46	85.81	96.67	83.39	**99.68**	84.65	94.20	83.95	91.59	84.88	96.92	76.51	80.63	83.20	93.28
Ours	96.85	99.31	**93.02**	99.29	**91.78**	99.44	**87.44**	95.54	**90.47**	97.34	**89.30**	97.13	**84.19**	92.82	**89.37**	**96.92**

Table 1: **ACC and AP (%) on MovieGen and Magic Videos.** Baseline results are quoted from Li et al. (2026b), except VINA, which is evaluated by us using the released checkpoint of Li et al. (2026a). Ours denotes V-FIND with the selected LAN subspace. Best results are in bold and second best are underlined.

normalized-shift signal

$$D_{\text{shift}}(\ell) = \frac{\|\mu_{\text{real}}^{(\ell)} - \mu_{\text{fake}}^{(\ell)}\|_2}{\sqrt{M\,\bar{\sigma}_\ell^2}}, \tag{7}$$

which measures the overall real–fake displacement after controlling for within-class variation. This gives the second candidate set

$$\mathcal{L}_{\text{shift}} = \{\ell \mid D_{\text{shift}}(\ell) > \tau_{\text{shift}}\}, \tag{8}$$

where τ_{shift} is determined from the full-layer statistics. This set highlights layers that still exhibit pronounced class displacement after variance normalization. Figure 3 further shows that $D_{\text{shift}}(\ell)$ weakens in the middle stage and then increases steadily in the later layers, suggesting that these layers retain more stable real–fake displacement after accounting for within-class fluctuation.

Finally, we define the critical-layer set as the intersection of the two candidate sets:

$$\mathcal{L}^\star = \mathcal{L}_{\text{sep}} \cap \mathcal{L}_{\text{shift}}. \tag{9}$$

The intersection retains layers consistently highlighted by both signals and provides a compact search space for subsequent LAN discovery. We perform all subsequent LAN discovery only within $\mathcal{L}^\star$, while the detailed statistical settings of τ_{sep} and τ_{shift} are given in the appendix.

3.2 Latent Anchor Neuron Discovery

After obtaining the critical-layer set $\mathcal{L}^\star$, we further identify sparse latent anchor neurons within these layers. Let the full candidate neuron set in the critical layers be

$$\mathcal{N}^\star = \{(\ell, k) \mid \ell \in \mathcal{L}^\star,\, 1 \le k \le M\}, \tag{10}$$

where k indexes the latent unit in the unit-level representation of layer ℓ. Since latent units may differ in scale across layers, we score them separately within each critical layer. For each $\ell \in \mathcal{L}^\star$, we train a linear probe on the pooled representation:

$$g^{(\ell)}(x) = \mathbf{w}^{(\ell)\top}\mathbf{a}^{(\ell)}(x) + b^{(\ell)}. \tag{11}$$

Each probe is independently trained on the discovery split using a standard cross-entropy loss for real–fake classification, with its weights characterizing the discriminative alignment of individual units. For a candidate neuron $n = (\ell, k)$, let its corresponding probe coefficient be $w_n = w_k^{(\ell)}$. For an input video x_i, we define the per-sample response of this neuron as

$$r_n(x_i) = |a_n(x_i)| \cdot |w_n|. \tag{12}$$

This response combines two factors: the activation magnitude of the neuron on the current sample and the discriminative weight assigned to that neuron by the layer-specific probe. It therefore provides a unified unit-level response measure for subsequent selection.

Per-sample response alone is insufficient to determine whether a neuron can separate real and fake videos consistently, so we further measure the response separation between the two groups. For class $c \in \{\text{real}, \text{fake}\}$, we define the mean response of neuron n as

$$\mu_{n,c} = \frac{1}{|\mathcal{D}_c|} \sum_{x \in \mathcal{D}_c} r_n(x), \tag{13}$$

where σ_n^{pool} denotes the pooled standard deviation across the real and fake response groups. It serves to normalize the response gap by the within-group variation, and its exact definition is deferred to the appendix. ϵ is used to prevent division by zero. Based on these statistics, we define the separation effect size of neuron n as

$$d_n = \frac{|\mu_{n,\text{fake}} - \mu_{n,\text{real}}|}{\sigma_n^{\text{pool}} + \epsilon}. \tag{14}$$

We then use an effect-size threshold τ_d as the selection criterion and retain

$$\mathcal{S} = \{n \in \mathcal{N}^\star \mid d_n \ge \tau_d\}, \tag{15}$$

as the final LANs. Here τ_d is the core hyperparameter in the neuron-selection stage, and its effect is analyzed in the ablation study.

Method	Video Crafter1	Zero Scope	Open Sora	Sora	Pika	Stable Diff.	Stable Video	AVG
CNNDet	87.4	88.2	78.0	63.8	77.3	73.5	78.9	78.2
DIRE	55.9	61.8	53.8	60.5	65.8	62.7	69.9	62.1
MM-Det	93.5	<u>94.0</u>	88.8	86.2	95.9	<u>95.7</u>	89.9	92.0
NPR	86.6	85.6	96.0	81.0	94.6	71.1	97.0	87.4
TALL	<u>95.4</u>	91.8	97.2	94.9	97.5	83.6	98.2	92.6
F3Net	90.4	90.2	95.9	90.1	97.8	93.1	98.5	93.7
TimeSformer	94.5	92.7	98.0	92.5	98.4	92.4	99.5	95.4
Qwen2.5-ViT	93.5	**99.8**	<u>98.6</u>	<u>96.4</u>	<u>99.1</u>	95.6	<u>99.7</u>	<u>97.6</u>
Ours	**99.4**	93.2	**100.0**	**98.0**	**99.9**	**97.0**	**99.9**	**98.2**

Table 2: **Benchmarking results (AUC, %) on DVF-Test.** Baseline results are quoted from Li et al. (2026b). Ours denotes V-FIND using the selected LAN subspace. Best results are in **bold** and second-best are <u>underlined</u>.

After obtaining the selected LAN set $\mathcal{S}$, we use their raw activations in the frozen backbone to construct the final feature space, rather than the discovery-stage response scores $r_n(x)$. Specifically, we concatenate these activations into a compact forensic subspace

$$\mathbf{h}_{\mathcal{S}}(x) = [a_n(x)]_{n\in\mathcal{S}}\,, \tag{16}$$

and train a lightweight linear classifier on top of it while keeping the original detector backbone frozen.

