StegaVAR: Privacy-Preserving Video Action Recognition via Steganographic Domain Analysis

Lixin Chen[1*], Chaomeng Chen[1,2,6*], Jiale Zhou[4,5], Zhijian Wu[5], Xun Lin[3†]

[1]School of Computing and Information Technology, Great Bay University
[2]Tsinghua Shenzhen International Graduate School, Tsinghua University
[3]School of Computer Science and Engineering, Beihang University
[4]College of Science and Technology, Zhejiang University
[5]School of Engineering, Westlake University
[6]Dongguan Key Laboratory for Intelligence and Information Technology
clxhsa@outlook.com, linxun@buaa.edu.cn

Abstract

Despite the rapid progress of deep learning in video action recognition (VAR) in recent years, privacy leakage in videos remains a critical concern. Current state-of-the-art privacy-preserving methods often rely on anonymization. These methods suffer from (1) low concealment, where producing visually distorted videos that attract attackers' attention during transmission, and (2) spatiotemporal disruption, where degrading essential spatiotemporal features for accurate VAR. To address these issues, we propose StegaVAR, a novel framework that embeds action videos into ordinary cover videos and directly performs VAR in the steganographic domain for the first time. Throughout both data transmission and action analysis, the spatiotemporal information of hidden secret video remains complete, while the natural appearance of cover videos ensures the concealment of transmission. Considering the difficulty of steganographic domain analysis, we propose *S*ecret Spatio-*Te*mporal *P*romotion (*STeP*) and *Cro*ss-Band *D*ifference *A*ttention (*CroDA*) for analysis within the steganographic domain. STeP uses the secret video to guide spatiotemporal feature extraction in the steganographic domain during training. CroDA suppresses cover interference by capturing cross-band semantic differences. Experiments demonstrate that StegaVAR achieves superior VAR and privacy-preserving performance on widely used datasets. Moreover, our framework is effective for multiple steganographic models.

Introduction

Video action recognition (VAR) aims to automatically identify the body's movement patterns and behaviors from videos (Sun et al. 2023). This technology is increasingly applied in real-world scenarios like video surveillance, which rely on extensive data transmission and cloud-based analysis (Pittaluga and Koppal 2017; Fitwi and Chen 2020; Deng, Gao, and Xu 2023; Xie et al. 2024). However, this process introduces a significant privacy concern (Dave, Chen, and Shah 2022; Chen et al. 2024; Lin et al. 2024a; Chen and Su 2025), as sensitive attributes such as gender, race, appearance, and the actions themselves can be exposed when videos are uploaded to remote servers. Consequently, privacy-preserving VAR solutions are urgently needed.

Initial attempts at privacy preservation, such as extreme downsampling (Dai et al. 2015; Liu and Zhang 2020; Ryoo et al. 2017; Srivastav, Gangi, and Padoy 2019; Chou et al. 2018) or naive region obfuscation (Ren, Lee, and Ryoo 2018; Zhang et al. 2021), proved insufficient, as they severely degrade VAR accuracy. Obfuscation-based techniques also use pretrained detectors to find and modify private regions, for instance by synthesizing fake faces (Ren, Lee, and Ryoo 2018) or applying segmentation-guided blurring (Zhang et al. 2021). However, these methods are fundamentally limited by poor generalization to unseen privacy attributes, the impracticality of frame-level annotation, and the negative impact of modifications on downstream task performance. To better address the privacy-utility trade-off, research shifted towards end-to-end, learning-based methods. Initial works leveraged supervised adversarial learning to obfuscate features (Wu et al. 2020, 2022), which was later improved by the MaSS framework for selective attribute preservation (Chen et al. 2022). STPrivacy designed a transformer-based model to remove action-irrelevant information at the video level (Li et al. 2023). More recently, self-supervised learning has eliminated the need for privacy labels, as demonstrated by methods that optimize privacy without annotations (Dave, Chen, and Shah 2022), use contrastive learning to mitigate spatial privacy leakage (Fioresi, Dave, and Shah 2023), or introduce penalty-based optimization algorithms to further balance privacy and task performance (Aslam and Nasrollahi 2025). Despite these advancements, existing anonymization techniques are designed to protect privacy, inadvertently introduce new risks, and performance ceilings. *How precisely does this dilemma arise?*

We argue this failure stems from two critical, unaddressed problems inherent to the anonymization process: **(1) Low Concealment.** Anonymization techniques can easily produce visually distorted videos. These alterations create visual artifacts that distinguish them from the torrent of natural video data. This conspicuousness is a security flaw rather

[*]These authors contributed equally.
[†]Corresponding author.

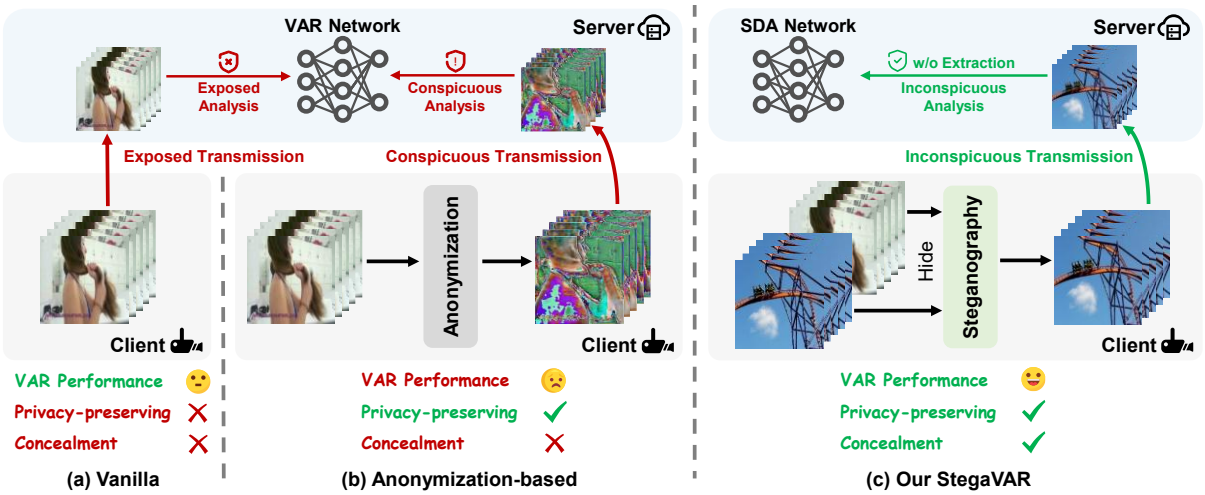

Figure 1: Different client-server VAR framework: (a) vanilla, (b) anonymization-based, (c) our proposed StegaVAR framework. StegaVAR can achieve accurate VAR without attracting the attention of attackers.

than a benign side effect. It acts as a red flag to network adversaries, signaling that the content is sensitive enough to warrant protection. This signal leads to a *cat-and-mouse game* of anonymization and de-anonymization, which may cause more aggressive server/client attack (Rosberg et al. 2023; Gadotti et al. 2024). **(2) Spatiotemporal Disruption.** The process of anonymization is inherently destructive, which irrevocably corrupts the video's pixel data to obscure private information. This corruption often disrupts the fine-grained spatiotemporal relationships and high-frequency details, which are necessary for accurate VAR (Wang et al. 2019; Feichtenhofer et al. 2019; Ye et al. 2025).

