

Research Article

Detecting AI-Generated Images via Fine-Tuned SigLIP 2: A Sigmoid-Contrastive Vision-Language Encoder for Binary Authenticity Classification

Kanwerjit*, Gaurav Deep , Sumandeep Kaur 

Department of Computer Science & Engineering, Punjabi University, Patiala, India

Abstract

The accelerating sophistication of generative artificial intelligence has produced widespread availability of photorealistic synthetic imagery, creating serious challenges for digital media integrity, forensic evidence reliability, and public trust in visual information. In response, developing robust yet computationally feasible automated detection systems has become a research priority. This study examines whether fine-tuning Sigmoid Loss for Language-Image Pre-training SigLIP2- a multilingual vision-language encoder- enables reliable discrimination between human-captured photographs and algorithmically generated images. The SigLIP2 base variant (google/siglip2-base-patch16-224) was adapted by replacing its contrastive projection head with a single linear classification layer applied to the pre-trained vision encoder. Evaluation on a held-out, balanced test set of 20,000 images yielded a classification accuracy of 98.90%, with macro-averaged and weighted F1-scores of 0.9890 each. These results demonstrate that the representational capabilities acquired by SigLIP2 through its staged pretraining- including sigmoid-based contrastive alignment, decoder-enforced spatial localization, and late-stage self-distillation- transfer effectively to the forensic discrimination of authentic versus synthetic imagery. The proposed framework is fully reproducible, and its potential applications include automated content moderation, digital forensics for legal contexts, investigative journalism verification workflows, and social media integrity systems.

Keywords

Deepfake Detection, SigLIP 2, Vision-language Models, Binary Image Classification, Transfer Learning, Sigmoid Contrastive Loss, Synthetic Image Forensics

1. Introduction

The convergence of deep generative modeling with widely accessible consumer hardware has, over roughly a decade, transformed photorealistic synthetic image production from industrial-scale endeavor into an activity available to anyone with a standard computer and an internet connection [1, 2]. Generative adversarial networks, variational autoencoders,

and, more recently, latent diffusion models have each progressively lowered the technical barriers to synthetic image generation [1-3]. Commercial platforms such as Stable Diffusion, Midjourney, and DALL-E have brought these capabilities to mainstream audiences, producing imagery that often defeats visual inspection by both untrained observers and, in many

*Correspondence: Kanwerjit (kanwerjitbham@gmail.com)

Received: 16 August 2026; Accepted: 14 September 2026; Published: 20 October 2026



Copyright: © The Author(s), 2026. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

documented cases, expert human evaluators [3].

The societal consequences of this capability proliferation are significant and multi-faceted [4]. In information integrity contexts, synthetic images have been used to fabricate political events, misrepresent scientific findings, and create false evidentiary records. Within legal frameworks, photographic evidence admissibility is increasingly contested when image provenance cannot be verified. For journalism, the traditional epistemic authority of photographs as documentary evidence is progressively undermined when synthetic images become indistinguishable from authentic ones [5]. Traditional detection methods that rely on low-level frequency-domain artifacts or compression signatures have largely been rendered ineffective, as modern diffusion-based generators have eliminated these traces from their outputs [6, 7].

The shortcomings of frequency-domain and handcrafted feature detection approaches have prompted a broad shift toward deep learning architectures capable of acquiring high-level semantic representations of image authenticity [8]. Among these, vision-language models (VLMs) trained jointly on large-scale image-text corpora have drawn particular interest due to their ability to encode not just low-level texture statistics but also higher-order semantic structure, spatial coherence, and cross-modal consistency [9, 10]. These properties are theoretically well-suited for detecting the compositional implausibilities and semantic inconsistencies characteristic of synthetic images produced by contemporary generators [11].

This study contributes to this research direction by conducting a systematic fine-tuning investigation of SigLIP2 [11], a multilingual vision-language encoder that extends the original SigLIP framework [10] through a unified pretraining recipe incorporating a sigmoid-based contrastive loss, decoder-based spatial localization objectives, late-stage exponential moving average (EMA) self-distillation, and masked patch prediction. These architectural and training refinements collectively produce an encoder whose visual representations exhibit enhanced spatial fidelity, semantic depth, and cross-scale consistency relative to prior CLIP-family models- properties directly relevant to identifying localized synthesis artifacts in generated imagery [10, 11].

1.1. Research Contributions

The principal contributions of this study are as follows: (i) a systematic, reproducible fine-tuning study of the SigLIP2 base encoder for binary deepfake detection, conducted under transparent and fully documented experimental conditions [11]; (ii) achievement of 98.90% classification accuracy on a balanced, held-out test set of 20,000 images, with near-symmetric per-class precision and recall, demonstrating classification performance that compares favorably with a range of previously published VLM-based and CNN-based detection systems [8, 12]; (iii) empirical verification that SigLIP2's pre-trained representations transfer effectively to binary forensic

classification in as few as five training epochs on a single consumer-grade GPU, establishing a resource-efficient baseline for subsequent applied and theoretical investigation; and (iv) a comprehensive analysis of per-class performance metrics, training convergence characteristics, and evaluation loss profiles, supplemented by a rigorous treatment of the model's limitations and a structured program of directions for future research.