4. Experiments And Analysis

4.1 Implementation Details

Datasets. We strictly separate neuron discovery from external evaluation throughout all experiments. The original detector is trained on the 15Model-140K dataset (Li et al. 2026b). From a smaller subset of this source distribution, we construct mutually disjoint discovery, training, and validation splits for layer localization, LAN discovery, and final linear-readout training, respectively. External evaluation is conducted on three benchmarks: Magic Videos (Li et al. 2026b), MovieGen (Polyak et al. 2024), and DVF (Song et al. 2024). Among them, Magic Videos focuses on recent high-quality generated videos, while MovieGen and DVF test generalization beyond the source dataset.

Evaluation Metrics. We report both ACC and AP for Magic Videos and MovieGen, where ACC measures standard classification accuracy and AP evaluates global ranking quality. For the DVF dataset, we report AUC to summarize detector performance across all decision thresholds.

Baselines. We compare against four categories of baselines: AI-generated video detectors, including MM-Det (Song et al. 2024), VINA (Li et al. 2026a), and WaveRep (Corvi et al. 2025); visual and video foundation backbones, including CLIP (Radford et al. 2021), X-CLIP (Ni et al. 2022), TimeSformer (Bertasius, Wang, and Torresani 2021), MoonViT (Team et al. 2025), and Qwen2.5-ViT (Li et al. 2026b); deepfake detectors, including F3Net (Qian et al. 2020) and TALL (Xu et al. 2023); and general AI-generated image detectors, including CNNDet (Wang et al. 2020), DIRE (Wang et al. 2023), NPR (Tan et al. 2024), RINE (Koutlis and Papadopoulos 2024), FatFormer (Liu et al. 2024), B-Free (Guillaro et al. 2024), and Effort (Yan et al. 2024).

Hyperparameters. Unless otherwise specified, we use the released Qwen2.5-ViT detector (Li et al. 2026b) as the frozen backbone and train only a lightweight linear classifier on the selected LAN subspace. In the main setting, we use the critical layers $\{28, 29, 30\}$ and select neurons solely by the discriminative effect size d_n with the default threshold $\tau_d = 1.5$, yielding 211 LANs. The discovery, training, and validation splits contain 2400, 2400, and 1200 videos, respectively. Each video is decoded at 2 FPS, and the center-aligned 8 frames are fed into the Qwen2.5-ViT visual encoder with dynamic resolution from 224p to 720p while preserving aspect ratio. We optimize both the layer-wise probes and the final linear classifier with AdamW using a learning rate of 1×10^{-2} and weight decay of 1×10^{-4}; the former is trained for 300 epochs and the latter for 200 epochs.

4.2 Compare With SOTA Methods

Quantitative Results. To verify whether the discovered forensic subspace supports strong detection across external benchmarks for generated videos, we conduct comparative experiments against state-of-the-art detectors on MovieGen, Magic Videos, and DVF. The results show that the compact forensic subspace delivers strong detection performance with the original backbone frozen and only a lightweight linear classifier trained. Specifically, V-FIND achieves 89.37 mACC and 96.92 mAP on Magic Videos and 96.85 ACC and 99.31 AP on MovieGen, as shown in Table 1, while improving the average AUC on DVF from 97.6 to 98.2, as shown in Table 2. Therefore, the discovered forensic subspace contains reusable forgery-discriminative knowledge that supports strong detection without full-model retraining.

Qualitative Result. To verify whether the discovered LANs consistently carry forgery-discriminative signals, we visualize their response maps on paired real and forged videos from Magic. The figure 4 shows both the response differences between real and fake videos and the activation patterns across consecutive frames. Compared with real videos, fake videos produce stronger, broader, and more continuous LAN responses around complex textures, object boundaries, repeated structures, and local visual anomalies. These responses also remain relatively stable across adjacent frames. In contrast, real videos show weaker and sparser activations. These observations suggest that LANs focus on local forgery cues. Within the forgery detector, latent anchor neurons can be organized into a compact forensic subspace, enabling the extraction of the forgery-discriminative knowledge.

4.3 Ablation Study

Layers Validation. To verify whether the critical layers carrying forgery-discriminative signals are located in the late part of the backbone, we conduct a depth-interval comparison by dividing the 32 layers into Q1 (0–7), Q2 (8–15), Q3 (16–23), and Q4 (24–31) under the same probing protocol. The results show that ACC and AP generally increase with layer depth across four representative DVF subsets, with Q4 consistently achieving the strongest performance, as shown in Figure 5. Specifically, ACC/AP increase from 79.89/67.15 to 99.34/98.74 on OpenSora and from 83.72/83.25 to 98.65/97.46 on VideoCrafter1, a trend

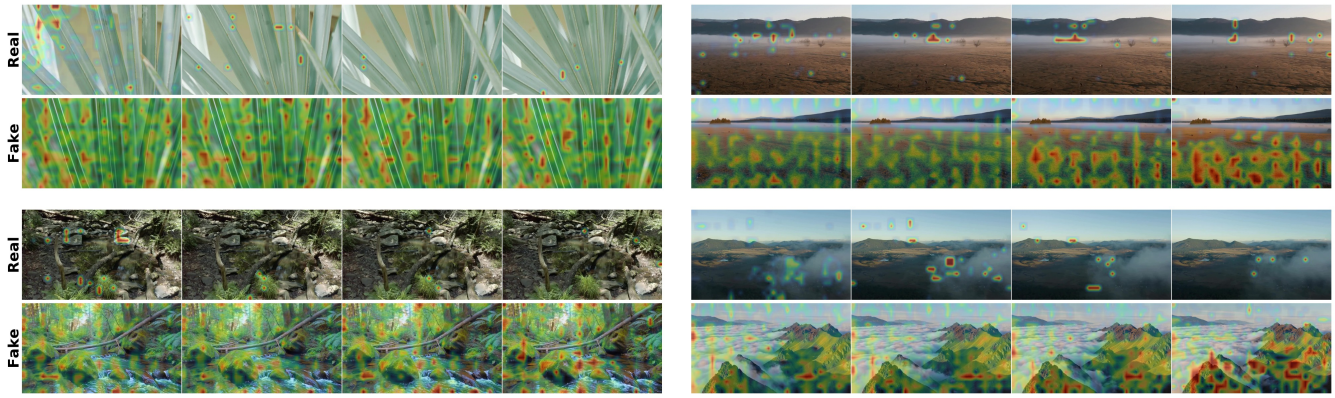

Figure 4: **Qualitative visualization of LAN responses on real and forged videos.** Response maps of latent anchor neurons (LANs) are shown for paired real and forged videos from the Magic dataset, with consecutive frames arranged from left to right.