The field of steganography, which provides this inspiration, has evolved significantly. Traditional steganographic methods offered limited payload capacity by modifying spatial or transform domains (Wu et al. 2005; Pan, Li, and Yang 2011; Imaizumi and Ozawa 2014). The advent of deep learning revolutionized this field: HiDDeN (Zhu et al. 2018) and SteganoGAN (Zhang et al. 2019) pioneered encoder-decoder structures for image hiding, while Deep Steganography (Baluja 2020) enabled full-size image embedding via convolutional networks. Discrete wavelet transform (DWT) (He et al. 2012) was subsequently introduced to further enhance the reversibility (Liu et al. 2021). A significant progression came with Invertible Neural Networks (INNs) (Dinh, Krueger, and Bengio 2015; Dinh, Sohl-Dickstein, and Bengio 2017), where HiNet (Jing et al. 2021) established parameter-shared bidirectional mapping for reversible transformations. Building upon image hiding, research has increasingly focused on video hiding, which requires greater embedding capacity (Weng et al. 2019). While separate encoder-decoder mechanisms (Islam et al. 2019) remain effective, they result in high model complexity. LF-VSN (Mou et al. 2023) achieved large-capacity multi-video hiding via a unified INN while reducing model complexity.

Drawing inspiration from these advancements in video steganography (Guan et al. 2022; Lu et al. 2021; Jing et al. 2021; Li et al. 2024; Mou et al. 2023; Deng, Gao, and Xu 2023), we propose a novel framework for privacy-preserving video action recognition via steganographic domain analysis (StegaVAR), reframing the problem from one of "editing" a video to one of "hiding" it. StegaVAR protects privacy information in videos without attracting attackers' attention. As shown in Fig. 1, our approach conceals the secret video within a natural cover video to generate the stego video for server upload, then directly performs VAR in the steganographic domain without video extraction or decryption. Since secret information is primarily embedded in high-frequency components of the stego video (Jing et al. 2021), these features remain subtle, while cover information introduces substantial interference that impedes discriminative feature extraction for action recognition. The subtlety of these features and interference from the cover video make it challenging to apply existing action models to the steganographic domain directly. We propose a Steganographic Domain Analysis network (SDANet) to solve the problem.

Within SDANet, we design two modules: **S**ecret Spatio-**Te**mporal **P**romotion (**STeP**) and **Cro**ss-Band **D**ifference **A**ttention (**CroDA**). Considering video steganography embeds secret videos into high-frequency components of cover videos, we decompose videos using Discrete Wavelet Transform (He et al. 2012) and analyze different frequency bands separately. During training, STeP utilizes the secret video's high-frequency components to guide feature extraction in the stego video's spatial (Zhang et al. 2023) and temporal dimensions. CroDA computes differences between frequency bands to suppress interference from the cover video (Cai et al. 2025; Lin et al. 2025), further enhancing performance with a shared position embedding to maintain temporal consistency across sub-bands.

Our StegaVAR framework circumvents the issues of anonymization entirely by losslessly embedding a secret video into a natural-looking cover video and performing analysis directly in the steganographic domain. Through-

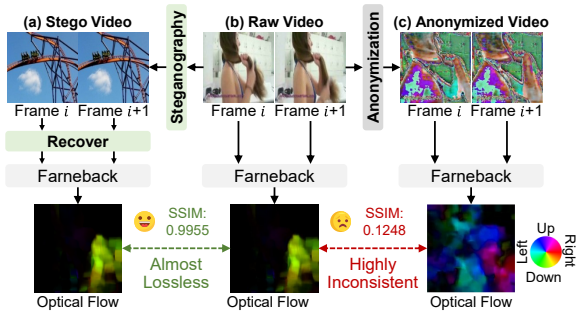

Figure 2: Comparative optical flow visualization: (a) stego video, (b) raw video, and (c) anonymized video. Raw video is embedded losslessly, maintaining coherent motion patterns, while anonymization disrupts temporal coherence.

out both data transmission and action analysis, the hidden secret video remains unexposed. As shown in Fig. 2, anonymization clearly disrupts the temporal information of video, while the process of steganography is lossless. StegaVAR ensures the data remains inconspicuous while preserving original spatiotemporal features, thereby maintaining high performance on the downstream task. Overall, our proposed StegaVAR conceals privacy attributes in action videos without raising suspicion and minimally disrupts spatiotemporal features in secret videos, thereby achieving accurate privacy-preserving VAR. Our contributions are summarized as follows:

- We propose the StegaVAR framework, a new paradigm for privacy-preserving VAR, which integrates video steganography with VAR for the first time, realizing accurate VAR without attracting attackers' attention.
- We design the STeP and CroDA modules to perform VAR directly in the steganographic domain, which fully leverages secret features in high-frequency components, protecting privacy while maintaining VAR performance.
- Our method achieves strong generalizability across multiple steganographic models in terms of VAR acc and concealment capability on six publicly available datasets.

