ViDToken: A Video-Transformer-Based Latent Token Defense for Adversarial Video Detection

Wei Song
University of New South Wales, Australia
wei.song1@unsw.edu.au

Zhenchang Xing
CSIRO's Data61, Australia
zhenchang.xing@data61.csiro.au

Liming Zhu
CSIRO's Data61, Australia
liming.zhu@data61.csiro.au

Yulei Sui
University of New South Wales, Australia
y.sui@unsw.edu.au

Jingling Xue
University of New South Wales, Australia
j.xue@unsw.edu.au

***Abstract*—We introduce VIDTOKEN, a transformer-based pre-inference defense against adversarial video attacks on Video Recognition Systems (VRSs). VIDTOKEN leverages spatial and temporal encoders to tokenize video frames and select a representative frame token (RTS), effectively distinguishing adversarial from clean videos. This approach secures both CNN- and transformer-based VRSs and outperforms existing defenses by neutralizing three advanced attack types across major video classifiers, achieving high detection rates, low false positives, and manageable overhead. To counter adaptive sparse attacks that alter only one frame, we propose a frame-replication countermeasure that significantly improves performance, positioning VIDTOKEN above nearly all defenses. Although this increases inference times, they remain under 1 second—comparable to modern facial recognition authentication systems. This study provides key insights into advancing transformer-based adversarial defenses and calls for further research.**

***Index Terms*—Machine Learning Security, Adversarial Video Attacks, Adversarial Video Defenses, Video Recognition Systems**

I. INTRODUCTION

Video Recognition Systems (VRSs) are deep neural networks that process sequential visual frames to produce categorical outputs for tasks such as face recognition [1], object detection in autonomous vehicles [2], [3], behavioral analysis in security systems [4], [5], and anomaly detection in surveillance [5], [6]. Architectures such as LRCN [7] and C3D [8] combine convolutions with temporal modeling; CNN+LSTM [9] integrates spatial and recurrent processing; two-stream networks like TSN [10] fuse RGB and flow; and inflated models such as I3D [11] set new benchmarks. However, DNNs are vulnerable to adversarial attacks [12]–[17], which introduce subtle, hard-to-detect perturbations that mimic natural inputs. The complexity of video data, spanning both spatial and temporal domains, allows adversaries to craft sophisticated attacks, increasing the challenge of defending VRSs against these threats [17]–[19].

Current defenses against adversarial attacks in VRSs, whether image-centric [20]–[24] or video-centric [19], [25]–[29], typically require model access/modifications or rely on resource-intensive adversarial training [22], [23], which lacks generalization across attack types. Image-centric defenses often overlook video-specific temporal dynamics [12], [13], [19], [29], and video-centric methods using optical flow [25], [26] falter against attacks maintaining temporal consistency [12]–[14]. Defenses tailored to specific classifiers or perturbations [27], [28] also exhibit limitations. This underlines the urgent need for more effective defenses in VRSs.

This paper initiates an exploration of transformer-based defenses against adversarial attacks on VRSs, delving into their strengths and limitations. Adversarial examples, derived through complex non-linear and non-convex optimization problems in DNNs [18], [30], [31], present formidable challenges. The absence of robust theoretical models to describe these optimization solutions amplifies the difficulty of defense over attack. Through this study, we aim to catalyze further research that strengthens transformer-based defenses, thereby advancing the broader field of machine learning defenses.

We introduce VIDTOKEN, a transformer-based pre-inference approach that secures both CNN- and transformer-based VRSs. VIDTOKEN builds on two key insights: (1) the limited transferability of adversarial samples from CNN-based VRSs to transformer models [32], [33] and (2) the strategic targeting of CLS-token perturbations in transformer-based attacks [34]–[36]. By tokenizing video frames with spatial and temporal encoders, VIDTOKEN distinguishes adversarial from clean videos. Its Representative Token Selection (RTS) strategy identifies the most indicative frame token, enhancing detection without accessing downstream models. Using tokenized frames and a one-class detector to analyze token distributions, VIDTOKEN efficiently detects adversarial videos across VRS architectures without relying on adversarial training.

VIDTOKEN neutralizes three advanced adversarial attacks—U3D [13], Geo-Trap [14], and StyleFool [12]—crafted under their respective protocols, outperforming existing defenses [21], [23], [25], [27], [37], [38]. It effectively protects two CNN-based classifiers (C3D [8], I3D [11]) and one transformer-based (VTN [39]) across the UCF-101 [40] and HMDB-51 [41]. VIDTOKEN achieves a 93.4% detection rate, keeps false positives at or below 1.3%, and increases inference

times by only 2–4× relative to unprotected VRSs.

Under adaptive sparse attacks [42] targeting the transformer's attention mechanism, VIDTOKEN's detection rates fall below 30%, while inference times remain 2–4×. The attention-focused design of VIDTOKEN is less effective with sparse perturbations, especially when only a single frame is altered. In response, we introduce a novel frame-replication countermeasure that boosts the average detection rate to 74.5%, surpassing nearly all existing defenses. While inference times can peak at 2–16× in the worst cases, false positives remain at or below 2.7%. This countermeasure keeps inference times under 1 second, aligning with the typical authentication times of around 1 second for real-world video applications such as face recognition systems [43]–[45]. Additionally, with frame replication, VIDTOKEN continues to outperform existing defenses [21], [23], [25], [27], [37], [38], achieving a 90.3% detection rate (down from 93.4%), with false positives slightly increasing from 1.3% to 2.7%.

In summary, this paper makes the following contributions, underscoring VIDTOKEN's superiority and providing key insights for advancing transformer-based defenses:

- **Adversarial Video Detection.** We present a transformer-based pre-inference framework, VIDTOKEN, to protect both CNN- and transformer-based VRSs. It operates independently, requiring no access or modifications to downstream classifiers, avoiding costly adversarial training [19], [29]. Unlike reconstruction methods [21], [46]–[48], VIDTOKEN directly detects adversarial videos while maintaining clean video recognition in downstream classifiers.
- **Representative Token Selection (RTS).** Our *RTS* approach identifies a single frame token, essential for capturing a video's key spatio-temporal attributes. VIDTOKEN's detection rate would drop by 71.8% with a non-representative frame token and by 41.9% with the CLS token in attacks on transformer-based VRSs.
- **Effective Defenses against Advanced Attacks.** Our approach surpasses existing defenses [21], [23], [25], [27], [37], [38] by effectively neutralizing existing advanced attacks [12]–[14] with low false positives, though at the cost of increased inference times.
- **Countermeasures for Adaptive Attacks.** Our frame-replication mitigation positions VIDTOKEN ahead of most defenses against adaptive sparse attacks, slightly increasing false positive rates. Inference times remain under 1 second, similar to modern facial recognition systems [43]–[45].

II. Ethical Consideration and Open Science

VIDTOKEN enhances VRS security by detecting adversarial videos, mitigating harm in urban surveillance and public safety. We have made all related resources—datasets, scripts, and source code—publicly available at https://github.com/anonymous-vidtoken/vidtoken to foster open science and ethical advancement in cybersecurity. This initiative supports the development of robust defenses and ensures advancements respect user privacy and data integrity.

III. Background and Motivation

A. Adversarial Video Attacks

Adversarial videos, designed to deceive VRSs, manipulate predictions to diverge from human perception. Created by adding perturbation $\mathcal{P}$ to an original video ($x^{adv} = x + \mathcal{P}$), these examples stay undetectable by constraining the perturbation magnitude ϵ with the ℓ_p-norm ($||x^{adv} - x||_p \leq \epsilon$), where p defines the norm type. The attacker aims for x^{adv} to either switch from the original class y_0 in untargeted attacks or meet a specific target class y_t in targeted attacks, while minimizing the queries to estimate the model's gradients. We focus on three state-of-the-art adversarial video attacks—U3D [13], Geo-Trap [14] and StyleFool [12].

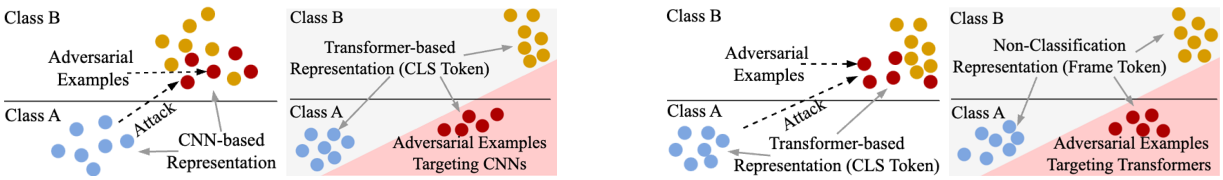
U3D [13]. This leading untargeted attack uses Perlin or Gabor noise under the ℓ_∞-norm, achieving an 87.8% success rate across video classifiers. It targets common DNN vulnerabilities, applying universally without specific customization. U3D incorporates temporal loss to smooth adversarial videos, enhancing broad applicability.

Geo-Trap [14]. A black-box framework for both untargeted and targeted attacks, Geo-Trap optimizes adversarial videos using gradients from geometric transformations within the ℓ_∞-norm. It distributes perturbations smoothly across frames, preserving temporal coherence and evading defenses by relying on temporal consistency.

StyleFool [12]. Unlike U3D and Geo-Trap, which rely on ℓ_p-norm constraints, employs stylistic transformations that preserve video semantics. It modifies stylistic attributes such as texture and color to all frames, bypassing the ℓ_p-norm's focus on pixel-level changes. Like U3D, StyleFool also incorporates a customized temporal loss to preserve temporal coherence in adversarial videos.

B. Defense Mechanisms

Image-Centric Defenses. In our context, image-centric defenses are considered as frame- or spatio-centric approaches. Adversarial training [22], [23] enriches datasets with adversarial examples [17], requiring extensive examples and retraining. ADDNP [16] uses discrepancies between clean and adversarial images for detection, relying on adversarial examples for learning. Input transformation [49] employs cropping and resizing, potentially impacting accuracy. ComDefend [21] and [46] use reconstruction to enhance image fidelity and defense effectiveness [12], but may reduce accuracy due to quality distortion. DiffPure [47] utilizes a diffusion model for image reconstruction, incurring high computational costs [50]. Techniques like random smoothing [38] and PixelDP [24] modify the model's architecture, which may compromise its integrity. Blacklight [51] identifies adversarial queries but is less effective against attackers using a shadow model.



(a) CNN-based Attacks (b) Transformer-based Attacks

Fig. 1: Two known insights exploited in designing VIDTOKEN: (1) the negligible transferability of adversarial examples targeting CNN-based VRSs to transformer models, and (2) targeted attacks on the CLS token in transformer-based VRSs.

Video-Centric Defenses. Video-specific defenses address the limitations of image-centric approaches that often miss temporal dynamics [12], [13], [19], [29], [48]. AdvIT [25] and the work reported in [26] detect anomalies by leveraging temporal consistency, assuming smooth transitions in genuine videos versus disrupted flow in adversarial ones. However, these are challenged by attacks like [13], [14], and [12], which maintain temporal flow with custom temporal loss adjustments. Techniques like JPEG compression and optical texture analysis [52] defend effectively against large, textured objects. Recent methods like OUDefend [27], using tailored feature restoration, and DP [28], customizing defense patterns based on prior examples, are effective but limited by their dependency on specific video classifier architectures or perturbation types. While adversarial training, initially developed for images [22], [23], has been adapted for videos [19], [29], it continues to incur high retraining costs and necessitates extensive adversarial examples.