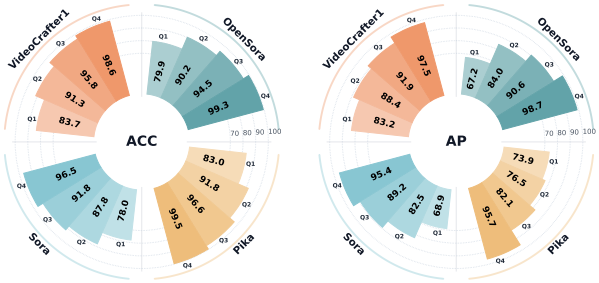

Figure 5: **Depth-wise validation.** The 32-layer backbone is partitioned into four quarters, Q1 (layers 0–7), Q2 (8–15), Q3 (16–23), and Q4 (24–31), and each interval is evaluated with the same all-channel quarter-level probe.

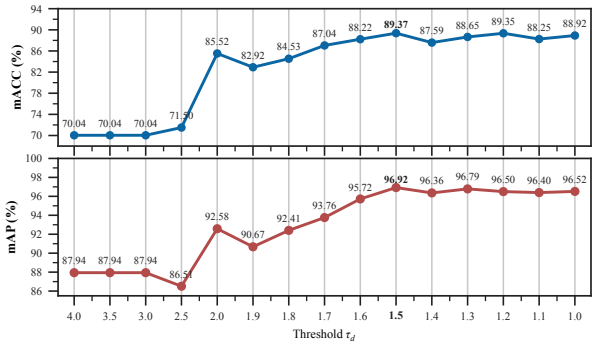

Figure 6: **Threshold sensitivity.** A coarse-to-fine sweep is used to select the default setting $\tau_d = 1.5$.

consistent with previous observations on temporal coherence and inconsistency cues (Guo et al. 2025) and with the role of deeper visual or video transformer layers in encoding more integrated temporal information (Chen et al. 2025; Shi et al. 2026b). Therefore, these results suggest late layers are more likely to carry forgery-discriminative knowledge, supporting their localization before LAN discovery.

Threshold Sensitivity. To verify whether the threshold for identifying LANs can be selected in a systematic and stable manner rather than being heuristically fixed, we conduct a coarse-to-fine sweep over τ_d. The results identify a relatively stable and consistently high-performing region around $\tau_d \approx 1.5$, as shown in Figure 6. Specifically, the coarse-to-fine search procedure first narrows the candidate range, and the subsequent fine-grained search shows that $\tau_d = 1.5$ achieves the best trade-off, reaching the highest mACC and mAP simultaneously. In contrast, larger thresholds consistently lead to degraded performance, while smaller thresholds do not provide additional improvements. Therefore, we select $\tau_d = 1.5$ based on the observed sensitivity behavior rather than an arbitrary heuristic choice.

Neuron Validation. To verify that LANs carry genuinely forgery-discriminative signals rather than simply benefiting from a constrained feature dimension, we evaluate them against several fixed-budget baselines in Figure 7. The comparison reveals two clear takeaways: first, layer localization isolates a more informative search space; second, LAN discovery successfully extracts a highly effective forensic subspace within these layers. Specifically, restricting random selection to localized layers improves Magic mACC/mAP from 78.99/89.15 to 83.04/91.60 and DVF AUC from 96.48 to 96.77. Under the same 211-dimensional budget, LANs achieve superior performance (89.37% mACC, 96.92% mAP, and 98.20% AUC), consistently outperforming both Same-layer Random and PCA. These findings confirm that the performance gains stem from isolating functionally specialized, forgery-discriminative neurons, rather than feature compactness alone.

4.4 Deeper Analysis

Cross-Architecture Validation. To verify the generalizability of V-FIND beyond Qwen2.5-ViT, we apply our LAN-discovery pipeline to three additional backbones: VINA, X-CLIP-B/16, and X-CLIP-L/14, evaluating their resulting forensic subspaces on Magic Videos (Table 3). The discovered subspaces consistently improve mACC across all three detectors, while boosting mAP on both X-CLIP variants and preserving competitive mAP on VINA. Specifically, mACC increases from 77.08% to 84.65% (+7.57%) on VINA, from 74.44% to 89.29% (+14.85%) on X-CLIP-B/16, and from 80.63% to 89.08% (+8.45%) on X-CLIP-L/14. These cross-architecture results confirm that sparse, extractable, and reusable forgery knowledge is a pervasive

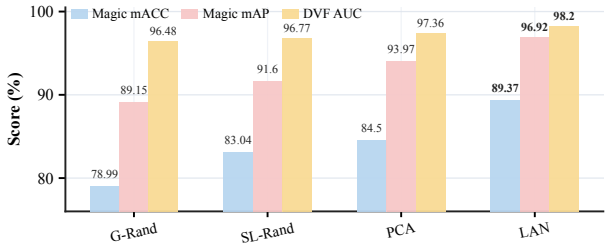

Figure 7: **Fixed-budget neuron controls.** All representations are 211-dimensional. G-Rand samples across the backbone, whereas SL-Rand preserves the LAN layer allocation.

Model	mACC (↑)	mAP (↑)
VINA	77.08	95.10
+ LAN Subspace Classifier	**84.65** (+7.6)	94.84 (-0.3)
X-CLIP-B/16	74.44	91.98
+ LAN Subspace Classifier	**89.29** (+14.9)	**96.16** (+4.2)
X-CLIP-L/14	80.63	94.39
+ LAN Subspace Classifier	**89.08** (+8.5)	**96.97** (+2.6)

Table 3: **Cross-architecture validation on Magic.** We compare each original detector with its LAN-subspace classifier; parentheses denote changes over the original detector.

property across diverse detector architectures.