Methodology

Our StegaVAR framework uses a client-side steganography network $\mathcal{S}(\cdot)$ and a server-side steganographic domain analysis network $\mathcal{A}(\cdot)$ for end-to-end privacy-preserving action recognition. As shown in Fig. 3, the entire pipeline can be conceptualized as a covert channel communication system. For an action video $\boldsymbol{x}_{secret}$ containing private information, the client embeds it into a visually natural cover video $\boldsymbol{x}_{cover}$ through the steganography network $\mathcal{S}(\cdot)$, generating a stego video containing secret information: $\boldsymbol{x}_{stego} = \mathcal{S}(\boldsymbol{x}_{cover}, \boldsymbol{x}_{secret})$. The stego video $\boldsymbol{x}_{stego}$ is visually indistinguishable from the original cover video $\boldsymbol{x}_{cover}$, avoiding attackers' attention. The server's analysis network, $\mathcal{A}(\cdot)$, then performs steganographic domain analysis on the stego video to yield the prediction $\hat{y} = \mathcal{A}(\boldsymbol{x}_{stego})$. Throughout this process, the private video $\boldsymbol{x}_{secret}$ always exists in concealed frequency-domain component form and is never exposed in transmission links or server memory. Compared to anonymization-dependent methods, StegaVAR attracts no adversarial attention, eliminating security risks induced by anonymization. Moreover, the server never reconstructs the original video, fundamentally preventing privacy leakage while avoiding disruption of original temporal information.

Steganographic Domain Analysis Network

Since the secret information is primarily embedded in high-frequency components, existing vanilla VAR networks struggle to analyze in the steganographic domain effectively (as demonstrated by experiments in Table. 2). This primarily occurs because the visual representation of $\boldsymbol{x}_{stego}$ mainly originates from $\boldsymbol{x}_{cover}$, while VAR-related features are concealed within the steganographic domain and difficult to extract. Therefore, we decompose $\boldsymbol{x}_{stego}$ using Discrete Wavelet Transform (DWT) to obtain four sub-bands: $\boldsymbol{x}^{LL}_{stego}, \boldsymbol{x}^{LH}_{stego}, \boldsymbol{x}^{HL}_{stego}, \boldsymbol{x}^{HH}_{stego} \in \mathbb{R}^{T\times\frac{H}{2}\times\frac{W}{2}\times 3}$, representing low-frequency and high-frequency features. Subsequently, we employ ResNet3D-18 (Tran et al. 2018) as the backbone to separately analyze these four sub-bands.

To resolve the misalignment between steganographic features and original action semantics, we design the Secret Spatio-Temporal Promotion (STeP) module, which explicitly guides feature learning of each sub-band in $\boldsymbol{x}_{stego}$ during training using spatiotemporal features from $\boldsymbol{x}_{secret}$, enforcing spatial layout and temporal evolution alignment between steganographic representations and secret actions (note that $\boldsymbol{x}_{secret}$ is excluded during inference). Since the cover video's semantics reside mainly in the low-frequency sub-band $\boldsymbol{x}^{LL}_{stego}$ (Jing et al. 2021; Lin et al. 2024b), the Cross-Band Difference Attention (CroDA) module suppresses interference by computing differences between it and the high-frequency sub-bands (LH/HL/HH). CroDA also uses a shared learnable position embedding to ensure temporal consistency across all frequency components. Finally, features from each sub-band are processed by Adaptive Average Pooling, and a two-layer MLP with a sigmoid activation generates weights for each sub-band. Weighted feature aggregation is then performed via a two-layer convolutional network.

Secret Spatio-Temporal Promotion. Steganographic models often embed different parts of the secret video into distinct frequency bands of the cover video using varied methods (Lin et al. 2024b), making directly learning secret features within these bands challenging. Beyond steganography, the Discrete Wavelet Transform (DWT) is also employed to extract high-frequency detail information from images, frequently used in tasks requiring attention to fine-grained texture information such as industrial defect detection (Zhang et al. 2023). Inspired by such methods, we introduce an additional STeP branch during training. Unlike traditional applications of DWT in vision tasks, we leverage it not only to decompose spatial features in horizontal and vertical directions but also to further decompose features along the temporal dimension. This temporal decomposition

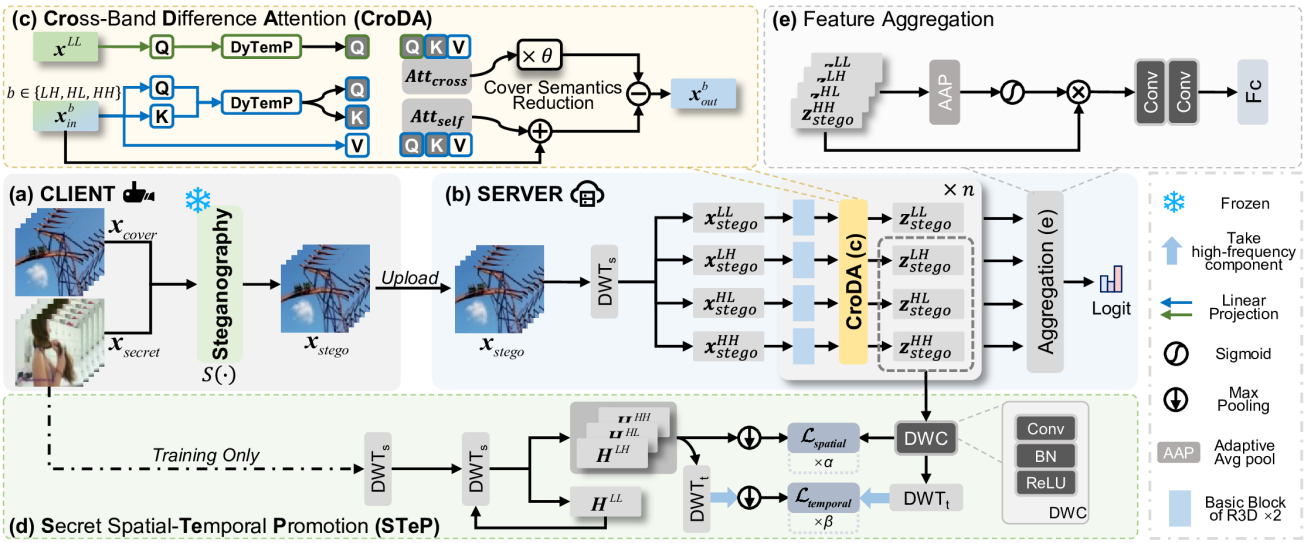

Figure 3: Illustration of StegaVAR's full pipeline. (a) At the client side, the steganography network $\mathcal{S}$ embeds $\boldsymbol{x}_{secret}$ into $\boldsymbol{x}_{cover}$ to generate $\boldsymbol{x}_{stego}$. (b) On the server side, SDANet decomposes $\boldsymbol{x}_{stego}$ into four sub-bands via DWT and extracts features from each sub-band separately. (c) The CroDA module computes cross-band feature differences to suppress interference from the cover video. (d) During training, the STeP module leverages the original secret video to guide feature learning separately along spatial and temporal dimensions. (e) Final classification is produced through feature aggregation and fully-connected layers.

allows the module to guide feature learning across temporal scales, which is critical for capturing fine-grained actions defined by subtle motion dynamics. With this spatiotemporal supervision, the network can better capture information in the steganographic domain (as shown in Fig. 3).

