Why DINOv2?

A Scientific Literature Review of Self-Supervised Vision Features for Compression-Robust Deepfake Video Detection

STF-Mamba V8.0 — Semantic Temporal Forensics via Hydra-Mamba and DINOv2

Abdel Rahman Madboly

Target Venue: CVPR 2026 / ICCV 2026

February 2026

Abstract

This report provides a scientific justification, grounded in the literature, for selecting DINOv2-ViT-B/14 as the backbone of STF-Mamba V8.0, a deepfake video detection system. We review three converging lines of evidence: (1) the systematic destruction of forensic artifacts by H.264 video compression, rendering statistical and frequency-domain detectors ineffective; (2) the unique properties of self-supervised vision features, particularly DINOv2, that encode semantic structure robust to compression; and (3) the empirical advantage of self-supervised ViTs over supervised CNNs for deepfake detection under realistic conditions. We further review recent state-space models (Mamba, Hydra) for temporal sequence modeling and justify their combination with DINOv2 for video-level identity consistency analysis. All claims are supported by published work from 2022–2025.

1. Introduction: The Compression Problem

The majority of deepfake detection methods published between 2018 and 2024 rely on statistical artifacts that exist in the pixel, frequency, or gradient domain of manipulated video. These include blending boundary inconsistencies, GAN fingerprints in spectral residuals, and temporal flicker patterns in wavelet sub-bands. While these signals are detectable in uncompressed or lightly compressed video, they are systematically destroyed by standard H.264 compression at the quality levels used in real-world deployment (CRF 23 or equivalent).

Nguyen et al. (WACV 2024) demonstrated this directly in their VideoFACT work: the distribution of forensic embedding distances between authentic and manipulated regions becomes indistinguishable after H.264 compression [1]. Vahdati et al. (CVPRW 2024) showed that image-trained deepfake detectors suffer severe performance degradation when applied to H.264 compressed video, with most detectors scoring AUC of 0.65 or lower [2]. Recent benchmark studies confirm that social media platforms apply aggressive compression that further launders forensic cues beyond what laboratory settings replicate [3].

This body of evidence establishes a fundamental constraint: **any detector intended for real-world deployment must use features that survive standard video compression**. This rules out pixel-level noise patterns, frequency-domain residuals, and handcrafted statistical signals as primary detection mechanisms. The question then becomes: what features are robust to compression?

2. DINOv2: Self-Supervised Visual Features

2.1 Architecture and Training

DINOv2 (Oquab et al., 2023; published in TMLR 2024) trains Vision Transformers on 142 million curated images using a self-supervised teacher-student distillation objective [4]. The student network learns to match the teacher network's output representations across different augmented views of the same image, without any human-annotated labels. The teacher is maintained as an exponential moving average of the student weights.

The resulting features are task-agnostic: the same frozen backbone achieves strong performance on classification, segmentation, depth estimation, and image retrieval without fine-tuning. This generality is critical for our application—DINOv2 was never trained on face data or forensic tasks, yet its features encode the visual structure needed to distinguish real from manipulated faces.

2.2 Why Self-Supervised Features Survive Compression

The key insight is in *what* the training objective forces the model to learn. Supervised ImageNet training (e.g., EfficientNet, ResNet) optimizes for category discrimination, which often relies on texture cues. H.264 compression specifically targets high-frequency texture information for removal through quantization of DCT coefficients. Self-supervised training, by contrast, forces the model to learn **intrinsic visual structure**—geometric relationships, object identity, and scene composition—that compression preserves.

Meta AI's own evaluation confirms this: DINOv2 features significantly outperform supervised features on out-of-domain tasks, including depth estimation on corrupted images [4]. The features exhibit strong robustness to image corruptions because the self-supervised objective makes the model invariant to the exact kind of information that compression removes.

2.3 CLS Token as Identity Embedding

The ViT architecture produces a CLS token that encodes a holistic representation of the input image. For DINOv2, this CLS token captures object-level identity without explicit recognition training. When applied to face images, the CLS token functions as a semantic face identity embedding that is stable across viewpoint changes, illumination variation, and—critically—compression artifacts.

This property is directly exploited by STF-Mamba V8.0: we track the consistency of CLS token embeddings across 32 video frames. For authentic video, these embeddings remain stable; for deepfakes, per-frame GAN synthesis introduces identity drift that manifests as measurable temporal variance in the embedding sequence.

3. DINOv2 in Deepfake Detection Literature

3.1 Self-Supervised ViTs for Forgery Detection

Nguyen et al. (2024) conducted a systematic comparison of self-supervised pre-trained ViTs versus supervised CNNs for deepfake detection [5]. Their study found that DINOv2-based feature extractors consistently outperform supervised alternatives, with DINOv2-ViT-L achieving the lowest Equal Error Rate across multiple test conditions. Critically, DINOv2 maintained its advantage even when evaluated on unseen manipulation methods (diffusion-based deepfakes), demonstrating superior generalization compared to supervised architectures.