C. Video Transformers

Originally transformative in NLP, the transformer model [53] has evolved into video transformers, effectively capturing long-range dependencies in sequences. This strength is evident in video analysis [39], [54], [55], where their self-attention mechanism evaluates and weighs all sequence positions as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

where Q, K, and V are the query, key, and value vectors derived from the input, and d_k is the dimensionality of K.

D. This Work

VIDTOKEN acts as a transformer-based, pre-inference framework for detecting adversarial videos. It protects both CNN- and transformer-based VRSs by integrating spatial and temporal dimensions, surpassing traditional image-centric and video-centric defenses [16], [21], [25], [26], [46], [47], [49]. Unlike adversarial training [19], [22], [23], [29] and ADDNP [16], VIDTOKEN utilizes a one-class detector [56]–[58], eliminating the need for model access, modifications, or adversarial examples. Distinct from reconstruction methods [21], [46]–[48], VIDTOKEN specializes in direct adversarial video detection and ensures that clean videos not flagged as adversarial maintain their recognition accuracy when passed to downstream classifiers.

IV. THE VIDTOKEN DESIGN

A. Threat Model

Attackers typically possess full knowledge of the target model's architecture, parameters, gradients, and training data. They deploy various strategies such as targeted and untargeted, black-box and white-box approaches, across different perturbation intensities. In defense, a detection layer is incorporated before inference to identify adversarial videos while ensuring accuracy on legitimate inputs, developed independently of the attackers' methods, algorithms, or constraints. We address three main attacks: U3D [13], Geo-Trap [14], and StyleFool [12], originally targeting CNN-based classifiers but now also adapted for transformer-based systems, evaluating VIDTOKEN's effectiveness (Sections VI-A and VI-B).

We thoroughly analyze potential adaptive attacks, including sparse and GAN-style attacks. Adaptive sparse attacks [42] perturb a single video frame, challenging VIDTOKEN's attention mechanism due to sparse perturbation reinforcement issues. Additionally, adaptive GAN-style attacks [16], [25], [59], common in adversarial defenses, aim to infer RTS tokens used by VIDTOKEN and deceive downstream video classifiers using the GAN architecture [60].

B. Design Rationale

VIDTOKEN is a transformer-based pre-inference framework that secures both CNN- and transformer-based VRSs. It relies on two insights: (1) adversarial examples for CNN-based VRSs transfer poorly to transformer models [32], [33], and (2) transformer-based VRSs are susceptible to targeted CLS token attacks [39], [61]–[63], as illustrated in Figure 1.

Figure 1(a) illustrates the rationale behind designing VIDTOKEN to counter adversarial attacks on CNN-based VRSs. It shows how attackers aim to shift classifications for some clean videos from class A (blue dots) to class B (red dots) through adversarial perturbations. However, prior studies [32], [33] reveal these perturbations have limited transferability to transformers, which can still classify them accurately despite noise-induced shifts in video distribution. Unlike CNNs that focus on local feature extraction, transformers utilize self-attention mechanisms to process data globally, enabling them to maintain distinct representations for adversarial and clean videos. Leveraging this, VIDTOKEN utilizes transformer-specific latent token representations for video frames to effectively differentiate between adversarial and clean videos.

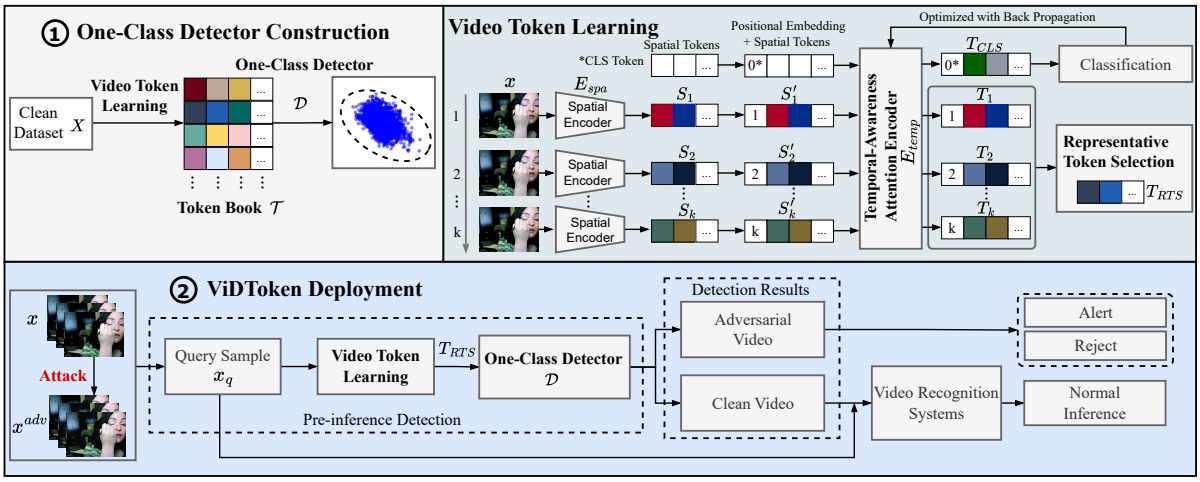


Fig. 2: Overview of the VIDTOKEN framework, which comprises two main phases: (1) One-Class Detector Construction via video token learning from a clean dataset, resulting in the formation of the Token Book $\mathcal{T}$; and (2) VIDTOKEN Deployment with VIDTOKEN utilized as a pre-inference layer to separate clean from adversarial video inputs.

Figure 1(b) extends the rationale to transformer-based VRSs, illustrating how attackers manipulate the CLS token—the video's global classification representation [39], [61]–[63]—to cause misclassifications. In this scenario, frame tokens from adversarial videos, which are not used for classification, differ from clean video tokens due to perturbations. VIDTOKEN, a transformer-based framework itself, leverages these differences by utilizing frame tokens to detect adversarial videos, specifically focusing on selecting the most representative frame token (RTS) for video identification.

We also address potential adaptive sparse attacks on VIDTOKEN's attention mechanism by employing a frame-replication countermeasure that preserves its effectiveness against state-of-the-art defenses.

C. The VIDTOKEN Architecture

VIDTOKEN, as illustrated in Figure 2, consists of two phases: One-Class Detector Construction and VIDTOKEN Deployment, where VIDTOKEN functions as a pre-inference detection layer. Central is the video token learning methodology, further detailed in Section V.

One-Class Detector Construction. We begin by building a latent token book $\mathcal{T}$ exclusively with clean videos via the video token learning outlined in Section V. This process encodes video frames, capturing their spatial attributes and integrating them within the temporal context. To boost VIDTOKEN's resilience against attacks targeting CNN- and transformer-based classifiers and minimize computational costs, we adopt a novel RTS (Representative Token Selection) strategy, inspired by key frame analysis techniques [42], [64]. Our approach selects a critical frame token from each video that encapsulates its essential spatio-temporal attributes, as detailed in Section V-C. The video token learning module processes each video in the dataset X, distilling spatial and temporal information into frame tokens. From these, we select and link a representative token $\mathcal{T}_{RTS}$ to its video in the latent token book $\mathcal{T}$.

Like [25], [59], to distinguish clean from adversarial examples in this abstract token space, we develop a one-class detector $\mathcal{D}$, utilizing principles from one-class classifiers [56]–[58]. $\mathcal{D}$ is trained on $\mathcal{T}$ to recognize the pattern of clean video tokens, thus identifying adversarial videos that diverge from the clean video distribution. The operating principle of $\mathcal{D}$ is simple: $\mathcal{D}(\mathcal{T}_{RTS}) = 1$ indicates a clean video and $\mathcal{D}(\mathcal{T}_{RTS}) = -1$ signifies an adversarial video.

VIDTOKEN Deployment. VIDTOKEN acts as a pre-inference layer to a VRS. Incoming videos are analyzed by VIDTOKEN's video token learning module (Section V) to identify a representative token T_{RTS}. This token is evaluated by the one-class detector; clean videos continue to standard VRS inference, while adversarial ones are flagged for alert or rejection. It operates independently from the VRS, ensuring the defense remains invisible and non-intrusive.

V. THE VIDTOKEN METHODOLOGY

At VIDTOKEN's core (shown in Figure 2) is the video token learning module. This module uses the video transformer network [39] to extract frame tokens through three key components: (1) a Spatial Encoder, (2) a Temporal-Awareness Attention Encoder, and (3) RTS (Representative Token Selection), described below.

A. Spatial Encoder

In VIDTOKEN, inspired by recent advances in video recognition [39], [55], [62], [63], we treat a video as a series of sequential frames, with each frame serving as a distinct spatial entity. This approach is facilitated by a spatial encoder E_{spa},

formulated as $S_i = E_{\text{spa}}(x_i)$, producing a feature vector for the i-th frame x_i. Selecting an appropriate encoder architecture is critical to ensure a balance between computational efficiency and adversarial detection capabilities. We consider the Vision Transformer (ViT) [65] given the negligible transferability of adversarial attacks from CNNs to transformers (Section IV-B), but also evaluate VIDTOKEN using CNN architectures like ResNet50 [66] and ResNet101 [66] (Section VI-D).

B. Temporal-Awareness Attention Encoder

To capture the temporal dynamics in a video alongside its spatial characteristics, a temporal-awareness attention encoder E_{temp} is employed. Each spatial token S_i is enhanced with a positional embedding PE_i, leading to $S'_i = PE_i + S_i$. In line with practices from [39], [61], we adopt the absolute frame position as its positional embedding. Such an approach allows E_{temp} to effectively understand the sequential order of frames within the video [39], [39], [61], facilitating a coherent interpretation of temporal information.

To convert each spatially enhanced token S'_i into a temporally enriched token T_i, we use E_{temp} with Longformer [67]. Unlike standard transformers, Longformer uses localized self-attention, focusing on $\frac{w}{2}$ adjacent tokens each side, where w is the window size. This approach handles long sequences more efficiently by addressing the standard transformer's issue of quadratic complexity. Longformer's sliding window attention, which scales linearly, allows VIDTOKEN to effectively capture temporal dynamics in long video sequences [39], [54], [55], boosting computational efficiency. Additionally, Longformer's multi-head attention, derived from the standard transformer, enhances E_{temp}'s parallel processing of video sequences, enabling VIDTOKEN to analyze complex spatio-temporal dynamics effectively [39], [53], [67].

To train E_{temp}, we follow Algorithm 1. For each video x, we begin by generating an embedding token array S' containing spatial representations S'_i of its k frames x_i (where $i \in [1..k]$), each enhanced with their positional embeddings (lines 4-11). Notably, S_0 is reserved for the CLS token. The training of E_{temp} is denoted by $T = E_{temp}(S')$ (line 12), where Longformer [67] transforms S'_i into T_i with enriched temporal information due to self-attention, as illustrated in Figure 2. We employ cross-entropy loss for classification (line 15), focusing on the CLS token (T_{CLS}), which is standard for classification tasks [61] (lines 13-14). This approach aligns with established video transformer networks [39], [54], [55]. Key training and system parameters, such as the number of epochs N, window size w, and the attention heads H, are detailed in Section VI-A.

C. Representative Token Selection

Selecting optimal frame tokens to differentiate clean and adversarial videos is challenging. The CLS token in video transformer networks [39], [54], [55] offers a global video representation but is vulnerable to targeted attacks, compromising its reliability. Moreover, using multiple frame tokens can significantly increase VIDTOKEN's inference times; for instance, using three tokens can quintuple the time compared to one. Selecting a non-representative frame also reduces detection effectiveness. All these challenges underscore the importance of careful token selection, as evaluated in Section VI-D.