The STeP branch uses high-frequency spatiotemporal features from $\boldsymbol{x}_{secret}$ to supervise SDANet's learning of high-frequency features in $\boldsymbol{x}_{stego}$. First, we apply a transform to $\boldsymbol{x}_{secret}$ and obtain its LL band to ensure dimensional alignment with subsequent feature maps:

$$H_0^{LL}, H_0^{LH}, H_0^{HL}, H_0^{HH} = \mathrm{DWT_s}(\boldsymbol{x}_{secret}), \tag{1}$$

where $\mathrm{DWT_s}(\cdot)$ denotes the discrete wavelet transform in spatial dimension. Then we perform four levels of DWT on H_0^{LL} (each subsequent transform is applied to the low-frequency component from the previous level). For $n \in \{1,2,3,4\}$:

$$H_n^{LL}, H_n^{LH}, H_n^{HL}, H_n^{HH} = \mathrm{DWT_s}(H_{n-1}^{LL}), \tag{2}$$

where $H_n^{LH}, H_n^{HL}, H_n^{HH} \in \mathbb{R}^{T\times\frac{H}{2^{n+1}}\times\frac{W}{2^{n+1}}\times 3}$ denote the high-frequency components at the n-th level. Subsequently, apply DWT along the temporal dimension to these components. Since the temporal dimension T is compressed during feature extraction, we use max pooling on the high-frequency components to create the spatial (G^s) and temporal (G^t) ground truth signals for promotion. This process can be represented as:

$$\begin{aligned} G_n^{s,b} &= \mathcal{P}_n(H_n^b), \quad b \in \mathcal{B}, \\ G_n^{t,b} &= \mathcal{P}_n(\mathrm{DWT_t}(H_n^b)_{high}), \end{aligned} \tag{3}$$

where $\mathrm{DWT_t}(\cdot)_{high}$ denotes the wavelet decomposition along the temporal dimension retaining only high-frequency components, and $\mathcal{B}$ denotes the set of three high-frequency sub-bands $\{LH, HL, HH\}$. $\mathcal{P}_n$ represents the level-dependent max pooling operation. It only selects significant features along the time dimension without changing height and width. Next, the sub-band features of the stego video are processed by a simple Deep Wavelet Compression (DWC) module (comprising $1\times1\times1$ convolution, batch normalization, and ReLU) to align channel dimensions with the high-frequency components of the secret video:

$$\begin{aligned} M_n^{s,b} &= \mathrm{DWC}_n\left(z_n^b\right), \quad b \in \mathcal{B}, \\ M_n^{t,b} &= \mathrm{DWT_t}(M_n^{s,b})_{high}, \end{aligned} \tag{4}$$

where z_n represents the the sub-band feature at the n-th layer. Then, calculate the spatial guidance loss $\mathcal{L}_{spatial}$ and the time guidance loss $\mathcal{L}_{temporal}$ respectively:

$$\begin{aligned} \mathcal{L}_{spatial}^b &= \sum_{n=1}^{N} \mathcal{L}_{mse}(G_n^{s,b}, M_n^{s,b}), \\ \mathcal{L}_{temporal}^b &= \sum_{n=1}^{N} \mathcal{L}_{mse}(G_n^{t,b}, M_n^{t,b}), \end{aligned} \tag{5}$$

where N represents the four DWT performed and $\mathcal{L}_{mse}$ denotes mean squared error loss. Finally, calculate $\mathcal{L}_{STeP}$:

$$\mathcal{L}_{STeP} = \alpha \cdot \sum_{b\in B} \mathcal{L}_{spatial}^b + \beta \cdot \sum_{b\in B} \mathcal{L}_{temporal}^b, \tag{6}$$

where α and β respectively represent the strength of spatial and temporal promotion.

Cross-Band Difference Attention. We frame the task of VAR in the steganographic domain as a signal denoising problem. The high-frequency sub-band of stego video contains a superposition of the desired action signal from $\boldsymbol{x}_{secret}$ and a small amount of unwanted noise from $\boldsymbol{x}_{cover}$.

Considering that the low-frequency sub-band $\boldsymbol{x}_{stego}^{LL}$ contains the majority of the cover video's semantic information (Jing et al. 2021; Guan et al. 2022), while the secret information is primarily embedded within the three high-frequency sub-bands ($\boldsymbol{x}_{stego}^{LH}, \boldsymbol{x}_{stego}^{HL}, \boldsymbol{x}_{stego}^{HH}$), we compute element-wise difference attention between each high-frequency sub-band and the low-frequency sub-band $\boldsymbol{x}_{stego}^{LL}$. This process aims to remove residual cover information lingering within the high-frequency sub-bands, effectively performing a content-adaptive filtering operation that can suppress the cover-related interference and enhance the underlying action signal.

Simultaneously, although the high-frequency sub-bands contain distinct features spatially, their subtle temporal variations are similar. Therefore, we employ two sets of position embeddings to separately encode the low-frequency sub-band and the high-frequency sub-bands.To enhance perception of subtle motion variations, we propose **Dy**namic **Tem**poral **P**erception (**DyTemP**). In order to adaptively perceive and unify temporal information in different sub-bands, we introduce a learnable position-specific offset based on RoPE (Su et al. 2024). This hybrid approach preserves relative position awareness while adaptively adjusting the absolute temporal landmarks essential for VAR. Specifically, for an input vector $x \in \mathbb{R}^{T\times d}$ partitioned into $x = [u, v]$ where $u, v \in \mathbb{R}^{T\times \frac{d}{2}}$, the DyTemP $E(x)$ is defined as below:

$$E(x) = \begin{bmatrix} \mathbf{u} \odot (\cos \boldsymbol{\Theta} + \varepsilon_{cos}) - \mathbf{v} \odot (\sin \boldsymbol{\Theta} + \varepsilon_{sin}) \\ \mathbf{u} \odot (\sin \boldsymbol{\Theta} + \varepsilon_{sin}) + \mathbf{v} \odot (\cos \boldsymbol{\Theta} + \varepsilon_{cos}) \end{bmatrix}, \quad (7)$$