Robustness to Train-Data Size. To verify whether the compact forensic subspace can support detection with limited training data, we evaluate V-FIND under different training-data sizes while keeping the original backbone frozen and training only a lightweight linear classifier. The results show that performance generally improves up to approximately 2400 samples and then exhibits only minor fluctuations, from Figure 8(a). Specifically, on Jimeng 2.0, accuracy increases from 79.07% with 300 samples to 90.23% with 2400 samples, before changing to 88.14% with 4800 samples. Therefore, these results suggest that the discovered forensic subspace can support strong detection with a relatively limited amount of training data in this evaluation setting.

Robustness Test. To verify whether the discovered forensic subspace remains effective under common input degradations, we evaluate it on Magic Videos with different levels of Gaussian blur, JPEG compression, and resizing. The results show that the relative mAP remains close to the clean baseline under mild perturbations and gradually decreases as perturbation strength increases, as shown in Figure 8(b). Specifically, the relative mAP reaches 99.62 under blur_1, 99.23 under jpeg_75, and 100.01 under resize_0.8. Therefore, these results suggest that the forgery-discriminative signals retained in the compact forensic subspace remain effective under the evaluated mild degradations.

Functional Specificity. To verify that the discovered LANs carry dedicated forgery-discriminative signals rather than arbitrary activations, we perform directional activation-swap interventions on paired real and synthetic videos, comparing them against budget-matched, same-layer random controls (Figure 9). Intervening on LANs induces substantially larger target-logit shifts and higher directional label-flip rates across both transfer directions. Specifically, LAN activation

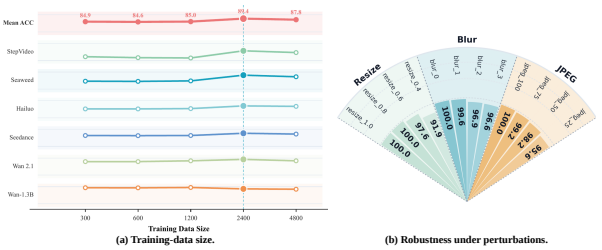

Figure 8: **Analysis of training-data efficiency and robustness. (a)** ACC (%) vs. training data size across Magic Video models using the LAN subspace. **(b)** Relative mAP (%) performance on Magic Videos under common visual perturbations (resizing, blur, and JPEG compression).

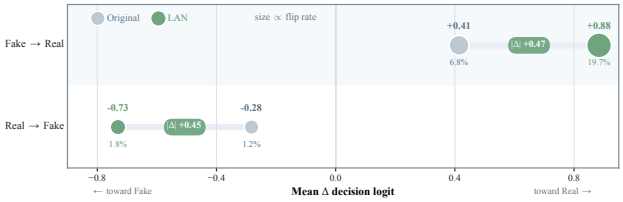

Figure 9: **Directional effects of paired activation swap.** Mean decision logit shifts (Δ) and label-flip rates under LAN vs. random neuron interventions. Random baseline preserves the same budget and layer allocation.

swaps drive logit shifts of $+0.884$ (vs. $+0.414$ for random controls) in Fake$\rightarrow$Real transfers and -0.730 (vs. -0.282) in Real$\rightarrow$Fake transfers, and increase flip rates from 6.8% to 19.7% and 1.2% to 1.8%, respectively. These causal intervention results provide direct evidence that LANs exhibit clear functional specificity for video forgery detection.

5. Conclusion

This paper studies video forgery detection from an intrinsic-knowledge perspective and proposes V-FIND, a video forgery-intrinsic neuron discovery framework. V-FIND first localizes critical layers that exhibit pronounced discrepancies between real and forged videos and then identifies latent anchor neurons that consistently carry forgery-discriminative signals, organizing them into a compact forensic subspace. With the original backbone frozen and only a lightweight linear classifier trained, this subspace delivers strong detection performance across multiple external benchmarks for generated videos, while neuron intervention experiments provide direct evidence for the functional specificity of the discovered neurons. Overall, our results suggest that forgery-discriminative knowledge is concentrated in a sparse set of functionally specialized neurons, indicating that video forgery detectors contain sparse, extractable, and reusable forgery-discriminative knowledge. Extending this neuron-level perspective to broader detector architectures and longer temporal settings remains a direction for future work.

References

Alain, G.; and Bengio, Y. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Bau, D.; Zhou, B.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6541–6549.

Bau, D.; Zhu, J.-Y.; Strobelt, H.; Lapedriza, A.; Zhou, B.; and Torralba, A. 2020. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48): 30071–30078.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is space-time attention all you need for video understanding? In *Icml*, volume 2, 4.

Bie, F.; Yang, Y.; Zhou, Z.; Ghanem, A.; Zhang, M.; Yao, Z.; Wu, X.; Holmes, C.; Golnari, P.; Clifton, D. A.; He, Y.; Tao, D.; and Song, S. L. 2025. RenAIssance: A Survey Into AI Text-to-Image Generation in the Era of Large Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3): 2212–2231.

Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; et al. 2024. Video generation models as world simulators. *OpenAI Blog*, 1(8): 1.

Cao, P.; Zhou, F.; Song, Q.; and Yang, L. 2026. Controllable Generation With Text-to-Image Diffusion Models: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(4): 4771–4791.

Chen, H.; Hong, Y.; Huang, Z.; Xu, Z.; Gu, Z.; Li, Y.; Lan, J.; Zhu, H.; Zhang, J.; Wang, W.; et al. 2024. Demamba: Ai-generated video detection on million-scale genvideo benchmark. *arXiv preprint arXiv:2405.19707*.

Chen, H.; Lin, J.; Chen, X.; Fan, Y.; Dong, J.; Jin, X.; Su, H.; Fu, J.; and Shen, X. 2025. Multimodal Language Models See Better When They Look Shallower. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 6688–6706.

Corvi, R.; Cozzolino, D.; Prashnani, E.; De Mello, S.; Nagano, K.; and Verdoliva, L. 2025. Seeing What Matters: Generalizable AI-generated Video Detection with Forensic-Oriented Augmentation. *arXiv preprint arXiv:2506.16802*.

Guillaro, F.; Zingarini, G.; Usman, B.; Sud, A.; Cozzolino, D.; and Verdoliva, L. 2024. A bias-free training paradigm for more general ai-generated image detection. *arXiv preprint arXiv:2412.17671*.