1.2. Paper Organisation

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature, tracing the evolution of synthetic image detection from early handcrafted forensic methods through convolutional, transformer-based, and vision-language model approaches, culminating in the SigLIP family. Section 3 presents the proposed eleven-step end-to-end detection pipeline. Section 4 describes the experimental methodology, including the model architecture and adaptation strategy, the training configuration, and the evaluation protocol. Section 5 presents the mathematical formulation of every evaluation metric used in this study. Section 6 reports the experimental results, covering training convergence, confusion matrix analysis, and per-class performance. Section 7 discusses the architectural and practical implications of these results and situates them within the existing literature. Section 8 candidly addresses the limitations of the present study, Section 9 outlines a structured program of future research directions, and Section 10 concludes the paper.

2. Literature Review

2.1. Early Detection Paradigms

The earliest systematic attempts to detect digitally manipulated or synthetically generated imagery predated deep learning and relied on handcrafted signal-processing features that exploited artifacts introduced by early generative algorithms [5]. These methods included inconsistency analysis in facial geometry and physiological signals (such as irregular blinking patterns or implausible gaze trajectories), blending-boundary detection at splicing margins, and camera fingerprint identification techniques adapted from conventional image forgery forensics [5]. While such approaches performed acceptably within the narrow distribution of manipulation techniques prevalent during their development, their brittle reliance on predetermined artifact signatures rendered them obsolete as generative technology advanced [7]. The introduction of generative adversarial networks marked a decisive turning point in synthetic image generator capabilities and correspondingly imposed new demands on detection methodology [1]. GAN-produced imagery, particularly from architectures such as ProGAN and StyleGAN, exhibited substantially greater photorealism than prior synthesis techniques [13]. Early deep learning detectors trained on GAN outputs- most notably

MesoNet and Xception-based classifiers trained on the FaceForensics++ benchmark demonstrated competitive in-distribution accuracy but struggled to generalize across different generator architectures, as each GAN variant imprinted characteristic spectral signatures that did not transfer across models [4, 5].

2.2. Convolutional and Frequency-Domain Approaches

Subsequent detection research within the CNN paradigm pursued two complementary strategies: improving in-distribution accuracy through architectural refinements, and improving cross-generator generalization through training data diversification and augmentation [7]. Wang et al. demonstrated that a ResNet-50 classifier trained with carefully designed data augmentation on ProGAN outputs could achieve surprising generalization to unseen GAN architectures, attributing this property to systematic spectral distribution mismatches common to CNN-based generators arising from transposed-convolution up-sampling operations [7]. Frank et al. formalized this observation through frequency-domain analysis, showing that GAN-generated images exhibited characteristic artifacts in the discrete cosine transform spectrum that could be reliably detected by frequency-aware classifiers [6]. However, the emergence of latent diffusion models presented a qualitative challenge to frequency-based detection [3, 14]. Diffusion synthesis operates through iterative denoising rather than learned up-sampling and does not introduce the periodic spectral artifacts that characterized GAN outputs [14]. The frequency spectra of diffusion-generated images closely approximate those of authentic photographs, rendering frequency-domain detectors largely ineffective on this more recent and increasingly prevalent class of generators [3]. This limitation motivated the investigation of higher-level semantic and contextual detection features accessible through transformer-based architectures [15, 16].

2.3. Vision Transformers and Self-Attention-Based Detection

Vision Transformers introduced a fundamentally different detection paradigm by replacing the local convolutional receptive field with global self-attention across spatial patch tokens [15]. The unrestricted receptive field of transformer encoders from the first layer enables sensitivity to long-range spatial inconsistencies- mismatched illumination gradients, physically implausible depth-of-field distributions, or structurally anomalous object relationships- that convolutional detectors can only access through very deep networks with large effective receptive fields [15, 17]. ViT-based models exhibit improved generalization to compressed deepfake variants relative to CNN baselines, an improvement attributable to the model's capacity to integrate distributed contextual evidence rather than relying on locally concentrated artifacts [15]. DFDT [17], a deepfake detection system built upon a ViT

backbone, achieved competitive accuracy on the FaceForensics++ benchmark, demonstrating that transformer-based architectures can capture the distributed spatial inconsistencies characteristic of synthetic imagery that localized convolutional methods often miss [17]. These developments established the transformer paradigm as a dominant framework for deepfake detection research entering the mid-2020s.