A comprehensive evaluation by Yermakov et al. (2025) in their Generalized Deepfake Detection (GenD) work further confirmed that models trained on older, diverse datasets (like FaceForensics++) with strong backbones generalize better across time than those trained only on recent deepfakes [6]. This validates our approach of using FF++ with SBI training and a powerful DINOv2 backbone.

3.2 DINOv2 as a Generalized Backbone

A 2025 study on generalized design choices for deepfake detectors evaluated DINOv2 against ResNet-50 CLIP and ViT-L CLIP across multiple augmentation strategies and training configurations [7]. DINOv2 achieved the highest average AUROC among all evaluated backbones. Furthermore, DINOv2 reached performance plateaus faster than smaller backbones, converging after only one or two training epochs, indicating that its pre-trained features already encode highly discriminative information for forgery detection.

The study also found that DINOv2 uniquely benefits from multi-crop evaluation strategies, suggesting that its patch-level features capture fine-grained spatial information that other backbones miss. This is relevant for face forgery detection where manipulation boundaries are often localized to specific facial regions.

4. Comparison with Alternative Backbones

Backbone	Type	Params	Compression Robustness	Identity Encoding	Key Limitation

EfficientNet-B0	Supervised CNN	5.3M	Low	Weak	Texture-dependent features destroyed by H.264
EfficientNet-B4 (AdvProp)	Supervised CNN	19.3M	Medium	Medium	Requires 200 epochs, overfits on small data
ResNet-50 (CLIP)	Weakly-sup. CNN	25.6M	Medium	Medium	Text-supervised features lack local detail
ViT-L/14 (CLIP)	Weakly-sup. ViT	304M	Medium	Medium	Captions miss spatial structure information
DINOv2 ViT-B/14	Self-sup. ViT	86M	High	Strong	Requires partial freeze for small datasets

Table 1: Comparison of backbone architectures for deepfake detection on compressed video. DINOv2 (highlighted) is the only backbone combining high compression robustness with strong identity encoding.

4.1 Our Empirical Evidence

Our own ablation experiments (documented in the STF-Mamba V8.0 Scientific Document) provide direct evidence for the DINOv2 choice:

Configuration	FF++ Val AUC	Celeb-DF AUC	Finding
EfficientNet-B0, frame-level	0.6850	0.6135	Honest baseline
EfficientNet-B0 + GRU	0.5954	0.5524	Temporal on weak features = worse
EfficientNet-B4, frame-level	0.5503	0.5812	Overfits with 600 videos
DINOv2-ViT-B/14 + Hydra (V8.0)	—	0.90+ target	Strong features + temporal

Table 2: STF-Mamba ablation results. EfficientNet backbones fail because their supervised features do not encode face identity strongly enough for temporal modeling to amplify. DINOv2 provides the semantic foundation that makes temporal consistency detection possible.

The critical finding is that **temporal modeling amplifies spatial signal—it cannot create signal that does not exist**. When EfficientNet-B0 features were fed to a GRU temporal module, performance degraded by 6.1% AUC because the spatial features did not encode face identity strongly enough. The temporal module amplified noise rather than signal. This directly motivates the DINOv2 backbone: its self-supervised CLS tokens encode holistic identity information that gives the Hydra-Mamba temporal module meaningful signal to track.

5. Temporal Modeling: Why Hydra-Mamba

5.1 State Space Models for Sequence Processing

Mamba (Gu and Dao, 2023) introduced selective state space models that achieve linear-time sequence modeling with input-dependent selectivity [8]. Unlike Transformers with quadratic attention complexity, Mamba processes sequences in O(L) time while maintaining long-range memory through its hidden state. This efficiency is critical for processing 32-frame video clips where each frame produces a 768-dimensional DINOv2 embedding.

5.2 Hydra: Principled Bidirectional SSM

Hwang et al. (NeurIPS 2024) proposed Hydra, a bidirectional extension of Mamba based on quasiseparable matrix mixers [9]. Unlike ad-hoc approaches that combine separate forward and backward SSMs through addition or concatenation, Hydra provides a mathematically principled bidirectional formulation. The quasiseparable matrix structure is proven to be strictly more expressive than addition-based bidirectional SSMs, with nearly half the parameters of concatenation-based alternatives.

For deepfake detection, bidirectionality is essential: a face identity inconsistency at frame *t* is informative both from the context of preceding frames (does the identity match what came before?) and following frames (does the identity match what comes after?). Hydra processes each embedding with simultaneous access to both past and future context, making it ideal for temporal consistency verification.

5.3 Mamba in Deepfake Detection

WMamba (Peng et al., 2025) demonstrated that the Mamba architecture achieves state-of-the-art performance for face forgery detection, reaching 96.29% AUC on Celeb-DF using wavelet features with VMamba backbone [10]. However, WMamba operates on individual frames (spatial domain) rather than across video frames (temporal domain), and relies on wavelet features that our experiments show are degraded by H.264 compression. STF-Mamba V8.0 takes a complementary approach: using Hydra-Mamba in the *temporal* domain on top of compression-robust DINOv2 features.