Guo, Z.; Liu, Y.; Zhang, J.; Zheng, H.; and Shan, S. 2025. Face forgery video detection via temporal forgery cue unraveling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 7396–7405.

He, Q.; Weng, J.; Tao, J.; and Xue, H. 2025. A single neuron works: Precise concept erasure in text-to-image diffusion models. *arXiv preprint arXiv:2509.21008*.

Ho, J.; Salimans, T.; Gritsenko, A. A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022. Video diffusion models. In *ICLR workshop on deep generative models for highly structured data*.

Huang, C.; Ye, Y.; Fu, B.; Su, Q.; and Shi, X. 2025. From Neurons to Semantics: Evaluating Cross-Linguistic Alignment Capabilities of Large Language Models via Neurons Alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 28956–28974.

Huang, V. S.-J.; Zhuo, L.; Xin, Y.; Wang, Z.; Wang, F.-Y.; Wang, Y.; Zhang, R.; Gao, P.; and Li, H. 2026. Tide: Temporal-aware sparse autoencoders for interpretable diffusion transformers in image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 435–443.

Huo, J.; Yan, Y.; Hu, B.; Yue, Y.; and Hu, X. 2024. Mmneuron: Discovering neuron-level domain-specific interpretation in multimodal large language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 6801–6816.

Ji, L.; Lin, Y.; Huang, Z.; Han, Y.; Xu, X.; Wu, J.; Wang, C.; and Liu, Z. 2024. Distinguish any fake videos: Unleashing the power of large-scale data and motion features. *arXiv preprint arXiv:2405.15343*.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, 2668–2677. PMLR.

Kling, A. 2024. Kling AI: Next-generation AI creative studio. *https://app. klingai. com/*.

Koutlis, C.; and Papadopoulos, S. 2024. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on computer vision*, 394–411. Springer.

Kumari, N.; Zhang, B.; Wang, S.-Y.; Shechtman, E.; Zhang, R.; and Zhu, J.-Y. 2023. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 22691–22702.

Kundu, R.; Xiong, H.; Mohanty, V.; Balachandran, A.; and Roy-Chowdhury, A. K. 2025. Towards a universal synthetic video detector: From face or background manipulations to fully ai-generated content. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 28050–28060.

Kwon, D.; Lee, S.; and Choi, J. 2025. Granular Concept Circuits: Toward a Fine-Grained Circuit Discovery for Concept Representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2356–2365.

Li, Z.; Jiang, C.; Su, L.; Shao, T.; Zhou, S.; Tao, M.; and Su, J. 2026a. Video as Natural Augmentation: Towards Unified AI-Generated Image and Video Detection. *arXiv preprint arXiv:2605.21977*.

Li, Z.; Jiang, C.; Zhao, H.; Zhou, S.; Mo, Y.; Gao, F.; Yang, F.; Shan, Q.; Wu, S.; and Su, J. 2026b. Preserving Forgery Artifacts: AI-Generated Video Detection at Native Scale. *arXiv preprint arXiv:2604.04634*.

Lim, H.; Choi, J.; Choo, J.; and Schneider, S. 2024. Sparse autoencoders reveal selective remapping of visual concepts during adaptation. *arXiv preprint arXiv:2412.05276*.

Liu, H.; Tan, Z.; Tan, C.; Wei, Y.; Wang, J.; and Zhao, Y. 2024. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, 10770–10780.

Ma, L.; Zhang, J.; Deng, H.; Zhang, N.; Liao, Y.; and Yu, H. 2024. Decof: Generated video detection via frame consistency. *arXiv preprint arXiv*, 2402.

Ni, B.; Peng, H.; Chen, M.; Zhang, S.; Meng, G.; Fu, J.; Xiang, S.; and Ling, H. 2022. Expanding language-image pretrained models for general video recognition. In *European conference on computer vision*, 1–18. Springer.

Ni, Z.-L.; Qiangyu, Y.; Yuan, T.; Huang, M.; Hu, H.; Chen, X.; and Wang, Y. 2025. Genvidbench: A challenging benchmark for detecting ai-generated video. *URL https://arxiv.org/abs/2501.11340*.

Olah, C.; Cammarata, N.; Schubert, L.; Goh, G.; Petrov, M.; and Carter, S. 2020. Zoom in: An introduction to circuits. *Distill*, 5(3): e00024–001.

Olah, C.; Mordvintsev, A.; and Schubert, L. 2017. Feature visualization. *Distill*, 2(11): e7.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.

Polyak, A.; Zohar, A.; Brown, A.; Tjandra, A.; Sinha, A.; Lee, A.; Vyas, A.; Shi, B.; Ma, C.-Y.; Chuang, C.-Y.; et al. 2024. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*.

Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; and Shao, J. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, 86–103. Springer.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Schwettmann, S.; Chowdhury, N.; Klein, S.; Bau, D.; and Torralba, A. 2023. Multimodal neurons in pretrained text-only transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2862–2867.

Shi, C.; Li, S.; Guo, S.; Xie, S.; Wu, W.; Dou, J.; Wu, C.; Xiao, C.; Wang, C.; Cheng, Z.; et al. 2026a. Where culture fades: revealing the cultural gap in text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14702–14712.

Shi, Y.; Long, Q.; Wu, Y.; and Wang, W. 2026b. Causality matters: How temporal information emerges in video language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 9006–9014.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

Song, X.; Guo, X.; Zhang, J.; Li, Q.; Bai, L.; Liu, X.; Zhai, G.; and Liu, X. 2024. On learning multi-modal forgery representation for diffusion generated video detection. *Advances in Neural Information Processing Systems*, 37: 122054–122077.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; and Wei, Y. 2024. 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*, 28130–28139.

Tang, T.; Luo, W.; Huang, H.; Zhang, D.; Wang, X.; Zhao, W. X.; Wei, F.; and Wen, J.-R. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5701–5715.

Team, K.; Du, A.; Yin, B.; Xing, B.; Qu, B.; Wang, B.; Chen, C.; Zhang, C.; Du, C.; Wei, C.; et al. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Wang, S.-Y.; Wang, O.; Zhang, R.; Owens, A.; and Efros, A. A. 2020. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8695–8704.

Wang, Z.; Bao, J.; Zhou, W.; Wang, W.; Hu, H.; Chen, H.; and Li, H. 2023. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22445–22455.