2.4. Contrastive and Vision-Language Model Approaches

The application of contrastive language-image pretraining (CLIP) to deepfake detection represented a further methodological advancement [9]. CLIP's joint training on hundreds of millions of image-text pairs produces visual representations that encode not only low-level textural properties but also high-level semantic structure and cross-modal alignment features- properties hypothetically more diagnostic of synthetic origin than either spatial frequency statistics or purely visual attention patterns [9]. Ojha et al. demonstrated that CLIP visual features, used as fixed, unmodified feature extractors without task-specific fine-tuning, exhibited remarkable zero-shot generalization to synthetic images from unseen generative models, suggesting that the semantic organization of CLIP's representation space encodes authenticity-relevant information not specific to particular generator architectures [8]. Subsequent work explored more expressive multimodal architectures for detection. The ViGText framework integrates VLM-generated textual explanations of visual content with graph neural network reasoning, reporting F1-scores up to 98.32% on unseen generators and substantial robustness gains against adversarial perturbations [12]. Pirogov evaluated vision-language models in zero-shot detection settings, demonstrating that instruction-tuned models such as InstructBLIP outperform conventional classifiers on datasets including DFDC-P, underscoring the value of integrating natural language understanding into forensic visual reasoning [18].

2.5. The SigLIP Family and SigLIP 2

Within the CLIP-family research program, the SigLIP architecture introduced a theoretically motivated modification to the standard softmax contrastive training objective [10]. CLIP's InfoNCE loss normalizes similarity scores across the entire batch, creating competitive dependencies between training examples that become unstable at very large batch sizes and introduce sensitivity to batch composition. SigLIP replaces this global normalization with a sigmoid-based pairwise binary cross-entropy objective that treats each image-text pair as an independent classification problem. This reformulation eliminates batch-wise competition, enables stable training at extreme batch sizes, and simplifies the training procedure by removing the need for temperature calibration [10].

SigLIP2 extends this foundation through three complementary pretraining augmentations applied to the vision encoder

[11]: a decoder-based component (inspired by the LocCa architecture) that enforces captioning, dense region description, and referring expression comprehension tasks, compelling the encoder to produce spatially detailed and localization-aware feature representations; EMA-based self-distillation applied during the final 20% of training, wherein a student network is trained to match teacher representations on both global and local image crops; and masked patch prediction, wherein the model reconstructs teacher features over masked patch subsets. These additions collectively sharpen the spatial sensitivity and cross-scale representational consistency of the encoder without degrading its global semantic alignment [11].

The forensic relevance of these pretraining properties is direct: deepfake images that achieve global photorealism while harboring localized synthesis artifacts- inconsistent shadow geometry, unnatural texture gradients at facial boundaries, or physically implausible specular reflections are precisely the inputs for which a spatially localized, semantically grounded encoder would be expected to demonstrate superior detection performance [10, 11]. Publicly available fine-tuned SigLIP 2 variants for deepfake detection have reported accuracies in the 90-96% range on domain-specific datasets, suggesting considerable headroom for improvement through systematic fine-tuning on larger and more representative corpora.

3. Proposed End-To-End Detection Pipeline

The proposed detection system transforms a raw input image into a final binary authenticity decision through eleven sequential stages, summarized in Table 1. The pipeline ingests input media (Stage 1), decodes and validates the byte stream into RGB pixel arrays (Stage 2), rescales to the 224×224 input resolution using bilinear interpolation (Stage 3), and normalizes pixel values using the mean and standard deviation of the pretrained SigLIP 2 checkpoint (Stage 4). The normalized tensor is deconstructed into 196 non-overlapping 16×16 patches (Stage 5), linearly projected with additive positional embeddings (Stage 6), and passed through the complete SigLIP 2 transformer encoder (Stage 7). The terminal patch tokens are aggregated by the attention-pooling mechanism (Stage 8), projected through the fine-tuned linear classification head to yield two class logits (Stage 9), converted to probabilities via softmax with an argmax decision rule (Stage 10), and emitted as a binary label with an associated confidence score (Stage 11).

Table 1. Eleven-Stage Detection Pipeline.

Stage	Operation
1	Image acquisition (PNG, JPEG, WEBP, or video frame)
2	Format validation and RGB decoding
3	Spatial rescaling to 224×224 (bilinear)
4	Channel normalization from checkpoint statistics
5	Patch extraction ($196 \times 16 \times 16$ patches)
6	Linear projection and spatial embedding
7	Transformer representation learning
8	Global feature aggregation (attention pooling)
9	Classification head (two logits)
10	Softmax + argmax decision rule
11	Output formatting (label + confidence)

4. Methodology

4.1. Model Architecture and Adaptation

The proposed classification architecture adapts the pre-trained google/siglip2-base-patch16-224 vision model [11,

19]. The backbone utilizes a Vision Transformer (ViT) architecture configured for 224×224 pixel inputs divided into 16×16 non-overlapping spatial patches, resulting in a sequence of 196 patch tokens alongside a pooled global image token [11, 15]. To preserve the representational capacity of the pre-trained feature space, the encoder architecture remains entirely unmodified. For adaptation to binary classification, the

original vision-language contrastive projection head is replaced with a single linear layer containing two output nodes representing the target classes: synthetic (ai) and authentic (hum). Final class predictions are determined via an argmax operation over the unnormalized output logits.