where Θ denotes the matrix of rotation angles pre-computed based on token position and feature dimension. $\varepsilon_{\cos}$ and $\varepsilon_{\sin}$ are learnable biases. The implementation of CroDA can be formulated as follows:

$$\mathrm{CA}(x^{LL}, x^{b}) = \sigma(\frac{E_{LL}(Q(x^{LL}))E_b(K(x^b))^\top}{\sqrt{d}})V(x^b), \quad (8)$$

$$\mathrm{SA}(x^{b}) = \sigma(\frac{E_b(Q(x^b))E_b(K(x^b))^\top}{\sqrt{d}})V(x^b), \quad (9)$$

$$x_{out}^b = x_{in}^b + \mathrm{SA}(x_{in}^b) - \theta \cdot \mathrm{CA}(x^{LL}, x_{in}^b), \quad b \in \mathcal{B}, \quad (10)$$

where $\mathrm{CA}(\cdot)$ and $\mathrm{SA}(\cdot)$ denote the standard implementation of cross-attention and self-attention. σ represents the Softmax function and θ is the subtraction strength for the difference component. E_{LL} and E_b correspond to position embedding operations applied to the low-frequency and high-frequency sub-bands, respectively, with parameters in E_b shared across LH, HL and HH. Q, K, V represent linearly projected values for query, key, and value, respectively.

Experiments

Datasets & Metrics

Datasets. For VAR evaluation, we select the two most widely adopted datasets: UCF101 (Soomro, Zamir, and Shah 2012) and HMDB51 (Kuehne et al. 2011). To quantitatively compare privacy preservation performance with other methods, we utilize two subsets of VISPR (Orekondy, Schiele, and Fritz 2017) and the privacy-annotated versions of UCF101 and HMDB51, namely VPUCF101 and VPHMDB51 (Li et al. 2023) (Further details are provided in Supp.). All datasets employ their official split protocols. We randomly sample 1000 video clips from YouTube-VIS (Yang, Fan, and Xu 2019) as cover videos, ensuring no overlap between the training and testing cover videos. For cover samples with fewer than 16 frames, we employ reverse-order padding to extend the sequence length.

Metrics. Following previous works (Dave, Chen, and Shah 2022; Aslam and Nasrollahi 2025), we employ Top-1 accuracy to evaluate VAR performance. For privacy preservation performance, we use classwise-mAP (mean Average Precision) and classwise-F1 score.

Implementation Details

Input Details. In all experiments, we first cropped each frame to 0.8× the original dimensions before resizing to an input resolution of 224×224. Each video clip consisted of 16 frames, sampled from random starting points at a frame skip rate of 4. During training, we applied standard augmentations to $\boldsymbol{x}_{secret}$, including random erasing, random cropping, horizontal flipping, and random color jittering.

Initialization and Training Details. The steganography model $\mathcal{S}$ used frozen DIV2K-pretrained weights (Agustsson and Timofte 2017); all VAR models trained from scratch. In StegaVAR, θ in Eq. 10 was set to 0.2. For raw data or other privacy methods, $\theta = 0$ retained only self-attention. This hyperparameter setting is based on CroDA's function of removing low-frequency information from high-frequency components, a difference computation strategy that is uniquely effective in the steganographic domain and would otherwise degrade performance. VAR models trained 150 epochs with Adam (lr=1e-4, batch=32), loss coefficients α=0.2 and β=0.3 in Eq. 6. Privacy evaluation used ResNet-50 (He et al. 2016) with ImageNet weights (Deng et al. 2009) trained 100 epochs under identical optimization. All experiments were implemented in PyTorch on four NVIDIA RTX 4090 GPUs.

Main Results

VAR Performance. As shown in Table. 1, our StegaVAR framework significantly outperforms existing privacy-preserving methods, achieving accuracy comparable to non-private baselines. Specifically, StegaVAR with LF-VSN (Mou et al. 2023) achieves 71.66% Top-1 accuracy on UCF101 and 43.66% on HMDB51, surpassing the state-of-the-art BPAP (Aslam and Nasrollahi 2025) by over 9% on both datasets while incurring only a minimal 0.32%/0.59% drop compared to the raw data baseline. The superiority of LF-VSN over other steganographic models like Weng (Weng et al. 2019) and HiNet (Jing et al. 2021) is attributed to its use of inter-frame information, which better preserves temporal features. Furthermore, the poor performance of vanilla VAR networks on stego videos (shown in Table. 2) confirms the necessity of a specialized network for effective analysis within the steganographic domain.

Notably, in Table. 2, SDANet substantially outperforms the vanilla ResNet3D on raw data (71.98% v.s. 62.33% on UCF101). This advantage stems from the wavelet decomposition in STeP, where high-frequency components provide

Method	VAR Performance		Privacy-Preserving Performance			
	UCF101	UCF101→HMDB51	VISPR1		VPHMDB	
	Top-1 ↑	Top-1 ↑	cMAP ↓	F1 ↓	cMAP ↓	F1 ↓
Raw data	71.98	44.25	64.41	0.555	76.62	0.684
Downsample-2×	54.11	24.10	57.23	0.483	71.35	0.601
Downsample-4×	39.65	16.80	50.07	**0.379**	69.79	0.594
Obf-Blackening	53.13	26.20	56.39	0.457	74.06	0.649
Obf-StrongBlur	55.59	26.40	55.94	0.456	74.33	0.655
Obf-WeakBlur	61.52	33.70	63.52	0.523	75.11	0.663
Noise-Features	61.90	31.20	62.40	0.531	-	-
VITA	62.10	33.20	55.32	0.461	73.89	0.638
SPAct	62.03	34.10	57.43	0.473	-	-
BPAP	62.11	34.52	57.10	0.450	69.95	**0.519**
StegaVAR (Weng)	70.32	42.88	51.66	0.459	59.78	0.549
StegaVAR (HiNet)	70.08	42.75	**47.10**	0.399	58.93	0.530
StegaVAR (LF-VSN)	**71.66**	**43.66**	47.87	0.507	**56.65**	0.531

Table 1: Comparison with existing privacy-preserving VAR frameworks across datasets, while VAR reports Top-1 (%) and Privacy reports cMAP (%) and F1. ↑ denotes higher is better, ↓ denotes lower is better.