6. The DINOv2 + Hydra-Mamba Synthesis

The combination of DINOv2 backbone and Hydra-Mamba temporal module in STF-Mamba V8.0 is motivated by three converging arguments from the literature:

Argument 1: Compression Robustness. H.264 compression destroys statistical forensic signals [1, 2, 3]. Self-supervised features survive compression because they encode structural rather than textural information [4, 7]. DINOv2 is the strongest available self-supervised vision backbone.

Argument 2: Semantic Identity Encoding. Deepfake detection through temporal identity consistency requires per-frame embeddings that faithfully encode face identity. DINOv2 CLS tokens encode holistic object identity without explicit recognition training [4, 5]. Supervised CNN features (EfficientNet) do not encode identity strongly enough for temporal modeling to work (our ablation, Table 2).

Argument 3: Efficient Bidirectional Temporal Modeling. Face identity consistency must be verified bidirectionally across the video sequence. Hydra provides mathematically principled bidirectional SSM with linear complexity [9], outperforming both Transformers and heuristic bidirectional SSMs on non-causal tasks. The variance of identity similarity across frames provides an interpretable detection signal.

7. Positioning Against Current SOTA

Method	Year	Backbone	Temporal	CDF AUC	Compression Robust?
SBI (Shiohara)	2022	EfficientNet-B4	None (frame)	93.82%	Partial
WMamba (Peng)	2025	VMamba-S	None (spatial)	96.29%	Unknown
GenD (Yermakov)	2025	DINOv2	None (frame)	Varies	Yes
UniForensics	2024	ViT (CLIP)	None (frame)	—	Partial
STF-Mamba V8.0	2026	DINOv2-ViT-B/14	Hydra-Mamba x2	0.90+ target	By design

Table 3: STF-Mamba V8.0 in context of current SOTA. Our contribution is the first to combine compression-robust self-supervised features (DINOv2) with principled bidirectional temporal modeling (Hydra-Mamba) for video-level identity consistency detection.

STF-Mamba V8.0 occupies a unique position: it is the first system that explicitly combines a compression-robust self-supervised backbone with temporal identity state modeling. Existing SOTA methods either use strong backbones without temporal modeling (SBI, GenD), or use temporal/spatial Mamba without compression-robust features (WMamba). Our architecture addresses both gaps simultaneously.

8. Conclusion

The selection of DINOv2-ViT-B/14 as the backbone for STF-Mamba V8.0 is supported by three independent lines of evidence from the literature: (1) H.264 compression destroys the statistical forensic signals that supervised CNN features rely on; (2) self-supervised training produces features encoding intrinsic visual structure that survives compression; and (3) DINOv2 specifically outperforms supervised and weakly-supervised alternatives in deepfake detection benchmarks. Combined with the Hydra-Mamba bidirectional temporal module, DINOv2 provides the semantic foundation needed to detect face identity inconsistency across video frames—a signal that compression cannot destroy because it operates at the identity level, not the pixel level.

References

[1] Nguyen, T. et al. “VideoFACT: Detecting Video Forgeries Using Attention, Scene Context, and Forensic Traces.” *WACV*, 2024.

[2] Vahdati, S. et al. “Beyond Deepfake Images: Detecting AI-Generated Videos.” *CVPR Workshop on Media Forensics*, 2024.

[3] Bridging the Gap: A Framework for Real-World Video Deepfake Detection via Social Network Compression Emulation. *arXiv:2508.08765*, 2025.

[4] Oquab, M. et al. “DINOv2: Learning Robust Visual Features without Supervision.” *Transactions on Machine Learning Research (TMLR)*, 2024. arXiv:2304.07193.

[5] Nguyen, H.H., Yamagishi, J., and Echizen, I. “Exploring Self-Supervised Vision Transformers for Deepfake Detection: A Comparative Analysis.” *arXiv:2405.00355*, 2024.

[6] Yermakov, A. et al. “Deepfake Detection that Generalizes Across Benchmarks.” *arXiv:2508.06248*, 2025.

[7] “Generalized Design Choices for Deepfake Detectors.” *arXiv:2511.21507*, 2025.

[8] Gu, A. and Dao, T. “Mamba: Linear-Time Sequence Modeling with Selective State Spaces.” *arXiv:2312.00752*, 2023.

[9] Hwang, S., Lahoti, A., Dao, T., and Gu, A. “Hydra: Bidirectional State Space Models Through Generalized Matrix Mixers.” *NeurIPS*, 2024. arXiv:2407.09941.

[10] Peng, S. et al. “WMamba: Wavelet-based Mamba for Face Forgery Detection.” *arXiv:2501.09617*, 2025.

[11] Shiohara, K. and Yamasaki, T. “Detecting Deepfakes with Self-Blended Images.” *CVPR*, 2022.

[12] Dao, T. and Gu, A. “Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality.” *ICML*, 2024. (Mamba-2).

[13] Chandra, A. et al. “DeepFake-Eval-2024: A Multi-Modal Deepfake Detection Benchmark.” 2025.

[14] Verdoliva, L. “Media Forensics and DeepFakes: An Overview.” *IEEE JSTSP*, vol. 14, no. 5, pp. 910–932, 2020.

[15] Dosovitskiy, A. et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” *ICLR*, 2021.