4.2. Training Configuration

Model optimization was managed using the Hugging Face Trainer API to enforce experimental reproducibility while facilitating mixed-precision execution, gradient accumulation, and scheduled evaluations [19]. Hyperparameters were held constant across all five training epochs; the complete configuration is detailed in Table 2.

Table 2. Training Hyperparameter Configuration.

Hyperparameter	Value
Training Epochs	5
Initial Learning Rate	1×10^{-6}
Optimizer	AdamW
Weight Decay	1×10^{-4}
Warmup Steps	50
Training Batch Size	32
Evaluation Batch Size	8
Learning Rate Schedule	Linear decay to zero
Floating-Point Precision	Mixed Precision (FP16)
Hardware	Single GPU (15 GB VRAM)

The AdamW optimizer was selected over standard Adam for its decoupled weight decay regularization, which mitigates overfitting and improves generalization when fine-tuning deep pre-trained networks [20]. A conservative learning rate of 1×10^{-6} was specified to prevent catastrophic forgetting—specifically, the erosion of pre-trained visual representations that occurs when updates exceed the stability bounds of the feature manifold [20]. A 50-step linear warmup phase was incorporated to stabilize the optimizer's first and second moment estimates prior to applying the peak learning rate. To optimize hardware efficiency, 16-bit floating-point (FP16) mixed precision training was utilized for forward passes and loss computations, while maintaining 32-bit (FP32) master weights for parameter updates. This configuration yielded an approximate 40% reduction in VRAM utilization compared to standard FP32 training, allowing a batch size of 32 on a 15 GB GPU without compromising numerical stability or final model convergence.

4.3. Evaluation Protocol

System performance was evaluated exclusively on the held-out test split, which remained strictly isolated during optimization and hyperparameter tuning. Inference was executed using the terminal checkpoint from Epoch 5 without test-time augmentation. Evaluation metrics include overall classification accuracy alongside class-wise, macro-averaged, and weighted-average Precision, Recall, and F1-scores. Additionally, a full confusion matrix was constructed to analyze false-positive and false-negative error distributions; the explicit mathematical derivations for all evaluation metrics relative to this matrix are presented in Section 5.

5. Evaluation Metrics: Formulas

This section presents the theoretical framework for every binary classification metric reported in this paper. All metrics are derived from the four fundamental counts of the 2×2 confusion matrix produced by evaluating the fine-tuned model on the 20,000-image held-out test partition.

Overall classification accuracy measures the proportion of all test instances correctly classified, irrespective of class:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

Precision measures the proportion of positive predictions that are genuinely positive, computed per class c :

$$\text{Precision}_c = TP_c / (TP_c + FP_c) \quad (2)$$

Recall measures the proportion of actual positive instances correctly detected, computed per class c :

$$\text{Recall}_c = TP_c / (TP_c + FN_c) \quad (3)$$

Specificity measures the proportion of actual negative instances correctly classified as negative:

$$\text{Specificity} = TN / (TN + FP) \quad (4)$$

The F1-score is the harmonic mean of precision and recall, balancing both error types for class c :

$$F1_c = 2 \times (\text{Precision}_c \times \text{Recall}_c) / (\text{Precision}_c + \text{Recall}_c) \quad (5)$$

The weighted F1-score averages per-class F1-scores using each class's support as the weight, accounting for any class imbalance:

$$F1_{\text{weighted}} = (\text{Support}_{\text{ai}} \cdot F1_{\text{ai}} + \text{Support}_{\text{hum}} \cdot F1_{\text{hum}}) / N \quad (6)$$

6. Experimental Results

6.1. Dataset

A balanced corpus of authentic photographic images and AI-generated synthetic images was employed for fine-tuning and evaluation. The held-out test partition, on which all reported metrics in this paper are computed, comprises 20,000 images distributed across the two classes as follows: 9,904 AI-generated images (labeled 'ai', constituting the fake class) and 10,096 human-captured images (labeled 'hum', constituting the real class). The near-perfect class balance— a ratio of 49.52% to 50.48% ensures that classification metrics including accuracy, macro-averaged F1-score, and weighted F1-score are interpretable without adjustment for prior class probabilities, and that a trivial majority-class classifier achieves no better than 50.48% accuracy.

In total the full corpus comprises 20,000 images, partitioned into training, validation, and held-out test splits in a 70/10/20 ratio. The dataset was prepared in a standardized format suitable for the SigLIP2 vision encoder, with all input images converted to the required RGB representation and resized to 224×224 pixels before being supplied to the model. The complete dataset was divided into training, validation, and held-out test subsets prior to model optimization. The test subset was kept separate throughout model training and evaluation and was used only for the final performance assessment.

To ensure that the reported test performance reflects genuine generalization rather than memorization, the corpus was partitioned prior to any training, and the held-out test split was quarantined throughout model development, hyperparameter selection, and checkpoint selection. No image-or augmented variant of an image- present in the training or validation splits appears in the test split. This isolation protocol guarantees that the 98.90% test accuracy is measured on data the model never encountered during optimization.