We tackle these challenges with Representative Token Selection (RTS), which strengthens detection against advanced attacks (including adaptive ones on transformer-based video classifiers) and enhances efficiency. Our approach selects the most indicative frame token (the ***RTS token***) by analyzing attention from adjacent frames. Building on VIDTOKEN's temporal-awareness attention encoder—based on Longformer [67] with local windowed self-attention—the RTS token integrates context from nearby frames [39], [64], capturing the video's essential spatio-temporal attributes.

After training E_{temp} with Algorithm 1, we derive the spatio-temporal representation T_i for each frame x_i in a video comprising k frames ($i \in [1..k]$), as shown in Figure 2. Let Q_{T_i} and K_{T_i} be the query and key vectors, respectively, generated in the last decoder layer of Longformer [67] for T_i. The *cumulative attention score* for the i-th token, aggregating attention scores between this frame and others in the same sliding window w, is computed as follows:

$$A_i = \sum_{j=i-\frac{w}{2}}^{i+\frac{w}{2}} \left(\frac{Q_{T_i} K_{T_j}^T}{\sqrt{d_{K_{T_j}}}} \right) \tag{2}$$

The token with the highest cumulative attention score is chosen as the comprehensive representation of the video:

$$T_{RTS} = T_i \quad \text{where } i = \arg \max_{j \in [1..k]} A_j \tag{3}$$

Thus, our approach intuitively selects the frame that most effectively encapsulates the spatio-temporal attributes of the video, enhancing the detection of any adversarial content.

In Figure 2, the token book $\mathcal{T}$, displayed in the top-left corner, links each video with its corresponding RTS token.

D. The VIDTOKEN-Protected Inference

As shown in Figure 2, the video token learning module is central to VIDTOKEN, enabling the one-class detector's construction and integration as a pre-inference layer in a VRS. Algorithm 2 details the VIDTOKEN workflow for detecting adversarial videos. Upon receiving a video x_q, VIDTOKEN processes its frames T using pre-trained E_{spa} and E_{temp} encoders (lines 1–9), identifies the representative token T_{RTS} (lines 10–11) following Equations 2 and 3, and then $\mathcal{D}$ evaluates T_{RTS} to determine if x_q is adversarial or clean.

VI. EVALUATION

Our evaluation addresses four key research questions:

- **RQ1:** How effective is VIDTOKEN compared to leading defenses [21], [23], [25], [27], [37], [38] in detecting advanced video attacks on both CNN- and transformer-based VRSs?
- **RQ2:** What is its computational efficiency?
- **RQ3:** How do setting adjustments affect performance?

Algorithm 1: Training of E_{temp}.

Input: Dataset X, pre-trained E_{spa}, epochs N, heads H, window size w.
Output: Trained parameters Θ for E_{temp}.

Initialize Θ for E_{temp} with w, H
for *epoch* $= 1$ *to* N **do**
  for *video* x *in* X **do**
    Initialize empty embedding token array S'
    Initialize the CLS token S_{CLS} for x
    $S'_0 \leftarrow PE_0 + S_{CLS}$
    Append S'_0 to S'
    for *frame* x_i *in video* x **do** // $i \in [1..k]$
      $S_i \leftarrow E_{spa}(x_i)$
      $S'_i = PE_i + S_i$
      Append S'_i to S'
    $T \leftarrow E_{temp}(S')$
    $T_{CLS} \leftarrow T_0$ // Extract CLS token
    $\mathcal{L} \leftarrow \texttt{classification}(T_{CLS})$
    $\Theta \leftarrow$ Minimize $\mathcal{L}$

Algorithm 2: VIDTOKEN as a pre-inference layer.

Input: Query video x_q, trained encoders E_{spa} and E_{temp}, and trained one-class detector $\mathcal{D}$.

Initialize empty embedding token array S'
Initialize the CLS token S_{CLS} for x^q
$S'_0 \leftarrow PE_0 + S_{CLS}$
Append S'_0 to S'
for *each frame* x_i *in video* x_q **do** // $i \in [1..k]$
  $S_i \leftarrow E_{spa}(x_i)$
  $S'_i = S_i + PE_i$
  Append S'_i to S'
$T \leftarrow E_{temp}(S')$
$T' \leftarrow T[1..k]$ // Exclude CLS token
$T_{RTS} \leftarrow \texttt{rep\_token\_selection}(T')$
if $\mathcal{D}(T_{RTS}) = -1$ **then** // x_q is adv video
  Alert or reject x_q
else // x_q is clean video
  Perform inference on x_q using VRS

- **RQ4:** How effectively does VIDTOKEN withstand adaptive attacks, including sparse attacks targeting its temporal-awareness attention mechanism and the well-known GAN-style attacks?

A. Experiment Setup

Video Recognition Systems (VRSs). We assess three video classifiers, with two CNN-based and one transformer-based:

- *CNN-based classifiers:* C3D [8] and I3D [11] are known for their exceptional classification abilities [12], [13]. C3D leverages 3D convolution to extract spatio-temporal features, while I3D utilizes optical flow to parse frame relationships.
- *Transformer-based classifier:* The Video Transformer Network (VTN) [39], built on transformers, has proven its outstanding effectiveness in classifying videos by leveraging transformer-based spatial and temporal encoders [63].

TABLE I: Adversarial videos created under the default attack setting for U3D, Geo-Trap, and StyleFool (Section III-A) targeting C3D, I3D, and VTN on UCF-101 and HMDB-51.

Attack Method	C3D		I3D		VTN	
	UCF-101	HMDB-51	UCF101	HMDB-51	UCF101	HMDB-51
U3D [13]	602	602	301	301	161	251
Geo-Trap [14]	76	53	97	63	14	35
StyleFool [12]	109	32	144	206	45	27

Datasets. We utilize UCF-101 [40] and HMDB-51 [41], two widely used datasets for video analysis. UCF-101, sourced from YouTube, contains $13,320$ videos across 101 classes. HMDB-51, collected from varied sources, comprises $6,849$ videos in 51 classes, encompassing activities such as sword fighting and climbing. Like previous studies [8], [12], [14], we segment all videos into 16-frame chunks. Both datasets are used to train the video classifiers and VIDTOKEN.

Attack Settings. We consider three state-of-the-art attacks—U3D [13], Geo-Trap [14], and StyleFool [12]—configured as untargeted black-box attacks. Following these earlier works, we set $\epsilon = 8/255$ by default, balancing attack strength with human imperceptibility. Recall that ϵ measures perturbation intensity on clean videos: higher values on an 8-bit scale yield more conspicuous alterations, though subtle changes are often preferred. We also test variations: targeted Geo-Trap (T) by selecting random non-original labels [14], white-box StyleFool (W), and differing perturbation intensities.

Adversarial Video Examples. To create adversarial examples for C3D and I3D, we used open-source frameworks for U3D [13], Geo-Trap [14], and StyleFool [12], following their protocols. For VTN, we correspondingly applied these attacks. Geo-Trap and StyleFool perturb all frames in a video, while U3D perturbs only 25%. We created 3,119 examples across UCF-101 and HMDB-51, as detailed in Table I, targeting these classifiers under the default attack setting. The decreasing example counts for C3D, I3D, and VTN suggest increasing difficulty in compromising these classifiers. To evaluate defenses, we ensured that classifiers correctly identified original videos while misclassifying their adversarial counterparts [68].

Adversarial examples are generated against video classifiers and then detected by defenses.

Baselines. AdvIT [25], the pioneering adversarial video detection method, leverages temporal consistency by estimating flow between a target frame and its previous m frames to generate a pseudo frame and calculate an inconsistency score c, with higher scores indicating adversarial content. We set $m = 1$ due to negligible performance differences

[13] and used SpyNet [69] for flow estimation. OUDefend [27] employs over/undercomplete features in a restoration network, following its official implementation. Additionally, we adapted four image-based defenses for video—AT [23], IT [37], ComDefend [21], and RS [38]. For AT, we adhered to its guidelines, for IT and ComDefend, we utilized their open-source codes, and for RS, we directly applied Gaussian noise within the ℓ_2-norm from its available implementation.

Computing Platform. We conduct all experiments on one NVIDIA RTX 4090 card with 24GB RAM and CUDA 12.0.

VIDTOKEN Training Protocols. VIDTOKEN training involves two components: the video token learning module and the one-class detector. The learning module uses a pre-trained ViT from HuggingFace [70] as the spatial encoder (E_{spa}) and a Longformer-based temporal encoder (E_{temp}) with three encoder and decoder layers, configured with $w = 8$ and $H = 8$ (trade-offs in Section VI-D). E_{temp} is trained with Adam [71] at a 0.001 learning rate and cross-entropy loss on the CLS token for 50 epochs, typically converging by epoch five. The one-class detector $\mathcal{D}$ uses OC-DNN [56] to learn normal data patterns, implemented via its open-source code [72].

B. *RQ1:* VIDTOKEN*'s Detection Capabilities*

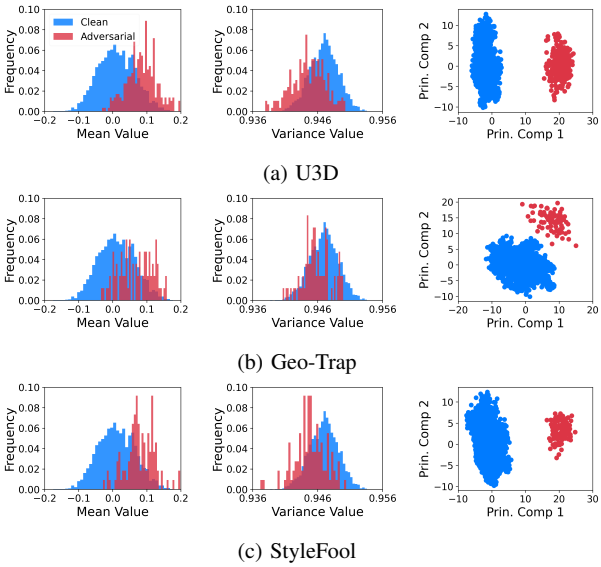
For a given defense solution, we define its *Detection Success Rate (DSR)* as the ratio of adversarial videos correctly identified to the total number of adversarial examples. VIDTOKEN intercepts adversarial videos in its pre-inference module before they reach the downstream classifier (Figure 2). For defenses like RS, IT, ComDefend, OUDefend, and AT, success is measured by whether an altered adversarial video is correctly classified as its original category.

We present RQ1 via four tables and four figures. Tables II and III report VIDTOKEN's DSRs against untargeted black-box attacks and false positive rates (FPRs) on clean videos, respectively. Table IV extends this to Geo-Trap (T), a targeted black-box attack, and StyleFool (W), an untargeted white-box attack, while Table V evaluates robustness under varying perturbation intensities. Figures 3 to 5 (main text), together with Figure 8 (appendix), analyze mean, variance, and PCA across these scenarios to illustrate and interpret VIDTOKEN's performance across scenarios.

Below we first assess VIDTOKEN's detection capability in the default attack setting before exploring its variations.

Default Attacking Setting. We used untargeted black-box adversarial videos with $\epsilon = 8/255$ for U3D [13], Geo-Trap [14], and StyleFool [12], as detailed in Section III-A. VIDTOKEN outperformed six defenses in protecting the C3D, I3D, and VTN classifiers, achieving an average DSR of 93.4% across all attack-classifier-dataset combinations while FPRs for clean videos (0.8% for UCF-101 and 1.3% for HMDB-51). Our discussion of the results, presented in Tables II and III, begins with the CNN-based classifiers C3D and I3D, followed by the transformer-based VTN.