Xie, W.; Feng, Y.; Gu, S.; and Yu, D. 2021. Importance-based neuron allocation for multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5725–5737.

Xu, Y.; Liang, J.; Jia, G.; Yang, Z.; Zhang, Y.; and He, R. 2023. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 22658–22668.

Yan, Z.; Wang, J.; Jin, P.; Zhang, K.-Y.; Liu, C.; Chen, S.; Yao, T.; Ding, S.; Wu, B.; and Yuan, L. 2024. Orthogonal subspace decomposition for generalizable ai-generated image detection. *arXiv preprint arXiv:2411.15633*.

Yan, Z.; Zhang, Y.; Yuan, X.; Lyu, S.; and Wu, B. 2023. Deepfakebench: A comprehensive benchmark of deepfake detection. *arXiv preprint arXiv:2307.01426*.

Zaigrajew, V.; Baniecki, H.; and Biecek, P. 2025. Interpreting CLIP with hierarchical sparse autoencoders. *arXiv preprint arXiv:2502.20578*.

Appendix

The appendices provide additional details and complementary analyses that support and extend the main paper. Appendix A clarifies the experimental protocol and internal data-split usage. Appendix B provides deferred mathematical definitions for layer localization and LAN discovery. Appendix C presents a theoretical interpretation of layer localization and sparse LAN readout. Appendix D reports additional sensitivity analyses of the selected LAN subspace. Finally, Appendices E and F discuss methodological design choices, limitations, and future research directions.

A Experimental Protocol Details

Role of the Internal Splits. This section only supplements protocol details that are not fully expanded in the main paper. From the source detector training distribution, we construct three mutually disjoint internal splits containing 2400, 2400, and 1200 videos, respectively. The discovery split is used exclusively for layer localization and LAN scoring, the training split is used exclusively to fit the final linear classifier after neuron selection, and the validation split is used only for model selection within the source distribution. No external benchmark is involved in layer localization, neuron discovery, threshold choice, or checkpoint selection.

Validation Usage and Checkpoint Selection. The validation split is not treated as an additional reporting benchmark, but only as a model-selection set for internal choices such as threshold comparison and checkpoint ranking. All final transfer results reported in the paper are obtained after fixing these choices and evaluating on the held-out external benchmarks, including Magic Videos (Li et al. 2026b), MovieGen (Polyak et al. 2024), and DVF (Song et al. 2024). In addition, all LAN-based experiments reuse the same released Qwen2.5-ViT detector checkpoint (Li et al. 2026b) and keep the backbone frozen, so the comparison focuses on whether a compact LAN subspace can preserve forgery-discriminative knowledge without additional backbone adaptation.

Fairness Constraints in Fixed-Budget Controls. The main paper already reports the common backbone, preprocessing settings, and optimization hyperparameters, so we only clarify here the fairness constraints used in the control experiments. All fixed-budget comparisons use the same feature budget as the main setting. In particular, the same-layer random baseline preserves the layer-wise allocation of the selected LAN set, which prevents the comparison from being affected by a different layer budget or a larger effective search space. This design makes the control experiments test neuron informativeness rather than feature dimensionality alone.

B Deferred Mathematical Definitions

This section provides the exact statistical definitions that are only briefly referenced in the main paper for readability. We specifically complete the formulas used in layer localization and latent anchor neuron discovery.

Definitions for Layer Localization. The normalized-shift metric $D_{\mathrm{shift}}(\ell)$ in the main paper depends on the mean within-class channel variance at layer ℓ, which is defined as

$$\bar{\sigma}_\ell^2 = \frac{1}{2M} \sum_{c\in\{\mathrm{real,fake}\}} \sum_{k=1}^{M} \mathrm{Var}_{x\in\mathcal{D}_c}\big[a_k^{(\ell)}(x)\big]. \tag{17}$$

This quantity averages channel-wise within-class variation over both classes and all M latent units, and is used to normalize centroid displacement across layers with different activation scales. The two layer-localization thresholds are computed from the layer-wise score distributions on the discovery split:

$$\tau_{\mathrm{sep}} = \mathrm{mean}(D_{\cos}) + \mathrm{std}(D_{\cos})\,. \tag{18}$$

$$\tau_{\mathrm{shift}} = \mathrm{mean}(D_{\mathrm{shift}}) + \mathrm{std}(D_{\mathrm{shift}})\,. \tag{19}$$

The critical-layer set is then obtained by intersecting the two thresholded sets.

Definitions for LAN Discovery. The neuron separation effect size d_n uses the pooled standard deviation of the real and fake response groups. For class $c \in \{\mathrm{real}, \mathrm{fake}\}$, we define the weight-normalized within-class response variance of neuron n as

$$\sigma_{n,c}^2 = \frac{1}{|\mathcal{D}_c| - 1} \sum_{x\in\mathcal{D}_c} \left(\frac{r_n(x) - \mu_{n,c}}{|w_n|} \right)^2. \tag{20}$$

The pooled standard deviation is then given by

$$\sigma_n^{\mathrm{pool}} = \sqrt{\frac{(|\mathcal{D}_{\mathrm{real}}| - 1)\,\sigma_{n,\mathrm{real}}^2 + (|\mathcal{D}_{\mathrm{fake}}| - 1)\,\sigma_{n,\mathrm{fake}}^2}{|\mathcal{D}_{\mathrm{real}}| + |\mathcal{D}_{\mathrm{fake}}| - 2}}. \tag{21}$$

C Theoretical Analysis of V-FIND

We provide a theoretical interpretation of V-FIND in a form aligned with the actual pipeline used in the paper. The video-forgery backbone is kept frozen, layer localization restricts the candidate space, LAN discovery selects a sparse subset of activations, and only a lightweight linear readout is optimized. The analysis shows that layer localization favors layers with large dimension-normalized class separability, whereas sparse LAN selection preserves the dominant real–fake signal when the mean-gap energy is concentrated on the selected coordinates.