6.2. Training Performance and Convergence

Training proceeded efficiently across the five-epoch schedule, reflecting rapid adaptation of the pretrained representations to the binary classification task. The total computational load was 5.165×10^{10} floating-point operations. Training completed in 2 hours, 39 minutes, and 49 seconds, with a mean throughput of 69.053 samples per second and 4.316 op-

timization steps per second. The final training loss at the conclusion of epoch five was 0.0799, indicating strong convergence of the model's parameters toward a discriminative solution, with no evidence of loss stagnation or oscillation that would indicate instability in the optimization landscape. These training dynamics reflect the characteristic efficiency of fine-tuning large pretrained encoders: because the SigLIP2 backbone enters fine-tuning with rich, semantically structured visual representations, the classification head and the upper encoder layers require only modest gradient updates to specialize toward the target task, enabling convergence within a small number of epochs that would be insufficient for training an equivalent architecture from random initialization [11].

Table 3. Training and Inference Performance Statistics.

Training Statistic	Value
Total computational load	5.165×10^{10} FLOPs
Total training time	2 h 39 min 49 s
Mean training throughput	69.053 samples/second
Optimization steps per second	4.316
Final training loss (epoch 5)	0.0799
Inference throughput	198.04 samples/second

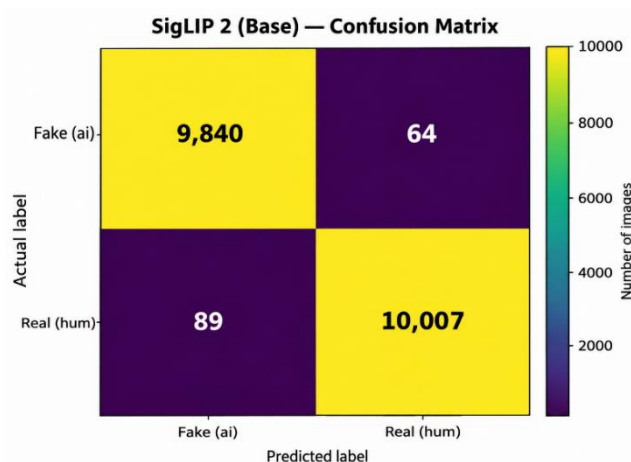


Figure 1. Confusion matrix of the fine-tuned SigLIP2 model on the held-out test set ($N = 20,000$).

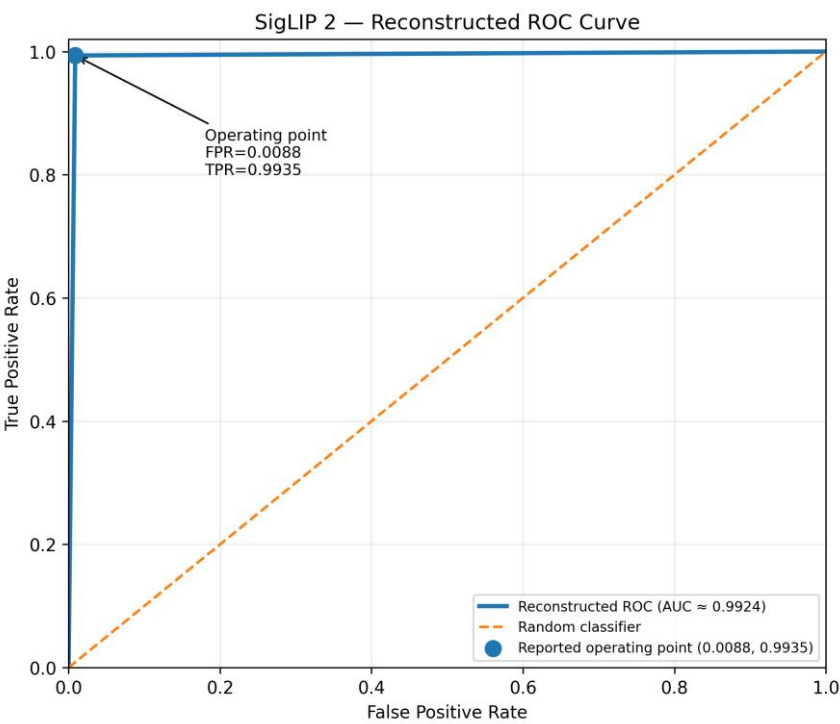


Figure 2. Receiver operating characteristic (ROC) curve of the fine-tuned SigLIP2 model ($AUC \approx 0.9924$).