Method		UCF101	UCF101→ HMDB51
Privacy-preserving	VAR Network	Top-1 ↑	Top-1 ↑
Raw data	ResNet3D	62.33	35.60
	SDANet	71.98	44.25
BPAP	ResNet3D	62.11	34.52
	SDANet	61.22	37.81
Weng	ResNet3D	59.08	33.13
HiNet	ResNet3D	58.69	33.28
LF-VSN	ResNet3D	58.88	32.67
StegaVAR (Weng)	SDANet	70.32	42.88
StegaVAR (HiNet)	SDANet	70.08	42.75
StegaVAR (LF-VSN)	SDANet	**71.66**	**43.66**

Table 2: Comparison of different privacy preserving methods and VAR network (our SDANet/ResNet3D) combinations, evaluated using Top-1 (%).

fine-grained supervisory signals that enhance focus on subtle temporal variations. Its effectiveness outside the steganographic domain confirms the cross-domain universality of this guidance mechanism. Conversely, SDANet's performance drops to 61.22% on UCF101 when applied to the anonymizing BPAP framework, falling below ResNet3D. This failure demonstrates the inability of DWT to extract meaningful signals, which validates that anonymization techniques inherently cause irreversible spatiotemporal disruption, whereas our steganography-based approach preserves video integrity through embedding.

Privacy-preserving. Although our primary objective is covert transmission and steganographic domain analysis, we quantitatively evaluate StegaVAR's privacy protection by extracting features from $\boldsymbol{x}_{stego}$ using ResNet-50 (He et al. 2016). As shown in Table. 1, ResNet-50 achieves significantly lower privacy recognition on StegaVAR-processed videos, attaining 47.10% cMAP and 0.399 F1 on VISPR1. This outperforms the strongest competitor (55.32% cMAP / 0.461 F1) by 8.22% (cMAP↓) and 0.062 (F1↓). StegaVAR also leads in transfer tasks: StegaVAR (Weng) achieves 45.93% cMAP/0.347 F1 on VISPR1→VISPR2, while StegaVAR (HiNet) attains 61.74% cMAP on VPHMDB→VPUCF (Table. 3). This demonstrates effective resistance against automated privacy inference attacks beyond human perception deception.

Method	VISPR1→2		VPHMDB→VPUCF	
	cMAP ↓	F1 ↓	cMAP ↓	F1 ↓
Raw data	57.63	0.434	76.62	0.699
VITA	49.60	0.399	76.02	0.669
SPAct	47.10	0.386	75.98	0.661
BPAP	49.50	0.352	70.91	0.612
StegaVAR (Weng)	**45.93**	**0.347**	62.54	0.532
StegaVAR (HiNet)	46.84	0.391	**61.74**	0.531
StegaVAR (LF-VSN)	47.04	0.389	61.83	**0.526**

Table 3: Comparison of privacy-preserving performance across datasets, evaluated using cMAP (%) and F1 score.

Ablation Studies

Effectiveness of STeP. Starting from a baseline accuracy of 63.15% on UCF101 without STeP or CroDA (Table. 4), the spatial promotion component alone improves performance to 66.29%, while the temporal promotion component achieves 66.16%. Integrating both spatial and temporal promotion yields 68.54% accuracy. When these components are individually combined with CroDA, performance is further improved to 68.86% (spatial promotion + CroDA) and 68.31% (temporal promotion + CroDA), respectively. These results confirm that explicit guidance from the secret video's high-frequency components is crucial. Furthermore, attention map visualizations in Fig. 4 reveal STeP's critical role in action region localization, which is absent without it.

Effectiveness of CroDA. CroDA substantially suppresses cover-induced interference. As shown in Table. 4, standalone CroDA improves the baseline by 2.66% (65.81% vs. 63.15%). Further analysis reveals that position embedding (PE) is essential for modeling temporal relationships. Models without PE achieve only 70.39% accuracy (Table. 5). Absolute PE (Devlin et al. 2019) marginally improves performance to 70.64% (+0.25%), while RoPE (Su et al. 2024) significantly enhances accuracy to 71.03% by capturing relative temporal dependencies. DyTemP further increases accuracy to 71.66% via the learnable offset ε. The visual evidence in Fig. 4 underscores CroDA's essential function in suppressing cover interference. Ultimately, integrating all components achieves the highest accuracy, confirming that optimal performance relies on the synergy between STeP's spatio-temporal guidance and CroDA's cross-band attention.

Spatial Promotion	Temporal Promotion	CroDA	Top-1 (%) ↑
×	×	×	63.15
✓	×	×	66.29
×	✓	×	66.16
×	×	✓	65.81
✓	✓	×	68.54
✓	×	✓	68.86
×	✓	✓	68.31
✓	✓	✓	**71.66**

Table 4: Ablation results of STeP and CroDA on UCF101, fixing α=0.2, β=0.3, θ=0.2.

Position Embedding	Top-1 (%) ↑
w/o PE	70.39
Absolute PE	70.64
RoPE	71.03
DyTemP	**71.66**

Table 5: Ablation results of position embedding strategies on UCF101.

Group of sub-bands. To validate the necessity of analysis for the four distinct frequency sub-bands using four separate ResNet3D, we conducted experiments with different sub-band grouping strategies (Table. 6). Concatenating all sub-bands for single-branch processing (retaining only CroDA's self-attention) yields the lowest VAR performance (58.03%). Significant improvement occurs when isolating the LL sub-band from others (62.73% v.s. 58.03%). While three-group configurations achieve comparable results across approaches, independent processing of all four sub-bands peaks at 71.66% accuracy. This demonstrates substantial spectral divergence of secret information across frequency bands and confirms that separate analysis optimally captures band-specific discriminative features.

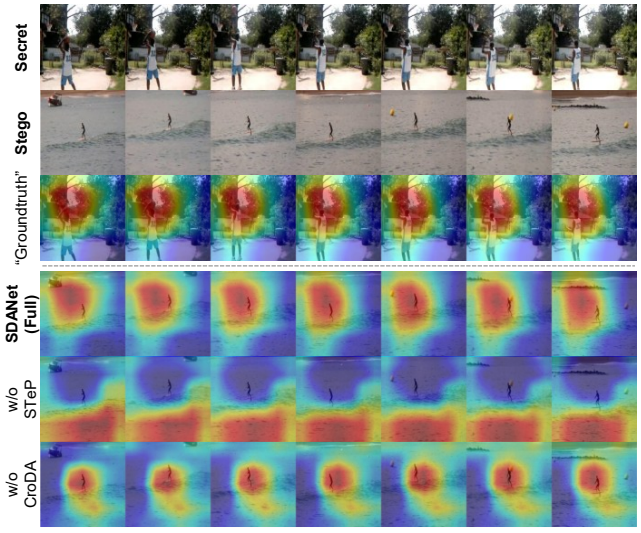

Figure 4: Visual attention of module ablation.