C3D and I3D. Table II shows VIDTOKEN's superior DSRs over six baselines in adversarial video detection across three



(a) U3D

(b) Geo-Trap

(c) StyleFool

Fig. 3: Mean values, variances, and PCA of RTS tokens for UCF-101 clean videos and adversarial videos from U3D, Geo-Trap, and StyleFool targeting C3D on UCF-101 (Table I).

attack types (Table I), with improvements ranging from $1.5\times$ to $7.8\times$ across various attack-classifier-dataset combinations. This underscores VIDTOKEN's ability to leverage latent frame tokens to identify distinct adversarial characteristics. Figure 3 further illustrates these distinctions with significant differences in means, variances, and PCA between clean and adversarial videos, with each represented by a single RTS token, focusing on C3D using UCF-101.

Additionally, Table II indicates AdvIT's low DSRs against U3D, Geo-Trap, and StyleFool, which preserve temporal consistency, highlighting the limitations of relying solely on temporal consistency for anomaly detection. OUDefend also underperforms due to its inability to handle subtle perturbations with its over/undercomplete features in a restoration network. Among image-centric defenses, AT and ComDefend perform better, yet AT relies on known adversarial examples, and ComDefend, which focuses on image reconstruction, diminishes recognition accuracy on clean videos.

While VIDTOKEN effectively defends against adversarial attacks, it also maintains high recognition accuracy for clean videos, as shown in Table III. Compared to AdvIT, which attains FPRs of 3.2% for UCF-101 and 8.2% for HMDB-51, VIDTOKEN reduces FPRs to 0.8% and 1.3%, respectively, though it may occasionally block clean videos. In contrast, defenses such as RS, IT, ComDefend, OUDefend, and AT, although designed to preserve classifier accuracy during attacks, tend to reduce recognition accuracy on clean videos. This superior performance stems from VIDTOKEN's one-class detector $\mathcal{D}$, trained exclusively on clean videos, which effectively discerns the boundaries of clean and adversarial videos. This robustness allows $\mathcal{D}$ to handle variations in

TABLE II: **DSRs (Detection Success Rates)** of VIDTOKEN and six baselines (AdvIT, RS, IT, ComDefend, OUDefend and AT) on adversarial videos from U3D, Geo-Trap, and StyleFool, aimed at **C3D**, **I3D** and **VTN** on UCF-101 and HMDB-51.

Model	Attack Method	AdvIT [25]		RS [38]		IT [37]		ComDefend [21]		OUDefend [27]		AT [23]		VIDTOKEN	
		UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51
C3D	U3D	12.0%	15.9%	16.1%	30.2%	50.0%	12.5%	70.9%	46.0%	57.1%	32.4%	47.3%	38.2%	92.0%	90.5%
	Geo-Trap	19.7%	10.5%	23.2%	32.1%	25.0%	0.0%	65.8%	56.6%	48.7%	22.6%	64.5%	58.5%	98.7%	98.1%
	StyleFool	31.2%	25.0%	13.8%	6.3%	8.3%	25.0%	42.2%	18.8%	37.8%	25.0%	62.4%	53.1%	96.3%	93.8%
I3D	U3D	12.3%	9.6%	25.2%	29.2%	20.2%	26.2%	18.3%	22.6%	34.9%	28.6%	49.8%	41.2%	94.7%	86.4%
	Geo-Trap	19.7%	9.2%	21.6%	27.0%	38.1%	0.0%	82.5%	63.5%	28.9%	23.8%	64.9%	61.9%	95.9%	96.8%
	StyleFool	13.7%	18.6%	9.0%	7.3%	16.7%	11.1%	75.0%	63.1%	43.1%	18.4%	57.0%	59.2%	96.5%	91.7%
VTN	U3D	17.4%	12.0%	5.6%	4.8%	15.5%	17.9%	4.4%	4.0%	34.2%	27.9%	62.1%	45.8%	90.0%	92.4%
	Geo-Trap	21.4%	17.1%	7.1%	5.7%	0.0%	5.7%	7.1%	2.9%	28.6%	20.0%	35.7%	37.1%	85.7%	91.4%
	StyleFool	15.6%	11.1%	6.7%	0.0%	11.1%	14.8%	4.8%	0.0%	24.4%	18.5%	42.2%	40.7%	97.8%	92.6%

TABLE III: Comparing VIDTOKEN with six baselines on **clean videos**, using **false positive rate** for VIDTOKEN and AdvIT (detection-only), and **accuracy** for RS, IT, ComDefend, OUDefend and AT (protection-oriented) across **C3D**, **I3D** and **VTN**.

Model	Recognition Accuracy												False Positive Rate			
	Unprotected		RS [38]		IT [37]		ComDefend [21]		OUDefend [27]		AT [23]		AdvIT [25]		VIDTOKEN	
	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51
C3D	78.3%	60.2%	65.6%	51.2%	49.7%	38.2%	73.1%	56.4%	59.7%	48.4%	76.2%	59.2%				
I3D	87.6%	62.5%	53.8%	54.3%	58.4%	42.3%	82.3%	59.6%	55.3%	42.9%	87.3%	62.5%	3.2%	8.2%	0.8%	1.3%
VTN	91.6%	73.1%	79.3%	65.8%	68.4%	52.5%	72.9%	61.6%	60.3%	45.7%	89.5%	73.1%				

TABLE IV: **DSRs (Detection Success Rates)** of VIDTOKEN and six baselines on **adversarial videos** from Geo-Trap's **targeted** mode (Geo-Trap (T)), and StyleFool's **white-box** mode (StyleFool (W)), aiming at C3D on UCF-101 and HMDB-51.

Model	Attack Method	AdvIT [25]		RS [38]		IT [37]		ComDefend [21]		OUDefend [27]		AT [23]		VIDTOKEN	
		UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51	UCF-101	HMDB-51
C3D	Geo-Trap (T)	16.7%	19.6%	14.3%	25.5%	0.0%	0.0%	63.1%	51.0%	48.8%	19.6%	61.9%	60.8%	97.6%	96.1%
	StyleFool (W)	17.6%	15.2%	11.2%	8.8%	20.0%	11.2%	51.2%	38.4%	43.2%	29.6%	52.8%	46.4%	94.4%	95.2%

clean videos and accurately detect subtle adversarial changes, effectively minimizing the FPR.

VTN. Tables II and III also demonstrate VIDTOKEN's consistent superiority over six baselines in defending the transformer-based video classifier VTN against three attack types (Table I), with DSR improvements ranging from $1.5\times$ to $21.0\times$ across various scenarios. Additionally, VIDTOKEN maintains low FPRs for clean videos (at or below 1.3%), highlighting its ability to use latent frame tokens to detect distinct adversarial characteristics targeting the CLS token in transformer-based attacks. In contrast, ComDefend performs significantly worse on VTN than on C3D and I3D. ComDefend's reconstruction optimization for CNNs focuses on localized feature extraction, effective for clean videos but less so against adversarial attacks exploiting global features, which compromises its efficacy against the sophisticated attack strategies employed on VTN.

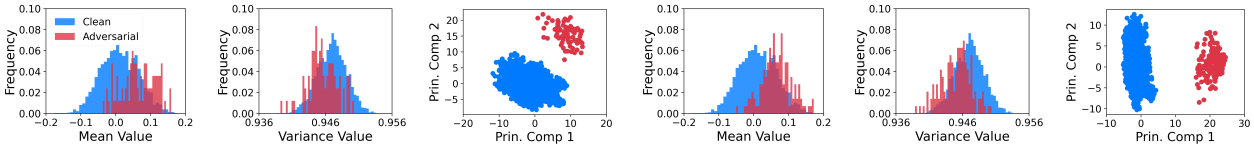
These findings highlight VIDTOKEN's effectiveness against advanced attacks on VRSs (listed in Table I).

Other Attack Settings. We show VIDTOKEN's effectiveness by varying its three parameters in the default setting.

Untargeted vs. Targeted Attacks. In our default attack setting (Table I), all attacks are untargeted (Section VI-A). We further evaluate VIDTOKEN against Geo-Trap's targeted mode (Geo-Trap T), generating 84 adversarial examples on UCF-101 and 51 on HMDB-51 for C3D. As shown in Table IV (in the "Geo-Trap (T)" row), VIDTOKEN achieves DSRs of 97.6% and 96.1% against targeted Geo-Trap on UCF-101 and HMDB-51, respectively—comparable to its performance against untargeted Geo-Trap attacks (Table II) and continues to outperform all six baselines. Figure 4(a), similar to Figure 3(b), shows VIDTOKEN's effective separation of clean videos and Geo-Trap (T) adversarial videos in the latent token space for C3D on UCF-101. This robust defense against both targeted and untargeted attacks stems from VIDTOKEN's ability to detect anomalies in the latent token space, regardless of attack intent.

Black-Box vs. White Box. In adversarial attacks, white-box settings grant attackers direct access to classifier's gradients, whereas black-box scenarios require gradient estimation. With StyleFool, we further explored its white-box scenario, StyleFool (W), by generating 125 adversarial videos for each dataset (UCF-101 and HMDB-51) targeting C3D. Transitioning to StyleFool (W) involves solely modifying how we access the target classifier's gradients. As shown in Table IV (in the "StyleFool (W)" row), VIDTOKEN achieves DSRs of 94.4% on UCF-101 and 95.2% on HMDB-51 against StyleFool (W), surpassing all six baselines. This effectiveness is further illustrated in Figure 4(b) for StyleFool (W), similar to Figure 3(c) for StyleFool. VIDTOKEN's robustness against both attack types stems from its token learning, which detects changes in the token space independently of the classifier's gradients.

Varying Perturbation Intensities. In our default setting (Ta-



(a) Geo-Trap (T) (b) StyleFool (W)

Fig. 4: Comparing mean values, variances, and distributions derived by PCA for clean videos in the UCF-101 dataset and adversarial examples generated using Geo-Trap (T) and StyleFool (W) with $\epsilon = 8/255$ for the C3D video classifier [8].

TABLE V: **DSRs** of VIDTOKEN and six baselines on adversarial examples generated for U3D and StyleFool attacks targeting the C3D classifier on UCF-101 dataset at four **perturbation intensities** $\epsilon \in \{8/255, 16/255, 32/255, 64/255\}$.

Model	Attack	Intensity (ϵ)	AdvIT	RS	IT	ComDefend	OUDefend	AT	VIDTOKEN
C3D	U3D	8/255	12.0%	16.1%	50.0%	70.9%	57.1%	47.3%	92.0%
		16/255	14.7%	13.3%	22.7%	80.0%	46.7%	56.0%	93.3%
		32/255	14.7%	19.3%	12.7%	76.0%	52.7%	45.3%	94.0%
		64/255	14.7%	10.7%	30.7%	72.7%	39.3%	46.7%	91.3%
	StyleFool	8/255	31.2%	13.8%	8.3%	42.2%	37.8%	62.4%	96.3%
		16/255	18.7%	16.0%	18.0%	40.0%	32.7%	59.3%	96.7%
		32/255	20.0%	9.3%	13.3%	38.0%	28.0%	56.7%	96.7%
		64/255	18.0%	12.0%	6.0%	37.3%	34.7%	58.0%	97.3%

ble I), we set $\epsilon = 8/255$ per prior studies [13], [14], as larger perturbations become noticeable. However, Table V shows VIDTOKEN's effectiveness across increased intensities for U3D and StyleFool. We generated 150 adversarial videos per attack type for C3D on UCF-101 at intensities $\epsilon \in \{16/255, 32/255, 64/255\}$, exceeding the roughly 50 examples used [12], [13]. VIDTOKEN maintains its effectiveness across all perturbation levels for both attacks, as detailed in Table V and illustrated in Figure 5 for StyleFool (following page) and Figure 8 for U3D (appendix). In contrast, the six baselines often show lower DSRs at higher intensities. VIDTOKEN 's robustness across varying perturbation intensities highlights its sophisticated adversarial modification detection capabilities, maintaining effectiveness even as ϵ increases, potentially making adversarial modifications more noticeable.