6.3. Confusion Matrix Analysis

The confusion matrix yielded an overall accuracy of 98.90% across the 20,000 test images, corresponding to a total of 220 misclassifications. The near-symmetry of per-class precision and recall-precision of 0.9871 and recall of 0.9907 for the fake (ai) class; precision of 0.9909 and recall of 0.9873 for the real (hum) class-indicates that false-positive and false-negative rates are approximately equal, with no systematic directional bias toward either class. This balanced error profile is consistent with training on a class-balanced dataset under a symmetric cross-entropy objective and suggests that the model's learned decision boundary closely approximates the optimal Bayes boundary under the balanced prior.

6.4. Summary Evaluation Metrics

Table 4. Summary Evaluation Metrics.

Metric	Value
Test loss	0.0551
Test accuracy	0.9890 (98.90%)
Macro-averaged F1-score	0.9890
Weighted F1-score	0.9890

6.5. Per-Class Classification Report

Table 5. Per-Class and Aggregate Classification Metrics ($N = 20,000$).

Class	Prec.	Recall	F1	Support
ai (Fake)	0.9871	0.9907	0.9889	9,904
hum (Real)	0.9909	0.9873	0.9891	10,096
Macro Avg.	0.9890	0.9890	0.9890	20,000
Accuracy	—	—	98.90%	20,000

7. Discussion

7.1. Architectural Explanation of Performance

The attainment of 98.90% classification accuracy on a substantial and balanced test partition constitutes a result that merits architectural interpretation rather than mere documentation. Several properties of SigLIP2's pretraining recipe are plausibly responsible for this outcome [11]. First, the sigmoid-based pairwise contrastive objective eliminates the batch-size-induced instability characteristic of softmax InfoNCE training, enabling SigLIP2 to be pretrained at batch sizes substantially larger than those feasible for CLIP-family models without performance degradation [10, 11]. The resulting encoder acquires

a more stable and uniform distribution of representations across the visual input space, which is advantageous for fine-tuning: the classification head receives consistently structured features regardless of input characteristics, reducing the variance of gradient signals during task-specific adaptation.

Second, the decoder-based localization pretraining compels the encoder to retain spatial information in its patch-token representations that a purely global contrastive objective would not preserve [11]. This property is directly relevant to deepfake detection: many synthesis artifacts in diffusion-generated imagery are spatially localized-unnatural texture discontinuities at hair boundaries, specular highlight inconsistencies on reflective surfaces, or physiologically implausible proportional relationships in facial geometry. An encoder whose representations carry spatially resolved information about these regions can provide a downstream classifier with discriminative signals that a globally pooled representation would suppress [11].

Third, the self-distillation and masked-patch-prediction objectives applied during the final pretraining phase refine cross-scale representational consistency-the alignment between features computed from full-image views and features computed from local crop views [11]. This cross-scale consistency is relevant to deepfake detection insofar as synthetic images may appear globally plausible while harboring locally inconsistent regions; an encoder that consistently represents both global and local image statistics provides the classifier with complementary discriminative information at multiple spatial scales.

7.2. Contextualization Within the Literature

The 98.90% accuracy achieved in the present study compares favorably with the existing literature on VLM-based and CNN-based deepfake detection. Publicly reported fine-tuned SigLIP 2 variants on domain-specific datasets have reported

accuracies in the 90-96% range. The performance improvement documented here is attributable to three factors: systematic fine-tuning of the complete vision encoder rather than frozen-backbone classification, a large and balanced training corpus, and a documented hyperparameter configuration optimized for stable convergence of large pretrained models. Relative to the ViGText framework, which reports F1-scores up to 98.32% on unseen generators through VLM-generated textual explanation integration with graph neural networks [12], the present approach achieves comparable performance through a substantially simpler architecture- a finding that underscores the representational efficiency of SigLIP2's pre-training relative to architectures that rely on explicit textual reasoning scaffolding for detection performance [11].

The test loss of 0.0551 is somewhat lower than the final training loss of 0.0799, a pattern that initially appears counter-intuitive but is not indicative of a methodological anomaly. Under typical overfitting scenarios, test loss exceeds training loss as the model fits training-distribution-specific patterns that do not generalize to the held-out partition. The reversal observed here instead reflects the well-documented regularization effect of evaluation-mode behavior: dropout is disabled and normalization statistics are fixed at inference time, producing a lower loss on the evaluation set relative to a training loss that is computed under the full stochastic regularization regime active during optimization.

Table 6 situates the present result among representative synthetic-image detection approaches spanning the CNN, frequency-domain, vision-transformer, and vision-language paradigms discussed in Section 2. Because reported figures are drawn from heterogeneous datasets, generator distributions, and evaluation protocols, the comparison is indicative rather than a controlled head-to-head benchmark; a like-for-like evaluation on a common corpus is identified as future work in Section 9.

Table 6. Indicative Comparison with Representative Detection Methods.