Group of sub-bands	Top-1 (%) ↑
{*LL, LH, HL, HH*}	58.03
{*LL*}, {*LH, HL, HH*}	62.73
{*LL*}, {*LH*}, {*HL, HH*}	66.68
{*LL*}, {*LH, HL*}, {*HH*}	66.74
{*LL*}, {*HL*}, {*LH, HH*}	66.27
{*LL*}, {*LH*}, {*HL*}, {*HH*}	**71.66**

Table 6: Ablation results of sub-band grouping strategies on UCF101.

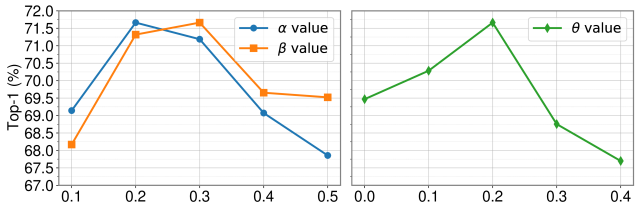

Figure 5: Effect of the three hyperparameters on UCF101. α and β respectively represent the strength of spatial and temporal guidance. θ is the subtraction strength for the difference component of CroDA.

Hyperparameter Analysis

Model sensitivity to α, β (in Eq. 6), and θ (in Eq. 10) necessitates strict calibration at optimal values (0.2, 0.3, 0.2). Fixing β=0.3 and θ=0.2, α=0.2 yields peak accuracy (71.66%), while α=0.1 (69.14%) or α=0.5 (67.86%) cause significant degradation. Similarly, with α=0.2 and θ=0.2, β=0.3 achieves maximum performance (71.66%), though β=0.2 remains competitive (71.32%); values beyond β=0.3 or below 0.2 reduce accuracy by 2-3.5%. Notably, θ exhibits an extremely narrow optimum: θ=0 degrades performance to 69.46% versus θ=0.2 (71.66%), proving cross-band difference computation's essential role. Minor deviations to θ=0.1 (70.28%) or 0.3 (68.76%) substantially harm accuracy, indicating precise calibration is required for limited low-frequency interference in high-frequency components.

Conclusion

We propose StegaVAR, a pioneering framework that rethinks privacy-preserving VAR by integrating steganographic principles. By embedding secret videos into cover videos and performing analysis directly in the steganographic domain, our strategy ensures content concealment while avoiding the spatiotemporal disruption of other methods. Experimental results testified that our STeP and CroDA modules effectively guide feature learning and suppress interference, enabling StegaVAR to achieve recognition performance rivaling non-private baselines with robust privacy protection against both human and automated attacks.

While StegaVAR demonstrates significant advantages over anonymization methods, it incurs minor information loss compared to raw video analysis. Future work should explore advanced reversible transformations or adaptive fusion mechanisms to bridge this fidelity gap.

Acknowledgements

This work was supported by Guangdong Research Team for Communication and Sensing Integrated with Intelligent Computing (Project No. 2024KCXTD047). The computational resources are supported by SongShan Lake HPC Center (SSL-HPC) in Great Bay University.

References

Agustsson, E.; and Timofte, R. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*.

Aslam, N.; and Nasrollahi, K. 2025. Balancing Privacy and Action Performance: A Penalty-Driven Approach to Image Anonymization. In *Proceedings of the Computer Vision and Pattern Recognition Conference Workshops*.

Baluja, S. 2020. Hiding Images within Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cai, R.; Cui, Y.; Yu, Z.; Lin, X.; Chen, C.; and Kot, A. 2025. Rehearsal-Free and Efficient Continual Learning for Cross-Domain Face Anti-Spoofing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chen, C.; and Su, S. 2025. PFDP: privacy-preserving federated distillation method for pretrained language models. *International Journal of Machine Learning and Cybernetics*.

Chen, C.; Zhou, Z.; Tang, P.; He, L.; and Su, S. 2024. Enforcing group fairness in privacy-preserving Federated Learning. *Future Generation Computer Systems*.

Chen, C.-F.; Hu, S.; Shi, Z.; Gulati, P.; Moriarty, B.; Pistoia, M.; Piuri, V.; and Samarati, P. 2022. MaSS: Multi-attribute Selective Suppression. arXiv:2210.09904.

Chou, E.; Tan, M.; Zou, C.; Guo, M.; Haque, A.; Milstein, A.; and Fei-Fei, L. 2018. Privacy-Preserving Action Recognition for Smart Hospitals using Low-Resolution Depth Images. arXiv:1811.09950.

Dai, J.; Saghafi, B.; Wu, J.; Konrad, J.; and Ishwar, P. 2015. Towards privacy-preserving recognition of human activities. In *2015 IEEE International Conference on Image Processing*.

Dave, I. R.; Chen, C.; and Shah, M. 2022. SPAct: Self-Supervised Privacy Preservation for Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*.

Deng, X.; Gao, C.; and Xu, M. 2023. PIRNet: Privacy-Preserving Image Restoration Network via Wavelet Lifting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Dinh, L.; Krueger, D.; and Bengio, Y. 2015. NICE: Non-linear Independent Components Estimation. arXiv:1410.8516.

Dinh, L.; Sohl-Dickstein, J.; and Bengio, S. 2017. Density estimation using Real NVP. arXiv:1605.08803.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. SlowFast Networks for Video Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Fioresi, J.; Dave, I. R.; and Shah, M. 2023. TeD-SPAD: Temporal Distinctiveness for Self-supervised Privacy-preservation for video Anomaly Detection. arXiv:2308.11072.

Fitwi, A.; and Chen, Y. 2020. Privacy-Preserving Selective Video Surveillance. In *2020 29th International Conference on Computer Communications and Networks*.

Gadotti, A.; Rocher, L.; Houssiau, F.; Creţu, A.-M.; and de Montjoye, Y.-A. 2024. Anonymization: The imperfect science of using data while preserving privacy. *Science Advances*.

Guan, Z.; Jing, J.; Deng, X.; Xu, M.; Jiang, L.; Zhang, Z.; and Li, Y. 2022. DeepMIH: Deep Invertible Network for Multiple Image Hiding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

He, Z.; Lu, W.; Sun, W.; and Huang, J. 2012. Digital image splicing detection based on Markov features in DCT and DWT domain. *Pattern Recognition*.