C. RQ2: ViDToken's Efficiency

We assess VIDTOKEN's computational efficiency, revealing that it outperforms the baselines in both training (pre-integration) and inference (post-integration) stages.

Training (Pre-Integration). The training of VIDTOKEN focuses on its video token learning module (Algorithm 1). AT's training involves enriching a classifier with additional data, ComDefend trains its CNN for reconstruction, and OUDefend trains its feature restoration CNN. Unlike AT and OUDefend, but like ComDefend, VIDTOKEN's training does not depend on the target classifier. AdvIT, IT, and RS require no training. For a fair comparison among VIDTOKEN, ComDefend, OUDefend, and AT, we trained each from scratch to convergence with a batch size of eight on the same device (Section VI-A). Trained on UCF-101 and HMDB-51, VIDTOKEN converges swiftly within five epochs, significantly faster than AT's 50 epochs, ComDefend's 10, and OUDefend's 50 epochs. Although VIDTOKEN introduces some training costs, unlike AdvIT, IT, and RS, it is more time-efficient than ComDefend, OUDefend, and AT, as shown in Table XIII (appendix).

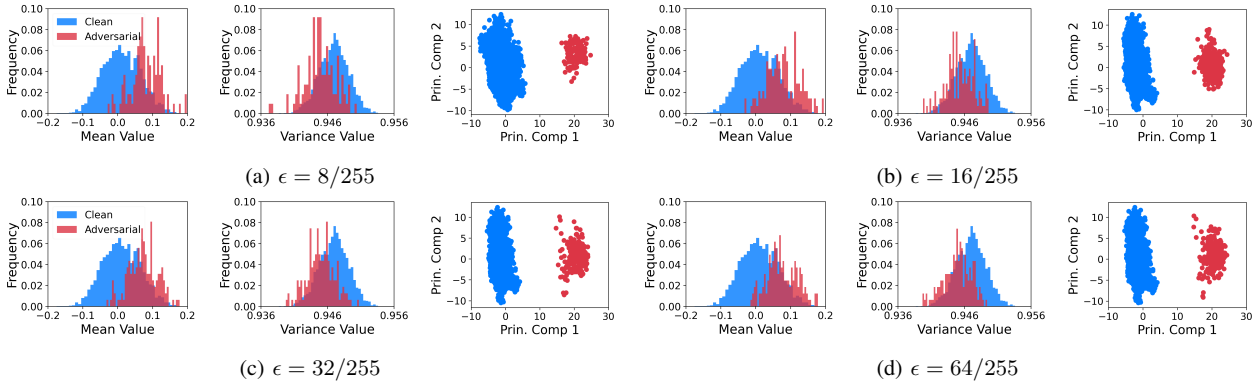
TABLE VI: Average PIIT (**Post-Integration Inference Times** in milliseconds) of VIDTOKEN, AdvIT, ComDefend, OUDefend, and "Unprotected" (representing RS, IT and AT) for 300 videos randomly selected from UCF-101 and HMDB-51.

Dataset	Model	Unprotected	AdvIT	OUDefend	ComDefend	VIDTOKEN
UCF-101	C3D	62.0	1987.8	121.5	131.8	135.0
	I3D	23.5	1949.3	51.7	93.3	96.4
HMDB-51	C3D	46.7	1974.5	118.6	128.3	122.7
	I3D	24.7	1950.5	76.6	106.3	98.9

Inference (Post-Integration). We assess the *Post-Integration Inference Time (PIIT)* of defense-enhanced video classifiers using 300 randomly selected clean videos from each dataset (UCF-101 and HDMB-51), measuring the average inference time per video. PIITs for RS, IT, and AT closely match those of an unprotected classifier; hence, we use the unprotected classifier's times, denoted "Unprotected", for these methods. According to Table VI, VIDTOKEN's inference times exceed those of RS, IT, AT (categorized under "Unprotected") and OUDefend, are comparable to ComDefend, and significantly shorter than AdvIT. VIDTOKEN increases the average PIIT for C3D on UCF-101 from 62.0 ms (unprotected) to 135.0 ms (protected), which is modest compared to AdvIT's increase to 1987.8 ms due to its high computational load from generating pseudo frames via optical flow [25]. Similar trends are observed across other model-dataset combinations. VIDTOKEN extends VRS inference times by $2\times$ to $4\times$, peaking at 135 milliseconds—still below the authentication times of modern facial recognition systems [43]–[45]. This highlights VIDTOKEN's effective management of computational overhead.

D. RQ3: Impact of VIDTOKEN's Settings

We investigate the impact of VIDTOKEN's configuration settings, specifically, one-class detector $\mathcal{D}$, spatial encoder E_{spa}, temporal-awareness attention encoder E_{temp}, and RTS token selection—on its detection capabilities and computational efficiency. This analysis is essential for optimizing VIDTOKEN for practical applications, aimed at enhancing adversarial detection accuracy, reducing FPRs on clean videos, and maintaining high computational efficiency. Additionally,



(a) $\epsilon = 8/255$ (b) $\epsilon = 16/255$

(c) $\epsilon = 32/255$ (d) $\epsilon = 64/255$

Fig. 5: Comparing mean values, variances, and distributions derived by PCA for clean videos in the UCF-101 dataset and adversarial examples generated for StyleFool at four different perturbation intensities for the C3D video classifier [8].

TABLE VII: Impact of different **one-class detectors** $\mathcal{D}$.

Model	Dataset	Attack Method	OC-SVM	LOF	OC-DNN
C3D	UCF-101	U3D	100.0%	99.0%	92.0%
		Geo-Trap	100.0%	98.7%	98.7%
		StyleFool	100.0%	98.2%	96.3%

(a) DSRs for Adversarial Videos

Model	Dataset	OC-SVM		LOF		OC-DNN	
		FPR	PIIT (ms)	FPR	PIIT (ms)	FPR	PIIT (ms)
C3D	UCF-101	3.0%	135.2	6.1%	135.2	0.8%	135.0

(b) FPRs and PIITs for Clean Videos

TABLE VIII: Impact of different **spatial encoders** E_{spa}.

Model	Dataset	Attack Method	ViT	ResNet50	ResNet101
C3D	UCF-101	U3D	92.0%	86.7%	76.2%
		Geo-Trap	98.7%	89.5%	76.3%
		StyleFool	96.3%	93.8%	77.1%

(a) DSRs for Adversarial Videos

Model	Dataset	ViT		ResNet50		ResNet101	
		FPR	PIIT (ms)	FPR	PIIT (ms)	FPR	PIIT (ms)
C3D	UCF-101	0.8%	135.0	6.3%	135.4	9.6%	153.2

(b) FPRs and PIITs for Clean Videos

the appendix delves into the effects of sliding window size w and attention head count H in E_{temp} (Tables XIV and XV). Our evaluation covers DSR, FPR, and PIIT (Post-Integration Inference Time, introduced in Section VI-C) across three advanced types of attacks: U3D, Geo-Trap, and StyleFool, targeting C3D on UCF-101. We assess DSR using all tailored adversarial examples for C3D and UCF-101, as detailed in Table I. FPR and PIIT measurements, presented in Table VII – Table X, and Figure 6, utilize a set of 300 clean videos randomly selected from the UCF-101 dataset.

One-Class Detector $\mathcal{D}$. We evaluated alternatives to our default OC-DNN [56] detector, including LOF (Local Outlier Factor) [58] and OC-SVM [57]. OC-DNN leverages DNNs to learn normal data patterns, LOF identifies anomalies by comparing data points to their neighbors based on deviation, and OC-SVM distinguishes normal from abnormal data using a hyperplane. We implemented LOF and OC-SVM using scikit-learn [73] and sourced OC-DNN from its open-source implementation [72] (Section VI-A).

Table VII presents our results. All three detectors exhibit similar PIITs. Across the three attack types, OC-DNN achieves DSRs ranging from 92.0% to 98.7% and attains the lowest FPR at 0.8%. LOF yields slightly higher DSRs (98.2% to 99.0%) but also the highest FPR at 6.1%. OC-SVM achieves a 100% DSR across all attack types with an intermediate FPR of 3.0%. These results highlight VIDTOKEN's robustness with different one-class detectors and the strong discriminative power of its RTS token representations, validating OC-DNN as the optimal default choice for balancing DSR and FPR.

Spatial Encoder E_{spa}. Beyond ViT [65], we evaluated two CNN-based encoders, ResNet50 and ResNet101, with different depths. As shown in Table VIII, VIDTOKEN achieves the best performance with ViT as E_{spa}, excelling in DSR, FPR, and PIIT. This outcome highlights the transferability of adversarial attacks among CNNs [13], [32], but their limited transferability to transformers [32], [33] (Section IV-B). Additionally, moving from ResNet50 to ResNet101 results in worsened DSR, FPR, and PIIT, suggesting that deeper networks may hinder the detection of adversarial perturbations.

Temporal-Awareness Attention Encoder E_{temp}. The dual-encoder setup of VIDTOKEN enables the extraction of latent tokens from video frames, covering both spatial and temporal dimensions. Adapting image-focused defenses to videos without considering temporal dimension is inadequate for effective adversarial video detection. We assessed the impact of removing the temporal-awareness encoder E_{temp} from VIDTOKEN. Without E_{temp}, VIDTOKEN relies solely on spatial tokens from E_{spa} (Section V-A), bypassing RTS token selection and using a randomly chosen spatial token (SPA-RTS) instead. Table IX shows the effect of this SPA-RTS variant on VIDTOKEN's DSR against U3D, Geo-Trap, and StyleFool targeting C3D on UCF-101, as well as its FPR on clean videos. Using SPA-RTS results in a significant

TABLE IX: Impact of modifying VIDTOKEN's architecture by bypassing the **temporal-awareness attention encoder** E_{temp} and substituting its RTS token with a SPA-RTS, a randomly selected spatial token from VIDTOKEN's spatial encoder E_{spa}, on video representation: This table presents VIDTOKEN's DSRs against adversarial videos from U3D, Geo-Trap, and StyleFool, targeting C3D on UCF-101, and its FPRs on clean videos from UCF-101, as detailed in Section VI-D.

Model	Dataset	Video Representation	FPR	DSRs		
				U3D	Geo-Trap	StyleFool
C3D	UCF-101	RTS	0.8%	92.0%	98.7%	96.3%
		SPA-RTS	13.3%	43.4%	63.2%	37.6%

TABLE X: Impact of selecting **a least representative token** (LST-RTS) for video representation on VIDTOKEN's DSRs, using adversarial videos from U3D, Geo-Trap, and StyleFool with C3D on UCF-101 (Table I), alongside VIDTOKEN's FPRs on UCF-101 clean videos, as detailed in Section VI-D.

Model	Dataset	Video Representation	FPR	DSRs		
				U3D	Geo-Trap	StyleFool
C3D	UCF-101	RTS	0.8%	92.0%	98.7%	96.3%
		LST-RTS	12.9%	9.7%	31.6%	30.3%

47.6% drop in average DSR across the attacks due to the loss of temporal information, while FPR rises from 0.8% to 13.3% on clean videos. Despite these changes, both setups maintain identical PIITs (Table VI). These results highlight the crucial role of E_{temp} in enhancing detection capabilities by incorporating temporal information.