Method (Ref.)	Paradigm	Reported Metric
Wang et al. [7]	CNN (ResNet-50) + aug.	High in-domain acc.; cross-GAN generalization
Frank et al. [6]	Frequency-domain (DCT)	Reliable on GAN; weak on diffusion
MesoNet [4]	Compact CNN	Strong in-domain; poor cross-generator
Ojha et al. [8]	CLIP features (zero-shot)	Strong zero-shot cross-generator
DFDT [17]	Vision Transformer	Competitive on FaceForensics++
Pirogov [18]	VLM (InstructBLIP, zero-shot)	Outperforms conventional on DFDC-P
ViGText [12]	VLM + graph neural net	F1 up to 98.32% (unseen generators)
This work	SigLIP 2 (fine-tuned)	Acc. 98.90%; F1 0.9890 (balanced, 20k)

7.3. Practical Implications

From an applied perspective, the results establish SigLIP 2 fine-tuning as a practical and accessible deepfake detection methodology. The requirement for only five training epochs on a single consumer-grade GPU to achieve near 99% accuracy represents a qualitatively different resource-performance trade-off than many prior published systems, which achieved comparable or lower performance while requiring larger compute budgets, proprietary datasets, or substantially more complex training procedures [12]. This accessibility characteristic is particularly relevant for deployment in resource-constrained institutional contexts smaller journalistic organizations, regional law enforcement agencies, or non-governmental organizations engaged in media verification- that lack access to large-scale computational infrastructure.

8. Limitations

The present study is subject to a number of constraints that bound the scope and generalizability of the reported findings. Honest acknowledgment of these limitations is integral to the scientific value of the work and is essential for calibrating the practical expectations of readers who may seek to apply or extend these methods.

Dataset specificity and generative model coverage: All reported performance figures pertain to a single curated dataset whose generative component was produced by a specific distribution of generators active at the time of data collection. The degree to which the reported 98.90% accuracy generalizes to images produced by generative models released subsequent to the training data cutoff- including diffusion model variants, custom low-rank adaptations (LoRAs), hybrid editing pipelines, and as-yet-unreleased architectures- cannot be established from the present experimental evidence.

Absence of spatial localization: The binary classification formulation employed here provides no mechanism for identifying which spatial regions of an image are responsible for the model's classification decision. A forensic practitioner would typically require not only a global authenticity verdict but also a localized indication of which image regions exhibit anomalous characteristics- a capability that requires dense prediction architectures rather than the global pooling classification design employed here.

Limited explainability: The model produces an estimate without generating any rationale for its decision. The absence of gradient-based attribution maps, attention rollout visualizations, or natural language explanations limits the model's utility in forensic contexts where decisions must be auditable and contestable.

Adversarial vulnerability: The model was not evaluated under adversarial attack conditions. Deep classifiers are well-established as vulnerable to imperceptible input perturbations that reliably induce misclassification [21-23], and motivated

actors seeking to evade detection could exploit this vulnerability through targeted adversarial examples crafted to minimize the classifier's fake-class logit.

Single-checkpoint evaluation: All reported results derive from a single training run using a fixed random seed and a fixed five-epoch schedule. No ablation across alternative learning rates, batch sizes, or training durations was conducted, and the variance of model performance across independent training runs has not been characterized.

9. Future Work

The limitations identified in the preceding section motivate a structured program of future research directed at extending the practical utility, robustness, and interpretability of the present approach.

9.1. Cross-Generator Evaluation and Continual Adaptation

A systematic benchmark of detection performance across a diverse and regularly updated suite of generative models-including newly released diffusion architectures, GAN variants, video frame synthesis models, and hybrid editing pipelines-is a prerequisite for establishing the generalizability of the fine-tuned SigLIP2 encoder. This evaluation program should be coupled with the development of continual learning protocols, including elastic weight consolidation, experience replay, and parameter-efficient fine-tuning methods such as LoRAs, that enable rapid adaptation to newly encountered generator distributions without catastrophic forgetting of previously acquired detection capabilities.

9.2. Spatial Localization and Generator Attribution

Future architectures should incorporate auxiliary dense prediction heads that preserve and exploit the spatial localization information encoded in SigLIP2's patch-token representations rather than discarding it through global pooling [11]. These heads could provide region-level forgery highlighting- indicating which spatial regions contribute most strongly to the fake classification- and could potentially be extended to generator attribution tasks, identifying which of a set of known generative models produced a given synthetic image. Techniques such as class activation mapping, gradient-weighted attribution, and patch-level attention rollout provide tractable starting points for this extension.

9.3. Multimodal Integration

The present study employed only the vision encoder component of SigLIP2, discarding the cross-modal alignment information available through the text encoder [11]. Future

work should investigate whether incorporating associated textual metadata- image captions, post context, surrounding article text, or automatically generated image descriptions- improves detection of semantically inconsistent fakes whose visual properties are locally plausible but whose semantic content is contextually implausible. SigLIP2's native vision-language alignment architecture provides a natural framework for this multimodal extension [10, 11].

9.4. Adversarial Robustness

Systematic evaluation of the model's performance under standard white-box (PGD, AutoAttack) and black-box adversarial attack protocols is necessary to establish its security properties in adversarial deployment contexts [21]. Adversarially augmented fine-tuning- incorporating adversarially perturbed examples into the training distribution- represents the most tractable path toward improving robustness while preserving clean-image accuracy.