Problem Formulation. Let

$$\mathbf{z}(x) = [a_n(x)]_{n\in\mathcal{N}^\star} \tag{22}$$

denote the concatenated activations from the localized candidate set, and let $P_{\mathcal{A}}$ denote the coordinate projection onto a neuron subset $\mathcal{A}$. For the selected LAN set $\mathcal{S}$, the final feature and readout are

$$\mathbf{h}_{\mathcal{S}}(x) = P_{\mathcal{S}}\mathbf{z}(x), \qquad f_{\mathcal{S}}(x) = \mathbf{w}_{\mathcal{S}}^\top \mathbf{h}_{\mathcal{S}}(x) + b_{\mathcal{S}}. \tag{23}$$

We denote the class means in the localized candidate space by $\boldsymbol{\mu}_{\mathrm{real}}$ and $\boldsymbol{\mu}_{\mathrm{fake}}$, define $\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_{\mathrm{fake}} - \boldsymbol{\mu}_{\mathrm{real}}$, and let Σ be the pooled covariance. For any coordinate subset $\mathcal{A}$, a Fisher-style separability score can be written as

$$J(\mathcal{A}) = (P_{\mathcal{A}}\Delta\boldsymbol{\mu})^\top \left(P_{\mathcal{A}}\Sigma P_{\mathcal{A}}^\top\right)^{-1} (P_{\mathcal{A}}\Delta\boldsymbol{\mu})\,. \tag{24}$$

The objective is to characterize when $J(\mathcal{S})$ remains large after projecting the localized representation onto a much smaller LAN subspace.

Layer Localization. For layer ℓ, let $\boldsymbol{\delta}^{(\ell)} = \boldsymbol{\mu}^{(\ell)}_{\text{fake}} - \boldsymbol{\mu}^{(\ell)}_{\text{real}}$. Under a locally isotropic approximation $\Sigma^{(\ell)} \approx \bar{\sigma}^2_\ell I$, the dimension-normalized layer separability satisfies

$$\begin{aligned} \frac{1}{M} J_\ell &= \frac{1}{M} \boldsymbol{\delta}^{(\ell)\top} \Sigma^{(\ell)-1} \boldsymbol{\delta}^{(\ell)} \\ &\approx \frac{\|\boldsymbol{\delta}^{(\ell)}\|_2^2}{M \bar{\sigma}^2_\ell} = D_{\text{shift}}(\ell)^2. \end{aligned} \tag{25}$$

Thus, $D_{\text{shift}}(\ell)$ estimates the real–fake displacement after accounting for within-class fluctuation and layer width. The complementary score $D_{\cos}(\ell)$ further measures whether the two class centroids diverge in direction. The intersection

$$\mathcal{L}^\star = \mathcal{L}_{\text{sep}} \cap \mathcal{L}_{\text{shift}} \tag{26}$$

therefore acts as a conservative filter that retains layers with both directional separation and normalized class displacement.

Sparse LAN Readout. Within the localized candidate layers, V-FIND retains neurons with large response effect size

$$d_n = \frac{|\mu_{n,\text{fake}} - \mu_{n,\text{real}}|}{\sigma_n^{\text{pool}} + \epsilon}, \tag{27}$$

which favors coordinates whose class gap is strong relative to their weight-normalized within-group variation. Let $\bar{\mathcal{S}}$ denote the complement of $\mathcal{S}$ within the localized candidate set. If the discarded coordinates contain only a bounded amount of mean-gap energy,

$$\|P_{\bar{\mathcal{S}}} \Delta\boldsymbol{\mu}\|_2 \le \eta \, \|P_{\mathcal{S}} \Delta\boldsymbol{\mu}\|_2 , \qquad \eta < 1, \tag{28}$$

then the selected coordinates preserve a lower-bounded fraction of the total mean-gap energy:

$$\frac{\|P_{\mathcal{S}} \Delta\boldsymbol{\mu}\|_2^2}{\|\Delta\boldsymbol{\mu}\|_2^2} \ge \frac{1}{1+\eta^2}. \tag{29}$$

This condition does not require all localized channels to be retained. It only requires that the main class-discriminative displacement be concentrated on the selected LAN coordinates, which is precisely what the effect-size criterion is designed to approximate.

Consider a reference linear separator on the localized representation, $f^\star(\mathbf{z}) = \mathbf{w}^{\star\top}\mathbf{z} + b^\star$. The discrepancy between the full localized separator and its restriction to the selected LAN coordinates is bounded by

$$\left| f^\star(\mathbf{z}(x)) - \left(\mathbf{w}_{\mathcal{S}}^{\star\top} \mathbf{h}_{\mathcal{S}}(x) + b^\star \right) \right| \le \left\|\mathbf{w}^\star_{\bar{\mathcal{S}}}\right\|_2 \left\|\mathbf{z}_{\bar{\mathcal{S}}}(x)\right\|_2 . \tag{30}$$

Therefore, if the omitted-coordinate contribution is smaller than the classification margin,

$$\left\|\mathbf{w}^\star_{\bar{\mathcal{S}}}\right\|_2 \left\|\mathbf{z}_{\bar{\mathcal{S}}}(x)\right\|_2 < \left| f^\star(\mathbf{z}(x)) \right| , \tag{31}$$

the sparse projection preserves the sign of the reference decision. This gives a margin-based condition under which a frozen backbone with a lightweight readout remains effective after projection onto the selected LANs.

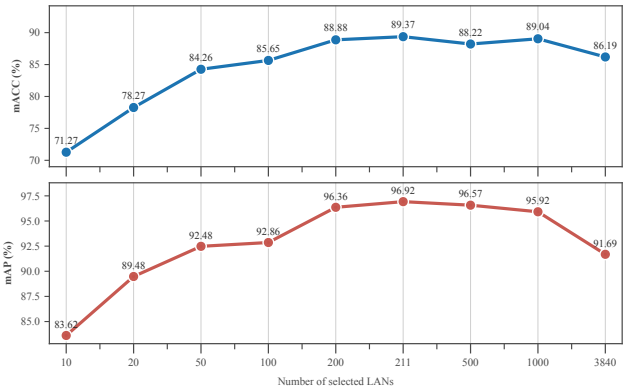

Figure 10: **Sensitivity to the number of selected LANs.** Performance improves substantially from very small subsets to moderate-size subsets, reaches its strongest region near the main setting, and drops again when all localized channels are used.

This analysis is consistent with the main observations in the paper and the appendix. Layer localization reduces the search space to a small set of discriminative layers, selected LANs outperform same-budget random neurons from the same layers, and the lightweight sparse readout maintains strong mACC and mAP with only a compact subset of activations. Together, these observations support the interpretation that V-FIND exposes a compact forensic subspace already encoded in the frozen detector. We regard this analysis as an explanatory approximation rather than a formal optimality proof for the nonlinear detector.