Imaizumi, S.; and Ozawa, K. 2014. Multibit Embedding Algorithm for Steganography of Palette-Based Images. In *Image and Video Technology*.

Islam, S.; Nigam, A.; Mishra, A.; and Kumar, S. 2019. VStegNET: Video Steganography Network using Spatio-Temporal features and Micro-Bottleneck. In *BMVC*.

Jing, J.; Deng, X.; Xu, M.; Wang, J.; and Guan, Z. 2021. HiNet: Deep Image Hiding by Invertible Network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; and Serre, T. 2011. HMDB: A large video database for human motion recognition. In *2011 International Conference on Computer Vision*.

Li, G.; Li, S.; Luo, Z.; Qian, Z.; and Zhang, X. 2024. Purified and Unified Steganographic Network. arXiv:2402.17210.

Li, M.; Xu, X.; Fan, H.; Zhou, P.; Liu, J.; Liu, J.-W.; Li, J.; Keppo, J.; Shou, M. Z.; and Yan, S. 2023. STPrivacy: Spatio-Temporal Privacy-Preserving Action Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Lin, X.; Liu, A.; Yu, Z.; Cai, R.; Wang, S.; Yu, Y.; Wan, J.; Lei, Z.; Cao, X.; and Kot, A. 2025. Reliable and Balanced Transfer Learning for Generalized Multimodal Face

Anti-Spoofing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lin, X.; Yu, Y.; Xia, S.; Jiang, J.; Wang, H.; Yu, Z.; Liu, Y.; Fu, Y.; Wang, S.; Tang, W.; et al. 2024a. Safeguarding Medical Image Segmentation Datasets against Unauthorized Training via Contour-and Texture-Aware Perturbations. *arXiv preprint arXiv:2403.14250*.

Lin, X.; Yu, Y.; Yu, Z.; Meng, R.; Zhou, J.; Liu, A.; Liu, Y.; Wang, S.; Tang, W.; Lei, Z.; and Kot, A. 2024b. HideMIA: Hidden Wavelet Mining for Privacy-Enhancing Medical Image Analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*.

Liu, J.; and Zhang, L. 2020. Indoor Privacy-preserving Action Recognition via Partially Coupled Convolutional Neural Network. In *2020 International Conference on Artificial Intelligence and Computer Engineering*.

Liu, L.; Meng, L.; Peng, Y.; and Wang, X. 2021. A data hiding scheme based on U-Net and wavelet transform. *Knowledge-Based Systems*.

Lu, S.-P.; Wang, R.; Zhong, T.; and Rosin, P. L. 2021. Large-Capacity Image Steganography Based on Invertible Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Mou, C.; Xu, Y.; Song, J.; Zhao, C.; Ghanem, B.; and Zhang, J. 2023. Large-capacity and Flexible Video Steganography via Invertible Neural Network. arXiv:2304.12300.

Orekondy, T.; Schiele, B.; and Fritz, M. 2017. Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images. In *Proceedings of the IEEE International Conference on Computer Vision*.

Pan, F.; Li, J.; and Yang, X. 2011. Image steganography method based on PVD and modulus function. In *2011 International Conference on Electronics, Communications and Control*.

Pittaluga, F.; and Koppal, S. J. 2017. Pre-Capture Privacy for Small Vision Sensors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ren, Z.; Lee, Y. J.; and Ryoo, M. S. 2018. Learning to Anonymize Faces for Privacy Preserving Action Detection. In *Proceedings of the European Conference on Computer Vision*.

Rosberg, F.; Aksoy, E. E.; Englund, C.; and Alonso-Fernandez, F. 2023. FIVA: Facial Image and Video Anonymization and Anonymization Defense. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.

Ryoo, M.; Rothrock, B.; Fleming, C.; and Yang, H. J. 2017. Privacy-Preserving Human Activity Recognition from Extreme Low Resolution. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. arXiv:1212.0402.

Srivastav, V.; Gangi, A.; and Padoy, N. 2019. Human Pose Estimation on Privacy-Preserving Low-Resolution Depth Images. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*.

Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; and Liu, Y. 2024. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*.

Sun, Z.; Ke, Q.; Rahmani, H.; Bennamoun, M.; Wang, G.; and Liu, J. 2023. Human Action Recognition From Various Data Modalities: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; and Paluri, M. 2018. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Van Gool, L. 2019. Temporal Segment Networks for Action Recognition in Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Weng, X.; Li, Y.; Chi, L.; and Mu, Y. 2019. High-Capacity Convolutional Video Steganography with Temporal Residual Modeling. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*.

Wu, H.-C.; Wu, N.-I.; Tsai, C.-S.; and Hwang, M.-S. 2005. Image steganographic scheme based on pixel-value differencing and LSB replacement methods. *IEE Proceedings - Vision, Image and Signal Processing*.

Wu, Z.; Wang, H.; Wang, Z.; Jin, H.; and Wang, Z. 2022. Privacy-Preserving Deep Action Recognition: An Adversarial Learning Framework and A New Dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wu, Z.; Wang, Z.; Wang, Z.; and Jin, H. 2020. Towards Privacy-Preserving Visual Recognition via Adversarial Training: A Pilot Study. arXiv:1807.08379.

Xie, X.; Cui, Y.; Tan, T.; Zheng, X.; and Yu, Z. 2024. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. *Visual Intelligence*.

Yang, L.; Fan, Y.; and Xu, N. 2019. Video Instance Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Ye, Q.; Yu, Z.; Shao, R.; Cui, Y.; Kang, X.; Liu, X.; Torr, P.; and Cao, X. 2025. CAT+: Investigating and Enhancing Audio-Visual Understanding in Large Language Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, K. A.; Cuesta-Infante, A.; Xu, L.; and Veeramachaneni, K. 2019. SteganoGAN: High Capacity Image Steganography with GANs. arXiv:1901.03892.

Zhang, Q.; Lai, J.; Zhu, J.; and Xie, X. 2023. Wavelet-Guided Promotion-Suppression Transformer for Surface-Defect Detection. *IEEE Transactions on Image Processing*.

Zhang, Z.; Cilloni, T.; Walter, C.; and Fleming, C. 2021. Multi-Scale, Class-Generic, Privacy-Preserving Video. *Electronics*.

Zhu, J.; Kaplan, R.; Johnson, J.; and Fei-Fei, L. 2018. HiDDeN: Hiding Data with Deep Networks. In *Proceedings of the European Conference on Computer Vision*.