RTS Token Selection. We evaluated the effectiveness of VIDTOKEN's RTS strategy, which identifies the most representative frame token for a video, against two alternatives: (1) selecting the least representative frame token (LST-RTS) and (2) using the CLS token.

RTS vs. LST-RTS. VIDTOKEN identifies a video using its most representative frame token $T_{RTS} = T_i$, where $i = \arg\max_{j \in [1..k]} A_j$ (Equation (3)). Alternatively, a worst-case scenario involves using the least representative token (LST-RTS) $T_{LST\text{-}RTS} = T_i$, where $i = \arg\min_{j \in [1..k]} A_j$. Table X contrasts these configurations by assessing DSRs with adversarial examples for C3D-UCF-101 (Table I) and FPRs with clean UCF-101 videos. The LST-RTS variant shows a drastic 71.8% drop in average DSR across all three attack types due to ineffective capture of spatio-temporal attributes [64]. Both configurations have the same PIIT (Table VI). These results emphasize the strategic value of the RTS mechanism in boosting detection capabilities while maintaining efficiency.

Furthermore, we analyzed the impact of using multiple frame tokens, ranked by attention scores (Equation (2)), on 300 clean UCF-101 C3D videos. As shown in Figure 6, C3D's PIITs increase with the number of tokens, resulting in multiplicative growth of VIDTOKEN's PIITs compared to the single-token baseline.

RTS vs. CLS. Table XI shows that the CLS token is ineffective, as it is the primary target for misleading transformer-based VRSs like VTN, which VIDTOKEN protects. With the RTS token, VIDTOKEN achieves $90.0\%-97.8\%$ DSRs against adversarial videos targeting VTN on UCF-101 and HMDB-51, with FPRs of 0.8% and 1.3%, respectively (with these results taken from Tables II and III). This significantly outperforms the CLS-token variant, which shows DSRs ranging from 31.3% to 76.1% across all attack-dataset scenarios—an average drop of 41.9%—and FPRs of 5.3% on UCF-101 and 46.4% on HMDB-51. Despite these lower DSRs, our one-class detector using the CLS token remains partially effective.

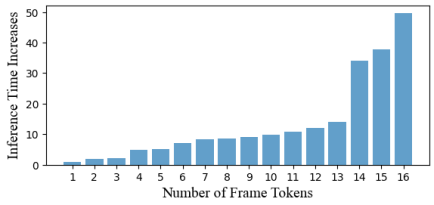


Fig. 6: C3D's **PIITs** on 300 clean videos selected randomly from the UCF-101 dataset with VIDTOKEN using multiple tokens (normalized to its single-token RTS baseline).

TABLE XI: VIDTOKEN's effectiveness in identifying adversarial videos from U3D, Geo-Trap, and StyleFool targeting **VTN**, and its FPRs for clean videos on UCF-101 and HMDB-51. Performance comparisons are based on VIDTOKEN's use of RTS versus CLS tokens for representing videos.

Model	Dataset	Video Representation	FPR	DSR		
				U3D	Geo-Trap	StyleFool
VTN	UCF-101	RTS	0.8%	90.0%	85.7%	97.8%
		CLS	5.3%	42.9%	68.4%	76.1%
	HMDB-51	RTS	1.3%	92.4%	91.4%	92.6%
		CLS	46.4%	38.1%	41.5%	31.3%

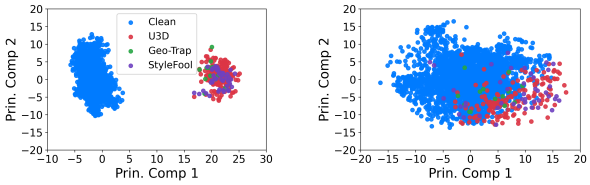
PCA on video representations in Figure 7 and Figure 9 (appendix) shows that the CLS-token variant struggles to distinguish clean from adversarial videos, leading to low DSRs and high FPRs, especially for HMDB-51. Despite this, the CLS token still offers some protection, as it can carry adversarial perturbations detected by VIDTOKEN's one-class classifier. (It is important to point out that PCA is well-known for potentially obscuring separable data points in the original, higher-dimensional spaces.) In contrast, the RTS-token approach effectively separates clean videos from adversarial ones, achieving significantly higher DSRs with lower FPRs.

This evaluation confirms VIDTOKEN's effectiveness across various configuration settings, highlighting it as a robust framework for adversarial video detection.

E. RQ4: Mitigating Adaptive Attacks

We assess VIDTOKEN's robustness against adaptive attacks, including sparse attacks [42] that challenge its temporal-awareness attention encoder, and GAN-style attacks [60]—applicable to all defenses—that directly infer the RTS token.

Adaptive Sparse Attacks. We evaluate VIDTOKEN's resilience against adaptive sparse attacks [42] that perturb a



(a) RTS Token (b) CLS Token

Fig. 7: PCA on RTS and CLS token representations for UCF-101 clean videos and adversarial videos from U3D, Geo-Trap and StyleFool, targeting VTN on UCF-101 (Table I).

TABLE XII: VIDTOKEN's performance with frame replication against adaptive sparse attacks on the UCF-101 and HMDB-51 datasets using C3D, I3D, and VTN.

Model	Attack Method	UCF-101		HMDB-51	
		DSR	FPR	DSR	FPR
C3D	Sparse Attack	78.5%	1.8%	81.3%	2.7%
I3D		76.9%		78.7%	
VTN		66.3%		65.5%	

single frame, challenging its temporal-awareness attention encoder. To maintain effectiveness, we introduce a novel frame-replication countermeasure. Although this increases VRS inference times, it ensures VIDTOKEN remains highly effective against both the sparse attacks assessed and the advanced attacks assessed earlier in RQ1–RQ3.

By selecting and perturbing exactly one frame at random within each 16-frame clip, attackers undermine VIDTOKEN 's temporal-awareness encoder E_{temp}, which amplifies adversarial features via neighboring tokens. With only a single altered frame, E_{temp} cannot effectively highlight sparse perturbations, causing the one-class detector to miss these adversarial videos—detection rates fall below 30% while inference times remain 2–4×.

To counter sparse attacks, we propose a frame-replication countermeasure, inspired by VIDTOKEN's success against U3D [13], which perturbs four frames uniformly. We divide each 16-frame video into four sub-videos, each containing four frames. Each frame in a sub-video is replicated three times to recreate a full 16-frame video. VIDTOKEN processes each sub-video separately, declaring success if any is detected as adversarial. This approach balances analysis time and DSR differently compared to other tested replication strategies, providing nearly the best empirical trade-off.

We tested our frame-replication approach against adaptive sparse attacks using an open-source implementation [74]. For C3D, I3D, and VTN on UCF-101, we generated 130, 130, and 80 adversarial examples, respectively, and for HMDB-51, we generated 150, 150, and 110, respectively, by perturbing one frame per video within $\epsilon = 8/255$ under the ℓ_2-norm [42]. Table XII shows that VIDTOKEN achieves an average DSR of 74.6% across all classifier-dataset combinations, outperforming all six baselines under adaptive sparse attacks—even when baselines are not subjected to their own adaptive attacks—except for I3D-Geo-Trap-UCF-101, where ComDefend achieves a DSR of 82.5%. Notably, AT struggles against unseen adversarial videos [13], and StyleFool [12] reports that ComDefend achieves DSRs of 36% (UCF-101) and 34% (HMDB-51) for C3D, and 53% (UCF-101) and 48% (HMDB-51) for I3D when operating under its adaptive attacks. Frame replication increases VRS inference times from $2-4\times$ to $2-16\times$ in the worst case, yet stays under 1 second, similar to modern facial recognition systems [43]–[45].

When deployed against sparse attacks, VIDTOKEN was reassessed alongside six leading defenses for frame-replication effectiveness against U3D [13], Geo-Trap [14], and StyleFool [12]. This evaluation extends prior results in Tables II and III without frame replication. As described in Section VI-A, Geo-Trap and StyleFool perturb all frames, while U3D perturbs only 25%. Despite increasing VRS inference times from $2-4\times$ to $2-16\times$ in the worst case, VIDTOKEN achieves an average DSR of 90.3% across all attack-classifier-dataset combinations. Against dense attacks like Geo-Trap and StyleFool, VIDTOKEN achieves DSRs of $92.0\% - 98.7\%$ across all scenarios. For U3D, which perturbs only four frames (a sparse attack under frame replication), DSRs drop slightly to 88.0%, 85.7%, and 86.3% for C3D, I3D, and VTN on UCF-101, and 85.5%, 85.0%, and 83.3% on HMDB-51, yet still surpass all six baselines. Frame replication slightly increases VIDTOKEN's FPRs to 1.8% on UCF-101 and 2.7% on HMDB-51, compared to 0.8% and 1.3% without replication (Table III).

Adaptive GAN-style Attacks. Such adaptive attacks, which are applicable to all adversarial defenses [16], [25], [59], attempt to infer VIDTOKEN's RTS tokens by using its one-class detector as a discriminator, mirroring GAN's design [60]. These attacks aim to deceive both the detector and the video classifier simultaneously, a requirement that significantly contributes to their low success rate ($< 0.08\%$) [59]. Attackers struggle to craft perturbations that flip the classifier's decision yet avoid triggering anomalous token patterns—misclassifying the video without leaving detectable traces in the latent token space is inherently conflicting [75]. Enhancing VIDTOKEN's robustness with a defense ensemble [76] can further complicate the attack, improving detection rates [59].

VII. DISCUSSION

Defending machine learning models is inherently more challenging than attacking them. Adversarial examples stem from complex non-linear, non-convex optimization problems in DNNs [18], [30], [31], compounded by the lack of robust theoretical models to describe their solutions. This paper explores transformer-based defenses, highlighting their strengths and limitations to inspire further research in enhancing these approaches. Several observations follow.

Superiority of VIDTOKEN over Existing Defenses. We designed VIDTOKEN using video transformers [39], [54], [55] as a pre-inference defense against adversarial video attacks. While a detailed theoretical analysis of VIDTOKEN's detection accuracy is lacking due to the challenges in modeling adver-

sarial videos [18], [30], [31], its effectiveness is demonstrated through established attack insights (Section IV-B), transformers' robust theoretical foundations, and comprehensive evaluation. Image-centric defenses like AT [23], IT [37], ComDefend [21], and RS [38] overlook temporal dynamics, while video-centric methods like AdvIT [25] and OUDefend [27], relying on optical flow, struggle with temporally consistent attacks. By integrating spatial and temporal dimensions, VIDTOKEN effectively detects perturbation-induced deviations.

VIDTOKEN's Computational Overhead. VIDTOKEN uses a frame-replication countermeasure against adaptive sparse attacks, increasing VRS inference times to 2–16× in extreme cases. However, VIDTOKEN's average PIIT remains below 1 second (Table VI), making it viable for high-risk environments such as urban surveillance and video verification for DSC, where the benefits outweigh the computational costs. For comparison, iPhone's Face ID takes approximately 1 second for user authentication [43], [44], and a recent anti-spoofing facial authentication system takes 1.26 seconds [45] on average. Therefore, given its improved defense against adaptive sparse attacks, we recommend enabling frame replication by default—especially in security-critical deployments where robustness outweighs minimal latency. Future work will focus on enhancing efficiency through model tuning, quantization, and hardware acceleration, aiming to advance transformer-based defenses by reducing costs while maintaining high DSRs under adaptive attacks.