D More Experiments

Sensitivity to the Number of Selected LANs. We first vary the number of selected LANs while keeping the rest of the evaluation protocol unchanged. Figure 10 shows that performance improves rapidly from very small subsets to moderate-size subsets, reaches its strongest region near the main setting, and then becomes comparatively stable. By contrast, replacing the sparse readout with all localized channels gives only 86.19 mACC and 91.69 mAP on Magic, which remains below the main sparse readout (89.37 mACC and 96.92 mAP). This result supports the view that V-FIND benefits from retaining a compact discriminative subset rather than simply enlarging the feature dimension.

Sensitivity to the Discovery Set Size. We next vary the size of the source-distribution split used for LAN discovery while keeping the downstream protocol unchanged. As shown in Figure 11, the Magic performance remains within a narrow range across discovery sizes from 150 to 2400 videos. Even the 150-video setting reaches 88.83 mACC and 96.17 mAP, while the larger settings remain close to this level. These results suggest that the discovered LAN subspace is not overly sensitive to the exact discovery split size under the current pipeline.

Sensitivity to the Input Frame Count. We further test whether the readout is sensitive to the number of input frames

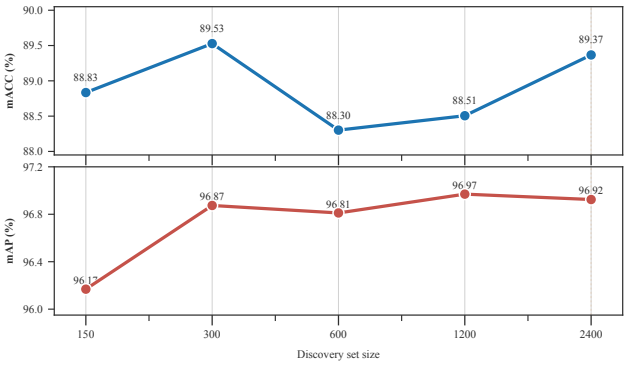

Figure 11: **Sensitivity to the discovery set size.** Magic mACC and mAP remain close across 150–2400 discovery videos, indicating that the downstream readout is relatively stable over a broad range of discovery sizes.

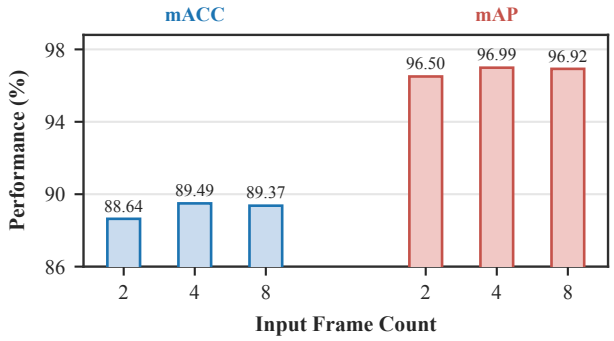

Figure 12: **Sensitivity to the input frame count.** Bars are grouped by metric: the left group compares mACC across 2, 4, and 8 frames, and the right group compares mAP across the same frame counts. The 4-frame and 8-frame settings remain very close.

used at inference time. In this analysis, only the sampled frame count is varied among 2, 4, and 8. Figure 12 reports the two metrics as separate groups, so that mACC is compared only across frame counts and mAP is compared in the same way. On Magic, the readout obtains 88.64 mACC and 96.50 mAP with 2 frames, 89.49 and 96.99 with 4 frames, and 89.37 and 96.92 with 8 frames. The 4-frame and 8-frame results remain very close, indicating that the observed performance is not tied to a single narrowly tuned frame count.

E Further Discussions

Q1: ***Why does V-FIND keep the detector backbone frozen?*** The objective of V-FIND is to examine whether a released video forgery detector already contains extractable forgery-discriminative knowledge, rather than to obtain another detector through additional end-to-end optimization. Updating the backbone would mix two effects: discovering existing internal knowledge and learning new representations from the source distribution. We therefore keep the original detector fixed during layer localization, neuron discovery, and final readout training. Under this setting, the performance of the lightweight classifier can be more directly attributed to the localized forensic subspace. This design also makes the intervention analysis more meaningful, because changes caused by manipulating selected LANs reflect the functional role of neurons already encoded in the frozen detector.

Q2: ***Why are both layer localization and LAN discovery necessary?*** The two stages address different granularities and are not redundant. Layer localization first asks where real–fake discrepancy is concentrated across the backbone, using complementary directional and shift-based signals to avoid searching all layers uniformly. However, a localized layer is still a high-dimensional representation that contains many units unrelated to the final forgery decision. LAN discovery therefore further asks which neurons within the localized layers consistently carry discriminative responses. This distinction is supported by the fixed-budget controls and the appendix sensitivity analysis: same-layer random features improve over global random features, but the selected LAN subspace remains stronger; using all localized channels also does not simply improve the result. Thus, V-FIND relies on both coarse layer filtering and fine neuron selection to isolate a compact and functional forensic subspace.

Q3: ***Is V-FIND a detector compression method or a forensic-subspace analysis method?*** V-FIND is closer to a forensic-subspace analysis method than to a detector compression method. Although the final readout uses only a compact set of LANs, the purpose is not to replace the original detector with the smallest possible model. Instead, the sparse readout serves as a controlled probe for testing whether forgery-discriminative information is already concentrated in a small internal subspace of the frozen backbone. This distinction is important because compression would primarily be evaluated by efficiency, latency, or parameter reduction, whereas V-FIND is evaluated by whether the selected neurons remain discriminative, outperform matched random controls, and produce meaningful intervention effects. Therefore, the compactness of the LAN readout should be interpreted as evidence about the organization of forensic signals inside the detector, rather than as a claim that the full detector can be universally compressed without loss.

F Limitations and Future Directions

Although our experiments cover multiple generator sources and complementary control settings, we do not conduct a systematic investigation of closely related detector variants within the same model family. Consequently, whether the localized LAN distributions and their relative layer-wise concentration remain fully consistent across detector checkpoints, training scales, and backbone variants has yet to be established. Future work will further examine the intra-family consistency and transferability of LANs across related video forgery detectors.