VIDTOKEN's Robustness Against Distribution Shift. Like many defenses [21], [23], [25], [27], [37], [38], [48], [59], VIDTOKEN relies on the classifier's training data and may misclassify out-of-distribution or unseen video samples as adversarial—a common limitation [16], [25], [59]. However, advanced methods for handling out-of-distribution data and adapting to dataset shifts [77]–[79] present promising opportunities to enhance VIDTOKEN. Notably, VRSs are often application-specific, making access to complete training datasets feasible for robust defense.

Adversarial Examples as Contrastive Instances. A promising direction for enhancing VIDTOKEN's defense capabilities lies in leveraging adversarial examples as contrastive instances during training. Currently, the one-class detector learns exclusively from benign video representations, identifying the RTS token distribution of clean samples. Incorporating adversarial examples as negative contrastive instances could improve discriminative power by explicitly learning boundaries between benign and adversarial representations. This contrastive learning approach would optimize the one-class detector to maximize the distance between benign and adversarial representations while maintaining tight clustering for benign examples, transforming VIDTOKEN into a more robust defense mechanism against adversarial video examples.

VIII. Related Work

The susceptibility of AI, especially DNNs, to adversarial attacks is well-known in the security and deep learning communities. Early adversarial video attacks on VRSs employed a white-box approach, manipulating video classifiers' gradient details to create adversarial examples [42], [80]–[83]. To overcome the unrealistic assumption of open access to classifier gradients, researchers developed black-box methods that approximate this information [12]–[15], [84], [85]. Current strategies refine perturbations within ℓ_p-norm limits and focus on gradient estimation techniques like FGSM [31] and PGD [30]. Additionally, methods like U3D [13], GeoTrap [14], and StyleFool [12] (Section III-A) explore specific attack vectors. C-DUP [80] creates universal perturbations in a white-box setting. V-BAD [15] targets black-box classifiers with query-limited attacks.

Most existing defenses against adversarial attacks are designed for images, enhancing model robustness through Adversarial Training [22], [23], [29], Input Transformations [37], Random Smoothing [38], ComDefend [21], [46], DiffPure [47]), detecting adversarial examples (e.g., [86]–[88], ADDNP [16]), or preventing the creation of adversarial examples (e.g., Blacklight [51]). For video-specific defenses, AdvIT [25] and [26] use temporal consistency to detect adversarial videos. DP [28] tailors defense patterns based on known adversarial examples, while [52] uses JPEG compression for defenses [52]. OUDefend [27] employs over/undercomplete features in a restoration network tailored for video classifiers. SecVID [48] uses video compressive sensing to defend video classifiers. Adversarial training, adapted for video, uses known adversarial examples to retrain video classifiers [19], [29].

IX. Conclusion

This paper introduces VIDTOKEN, a novel adversarial video detection approach leveraging video transformers to enhance the security of both CNN- and transformer-based VRSs. VIDTOKEN tokenizes video frames to capture spatial and temporal information, serving as a non-intrusive pre-inference layer without requiring classifier access or prior knowledge of attacks. It outperforms existing defenses with higher detection accuracy for adversarial examples while preserving recognition accuracy on clean videos. Although VIDTOKEN increases inference times (under 1 second in the worst case), particularly against adaptive sparse attacks, it is essential for applications like urban surveillance, anomaly detection, and video verification for DSC. Future work aims to optimize computational costs while maintaining detection efficacy and to address challenges posed by LLM-based video understanding models with multi-modal inputs. Leveraging ViT [65] and Longformer [67], VIDTOKEN can scale to large datasets, albeit with longer training times.

X. Acknowledgments

We express our gratitude to the reviewers for their constructive feedback, which has improved both the presentation and quality of our paper. This research is supported by CSIRO's Data61 HSE project.

REFERENCES

[1] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua, "Neural aggregation network for video face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4362–4371.

[2] V. Rausch, A. Hansen, E. Solowjow, C. Liu, E. Kreuzer, and J. K. Hedrick, "Learning a deep neural net policy for end-to-end control of autonomous vehicles," in *American Control Conference*, 2017, pp. 4914–4919.

[3] T. Onishi, T. Motoyoshi, Y. Suga, H. Mori, and T. Ogata, "End-to-end learning method for self-driving cars with trajectory recovery using a path-following function," in *International Joint Conference on Neural Networks*, 2019.

[4] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1933–1941.

[5] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," *arXiv:1901.03407*, 2019.

[6] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6479–6488.

[7] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2625–2634.

[8] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.

[9] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4694–4702.

[10] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in Neural Information Processing Systems*, 2014.

[11] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[12] Y. Cao, X. Xiao, R. Sun, D. Wang, M. Xue, and S. Wen, "Stylefool: Fooling video classification systems via style transfer," in *IEEE Symposium on Security and Privacy*, 2023, pp. 1631–1648.

[13] S. Xie, H. Wang, Y. Kong, and Y. Hong, "Universal 3-dimensional perturbations for black-box attacks on video recognition systems," in *IEEE Symposium on Security and Privacy*, 2022, pp. 1390–1407.

[14] S. Li, A. Aich, S. Zhu, S. Asif, C. Song, A. Roy-Chowdhury, and S. Krishnamurthy, "Adversarial attacks on black box video classifiers: Leveraging the power of geometric transformations," in *Advances in Neural Information Processing Systems*, 2021, pp. 2085–2096.

[15] L. Jiang, X. Ma, S. Chen, J. Bailey, and Y.-G. Jiang, "Black-box adversarial attacks on video recognition models," in *ACM International Conference on Multimedia*, 2019, pp. 864–872.

[16] Y. Qing, T. Bai, Z. Liu, P. Moulin, and B. Wen, "Detection of adversarial attacks via disentangling natural images and perturbations," *IEEE Transactions on Information Forensics and Security*, 2024.

[17] H. Zhu, S. Zhang, and K. Chen, "Ai-guardian: Defeating adversarial attacks using backdoors," in *IEEE Symposium on Security and Privacy*, 2023, pp. 701–718.

[18] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in *ACM workshop on Artificial Intelligence and Security*, 2017, pp. 3–14.

[19] K. A. Kinfu and R. Vidal, "Analysis and extensions of adversarial training for video classification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3416–3425.

[20] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *International Conference on Machine Learning*, 2019, pp. 1310–1320.

[21] X. Jia, X. Wei, X. Cao, and H. Foroosh, "Comdefend: An efficient image compression model to defend adversarial examples," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6084–6092.

[22] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv:1412.6572*, 2014.

[23] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!" in *Advances in Neural Information Processing Systems*, 2019.

[24] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, "Certified robustness to adversarial examples with differential privacy," in *IEEE Symposium on Security and Privacy*, 2019, pp. 656–672.

[25] C. Xiao, R. Deng, B. Li, T. Lee, B. Edwards, J. Yi, D. Song, M. Liu, and I. Molloy, "Advit: Adversarial frames identifier based on temporal consistency in videos," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3968–3977.

[26] X. Jia, X. Wei, and X. Cao, "Identifying and resisting adversarial videos using temporal consistency," *arXiv:1909.04837*, 2019.

[27] S.-Y. Lo, J. M. J. Valanarasu, and V. M. Patel, "Overcomplete representations against adversarial videos," in *IEEE International Conference on Image Processing*, 2021, pp. 1939–1943.

[28] H. J. Lee and Y. M. Ro, "Defending video recognition model against adversarial perturbations via defense patterns," *IEEE Transactions on Dependable and Secure Computing*, 2023.

[29] N. Thakur and B. Li, "Pat: Pseudo-adversarial training for detecting adversarial videos," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 131–138.

[30] G. Sriramanan, S. Addepalli, A. Baburaj *et al.*, "Guided adversarial attack for evaluating and enhancing adversarial defenses," in *Advances in Neural Information Processing Systems*, 2020, pp. 20 297–20 308.

[31] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4312–4321.

[32] Z. Wei, J. Chen, M. Goldblum, Z. Wu, T. Goldstein, and Y.-G. Jiang, "Towards transferable adversarial attacks on vision transformers," in *AAAI Conference on Artificial Intelligence*, 2022, pp. 2668–2676.

[33] Z. Wei, J. Chen, M. Goldblum, Z. Wu, T. Goldstein, Y.-G. Jiang, and L. S. Davis, "Towards transferable adversarial attacks on image and video transformers," *IEEE Transactions on Image Processing*, pp. 6346–6358, 2023.

[34] J. Zhang, Q. Yi, and J. Sang, "Towards adversarial attack on vision-language pre-training models," in *ACM International Conference on Multimedia*, 2022, pp. 5005–5013.

[35] W. Yang, Y. Lin, P. Li, J. Zhou, and X. Sun, "Rethinking stealthiness of backdoor attack against nlp models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021, pp. 5543–5557.

[36] J. Deng, Y. Wang, J. Li, C. Shang, H. Liu, S. Rajasekaran, and C. Ding, "Tag: Gradient attack on transformer-based language models," *arXiv:2103.06819*, 2021.

[37] C. Guo, M. Rana, M. Cisse, and L. Van Der Maaten, "Countering adversarial images using input transformations," *arXiv:1711.00117*, 2017.

[38] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *International Conference on Machine Learning*. PMLR, 2019, pp. 1310–1320.

[39] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video transformer network," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3163–3172.

[40] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, 2012.

[41] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: A large video database for human motion recognition," in *International Conference on Computer Vision*, pp. 2556–2563.

[42] X. Wei, J. Zhu, S. Yuan, and H. Su, "Sparse adversarial perturbations for videos," in *AAAI Conference on Artificial Intelligence*, 2019, pp. 8973–8980.

[43] X. Xu, J. Yu, Y. Chen, Q. Hua, Y. Zhu, Y.-C. Chen, and M. Li, "Touchpass: Towards behavior-irrelevant on-touch user authentication on smartphones leveraging vibrations," in *International Conference on Mobile Computing and Networking*, 2020, pp. 1–13.

[44] Y. Mizuho, Y. Kawasaki, T. Amesaka, and Y. Sugiura, "Earauthcam: Personal identification and authentication method using ear images acquired with a camera-equipped hearable device," in *Augmented Humans International Conference*, 2024, pp. 119–130.

[45] W. Xu, J. Liu, S. Zhang, Y. Zheng, F. Lin, J. Han, F. Xiao, and K. Ren, "Rface: Anti-spoofing facial authentication using cots rfid," in *IEEE INFOCOM Conference on Computer Communications*, 2021, pp. 1–10.

[46] C. Ferrari, F. Becattini, L. Galteri, and A. D. Bimbo, "A robust defense against adversarial attacks on image classification," *ACM Transactions on Multimedia Computing, Communications and Applications*, pp. 1–16, 2023.
[47] W. Nie, B. Guo, Y. Huang, C. Xiao, A. Vahdat, and A. Anandkumar, "Diffusion models for adversarial purification," *arXiv:2205.07460*, 2022.
[48] W. Song, C. Cong, H. Zhong, and J. Xue, "Correction-based defense against adversarial video attacks via {Discretization-Enhanced} video compressive sensing," in *USENIX Security Symposium*, 2024, pp. 3603–3620.
[49] X. Wang and K. He, "Enhancing the transferability of adversarial attacks through variance tuning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1924–1933.
[50] Y. Gao, I. Shumailov, K. Fawaz, and N. Papernot, "On the limitations of stochastic pre-processing defenses," in *Advances in Neural Information Processing Systems*, 2022, pp. 24 280–24 294.
[51] H. Li, S. Shan, E. Wenger, J. Zhang, H. Zheng, and B. Y. Zhao, "Blacklight: Scalable defense for neural networks against {Query-Based}{Black-Box} attacks," in *USENIX Security Symposium*, 2022, pp. 2117–2134.
[52] Y. Cheng, X. Wei, H. Fu, S.-W. Lin, and W. Lin, "Defense for adversarial videos by self-adaptive jpeg compression and optical texture," in *ACM International Conference on Multimedia in Asia*, 2021, pp. 1–7.
[53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.
[54] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 244–253.
[55] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6836–6846.
[56] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *International conference on machine learning*, 2018, pp. 4393–4402.
[57] L. M. Manevitz and M. Yousef, "One-class svms for document classification," *Journal of machine Learning research*, pp. 139–154, 2001.
[58] E. H. Budiarto, A. E. Permanasari, and S. Fauziati, "Unsupervised anomaly detection using k-means, local outlier factor and one class svm," in *IEEE International Conference on Science and Technology*, 2019, pp. 1–5.
[59] S. Wang, H. Hu, J. Chang, B. Z. H. Zhao, Q. A. Chen, and M. Xue, "Dnn-gp: diagnosing and mitigating model's faults using latent concepts," in *USENIX Security Symposium*, 2024, pp. 1297–1314.
[60] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014.
[61] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv:1810.04805*, 2018.
[62] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6202–6211.
[63] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3202–3211.
[64] J. Calic and E. Izuierdo, "Efficient key-frame extraction and video analysis," in *International Conference on Information Technology*, 2002, pp. 28–33.
[65] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv:2010.11929*, 2020.
[66] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[67] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv:2004.05150*, 2020.
[68] M. Weber, X. Xu, B. Karlaš, C. Zhang, and B. Li, "Rab: Provable robustness against backdoor attacks," in *IEEE Symposium on Security and Privacy*, 2023, pp. 1311–1328.
[69] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4161–4170.
[70] Huggingface, "Vit pre-trained model," 2019. [Online]. Available: https://huggingface.co/google/vit-base-patch16-224
[71] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.
[72] Ruff, "Open source implementation of oc-dnn," 2018. [Online]. Available: https://github.com/lukasruff/Deep-SVDD
[73] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, pp. 2825–2830, 2011.
[74] N. Thakur, "Sparse adversarial perturbations," 2020. [Online]. Available: https://github.com/thakurnupur/Sparse-Adversarial-Perturbations-PyTorch
[75] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv:1312.6199*, 2013.
[76] S. Kariyappa and M. K. Qureshi, "Improving adversarial robustness of ensembles with diversity training," *arXiv:1901.09981*, 2019.
[77] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo *et al.*, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8340–8349.
[78] H. Ye, C. Xie, T. Cai, R. Li, Z. Li, and L. Wang, "Towards a theoretical framework of out-of-distribution generalization," *Advances in Neural Information Processing Systems*, pp. 23 519–23 531, 2021.
[79] X. Zhang, P. Cui, R. Xu, L. Zhou, Y. He, and Z. Shen, "Deep stable learning for out-of-distribution generalization," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5372–5382.
[80] S. Li, A. Neupane, S. Paul, C. Song, S. V. Krishnamurthy, A. K. R. Chowdhury, and A. Swami, "Adversarial perturbations against real-time video classification systems," *arXiv:1807.00458*.
[81] J.-W. Chang, M. Javaheripi, S. Hidano, and F. Koushanfar, "Adversarial attacks on deep learning-based video compression and classification systems," *arXiv:2203.10183*, 2022.
[82] N. Inkawhich, M. Inkawhich, Y. Chen, and H. Li, "Adversarial attacks for optical flow-based action recognition classifiers," *arXiv:1811.11875*, 2018.
[83] M. Andriushchenko and N. Flammarion, "Understanding and improving fast adversarial training," in *Advances in Neural Information Processing Systems*, 2020, pp. 16 048–16 059.
[84] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1765–1773.
[85] K. T. Co, L. Muñoz-González, S. de Maupeou, and E. C. Lupu, "Procedural noise adversarial examples for black-box attacks on deep convolutional networks," in *ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 275–289.
[86] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner, "Detecting adversarial samples from artifacts," *arXiv:1703.00410*, 2017.
[87] D. Meng and H. Chen, "Magnet: a two-pronged defense against adversarial examples," in *ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 135–147.
[88] X. Ma, B. Li, Y. Wang, S. M. Erfani, S. Wijewickrema, G. Schoenebeck, D. Song, M. E. Houle, and J. Bailey, "Characterizing adversarial subspaces using local intrinsic dimensionality," *arXiv:1801.02613*, 2018.

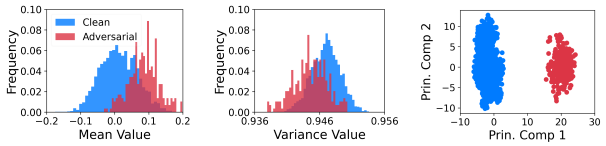
APPENDIX

A1. Adversarial Video Examples

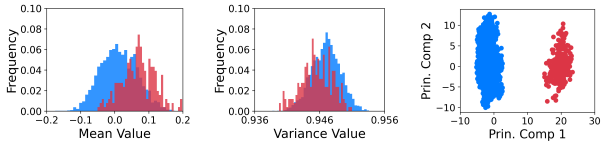
We have made available a set of adversarial video examples via an anonymous GitHub repository: adversarial video examples. These examples cover the three primary adversarial attacks—U3D [13], Geo-Trap [14], and StyleFool [12]—as discussed in Section III-A, as well as the adaptive sparse attack [42] detailed in Section VI-E. The repository provides representative perturbed videos alongside their corresponding clean versions, enabling readers to visually assess the perceptibility of the perturbations and their impact on

classification results. We hope that this resource facilitates further research in understanding the visual characteristics of adversarial perturbations and in developing more effective defense mechanisms.

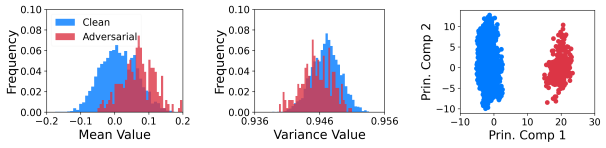
A2. PCA for Analyzing Perturbation Intensities Referenced in Section VI-B (RQ1).



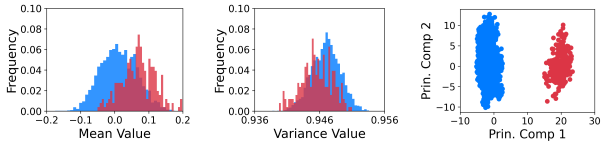
(a) $\epsilon = 8/255$



(b) $\epsilon = 16/255$



(c) $\epsilon = 32/255$



(d) $\epsilon = 64/255$

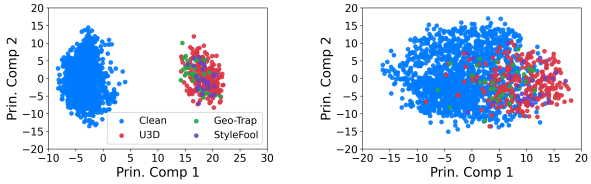
Fig. 8: Comparing mean values, variances, and distributions derived by PCA for clean videos in the UCF-101 dataset and adversarial examples generated for U3D attack at four different perturbation intensities for C3D.

A3. VIDTOKEN's Computational Efficiency in Training times Referenced in Section VI-C (RQ2).

TABLE XIII: Comparing VIDTOKEN with AT, ComDefend, and OUDefend in **training times** (in hours) on UCF-101 and HMDB-51 for C3D and I3D (where VIDTOKEN, and ComDefend, unlike AT and OUDefend, are independent of the target video classifier considered).

Dataset	Model	OUDefend	AT	ComDefend	VIDTOKEN
UCF-101	C3D	3.3	2.6	4.0	0.28
	I3D	4.5	3.8		
HMDB-51	C3D	2.4	2.1	2.3	0.16
	I3D	3.7	2.3		

A4. PCA for Analyzing Adversarial Attacks Targeting VTN Referenced in Section VI-D (RQ3).



(a) RTS Token (b) CLS Token

Fig. 9: PCA on RTS and CLS token representations for UCF-101 clean videos and adversarial videos from U3D, Geo-Trap and StyleFool, targeting VTN on HMDB-51 (Table I).

A5. Further Analysis of E_{temp}'s Sliding Window Size and Attention Head Count on VIDTOKEN's Performance as Referenced in Section VI-D (RQ3).

TABLE XIV: Impact of varying **window sizes** w **in** E_{temp} on VIDTOKEN's performance.

Model	Dataset	Attack Method	$w = 4$	$w = 8$	$w = 16$
C3D	UCF-101	U3D	91.5%	92.0%	93.2%
		Geo-Trap	97.4%	98.7%	96.1%
		StyleFool	97.2%	96.3%	96.3%

(a) DSRs for Adversarial Videos

Model	Dataset	$w = 4$		$w = 8$		$w = 16$	
		FPR	PIIT (ms)	FPR	PIIT (ms)	FPR	PIIT (ms)
C3D	UCF-101	1.1%	134.7	0.8%	135.0	1.3%	136.3

(b) FPRs and PIITs for Clean Videos

Sliding Window Size w **in** E_{temp}**.** The choice of w influences E_{temp}'s ability to recognize temporal patterns, balancing contextual understanding with computational efficiency. We evaluated VIDTOKEN using $w = 4$, 8, and 16 to determine its impact, noting that $w = 4$ is the minimum required by Longformer [67]. Table XIV presents the results, highlighting $w = 8$ as the optimal setting for performance. Increasing the window size from $w = 4$ to $w = 16$ led to a rise in PIIT, suggesting that larger windows capture temporal patterns more effectively but are computationally demanding. However, a window size of $w = 8$, which is half the number of input frames as per Section VI-A, proved best, achieving 92.0%–98.7% DSRs, the lowest FPR at 0.8%, and a PIIT comparable to $w = 4$, thus offering the most favorable balance across all performance metrics.

Attention Head Count H **in** E_{temp}**.** Attention heads in E_{temp} operate simultaneously, enhancing VIDTOKEN's video analysis efficiency by learning across multiple sub-spaces, crucial for detecting adversarial samples with minimal computational load. As shown in Table XV, multiple heads significantly

TABLE XV: Impact of **attention head counts H in E_{temp}** on VIDTOKEN's performance.

Model	Dataset	Attack	$H=1$	$H=2$	$H=4$	$H=8$	$H=16$
C3D	UCF-101	U3D	92.0%	93.2%	91.0%	92.0%	93.0%
		Geo-Trap	96.0%	97.4%	98.7%	98.7%	98.7%
		StyleFool	97.2%	98.2%	94.5%	96.3%	96.3%

(a) DSRs for Adversarial Videos

$H=1$		$H=2$		$H=4$		$H=8$		$H=16$	
FPR	PIIT (ms)	FPR	PIIT (ms)	FPR	PIIT (ms)	FPR	PIIT (ms)	FPR	PIIT (ms)
1.3%	228.3	1.1%	158.6	0.8%	132.5	0.8%	135.0	1.5%	139.0

(b) FPRs and PIITs for Clean Videos

reduce PIIT compared to a single-head setup ($H=1$), aligning with findings from [39], [53] and reducing computational demands. VIDTOKEN demonstrates consistently high DSRs across all configurations against adversarial examples. For clean videos, minor FPR increases are observed at both ends of the H range. The optimal setting, $H = 8$, recommended by [39], proves effective, demonstrating VIDTOKEN's efficient handling of multiple analytical sub-spaces.