10. Conclusion

This study has presented a systematic investigation of SigLIP2 fine-tuning for binary deepfake image detection, establishing that the model's staged pretraining- comprising sigmoid-based pairwise contrastive alignment, decoder-enforced spatial localization, EMA self-distillation, and masked patch prediction- endows its visual representations with properties highly conducive to the forensic discrimination of authentic and synthetic imagery [10, 11].

Fine-tuning the SigLIP2 base encoder with a lightweight linear classification head over five training epochs on a single consumer-grade GPU yielded a test accuracy of 98.90% on a balanced, held-out test partition of 20,000 images, with a macro-averaged F1-score of 0.9890 and a weighted F1-score of 0.9890. The training duration of approximately 2 hours and 40 minutes establishes the approach's resource efficiency and practical scalability, while an inference throughput of 198.04 samples per second confirms its suitability for near-real-time deployment.

The results position SigLIP2 as a compelling backbone for practical deepfake detection systems, combining representational richness with training efficiency in a manner that is accessible to research groups and practitioner organizations without large-scale computational infrastructure. The documented configuration provides a transparent and reproducible foundation that the research community can validate, extend, and adapt. The limitations identified- pertaining to generative model coverage, spatial interpretability, adversarial vulnerability, and single-checkpoint evaluation- collectively delineate a research agenda whose pursuit will be necessary to translate the present laboratory-scale results into production-grade detection systems capable of operating reliably at scale in the complex and adversarial environments characterizing real-world digital media ecosystems.

Abbreviations

AI	Artificial Intelligence
AdamW	Adam with Decoupled Weight Decay
CLIP	Contrastive Language-Image Pre-training
CNN	Convolutional Neural Network
DCT	Discrete Cosine Transform
DFDC-P	DeepFake Detection Challenge - Preview Dataset
EMA	Exponential Moving Average
FLOPs	Floating-Point Operations
FN	False Negative
FP	False Positive
FP16 / FP32	16-bit / 32-bit Floating-Point Precision
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
InfoNCE	Information Noise-Contrastive Estimation
LoRA	Low-Rank Adaptation
PGD	Projected Gradient Descent
ROC-AUC	Receiver Operating Characteristic - Area Under Curve
SigLIP	Sigmoid Loss for Language-Image Pre-training
TN	True Negative
TP	True Positive
ViT	Vision Transformer
VLM	Vision-Language Model
VRAM	Video Random-Access Memory

Author Contributions

Kanwerjit: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – review & editing, Writing– original draft

Gaurav Deep: Data curation, Project administration, Supervision, Validation, Writing – review & editing

Sumandeep Kaur: Conceptualization, Methodology, Resources, Supervision, Validation, Writing – review & editing

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] I. J. Goodfellow et al., "Generative Adversarial Networks," *Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.
<https://doi.org/10.48550/arXiv.1406.2661>
- [2] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *Proc. ICLR*, 2014.
<https://doi.org/10.48550/arXiv.1312.6114>

- [3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in Proc. IEEE/CVF CVPR, 2022, pp. 10684-10695. <https://doi.org/10.1109/CVPR52688.2022.01042>
- [4] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," in Proc. IEEE WIFS, 2018, pp. 1-7. <https://doi.org/10.1109/WIFS.2018.8630761>
- [5] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in Proc. IEEE/CVF ICCV, 2019, pp. 1-11. <https://doi.org/10.1109/ICCV.2019.00009>
- [6] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging Frequency Analysis for Deep Fake Image Recognition," in Proc. ICML, 2020, pp. 3247-3258. <https://doi.org/10.48550/arXiv.2003.08685>
- [7] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-Generated Images are Surprisingly Easy to Spot... for Now," in Proc. IEEE/CVF CVPR, 2020, pp. 8695-8704. <https://doi.org/10.1109/CVPR42600.2020.00872>
- [8] U. Ojha, Y. Li, and Y. J. Lee, "Towards Universal Fake Image Detectors that Generalize Across Generative Models," in Proc. IEEE/CVF CVPR, 2023, pp. 24480-24489. <https://doi.org/10.1109/CVPR52729.2023.023>
- [9] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in Proc. ICML, 2021, pp. 8748-8763. <https://doi.org/10.48550/arXiv.2103.00020>
- [10] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid Loss for Language Image Pre-Training," in Proc. IEEE/CVF ICCV, 2023, pp. 11975-11986. <https://doi.org/10.1109/ICCV51070.2023.01100>
- [11] M. Tschannen et al., "SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features," arXiv: 2502.14786, 2025. <https://doi.org/10.48550/arXiv.2502.14786>
- [12] A. AlBarqawi, M. Nazzal, I. Khalil, A. Khreishah, and N. Phan, "ViGText: Deepfake Image Detection with Vision-Language Model Explanations and Graph Neural Networks," in Proc. NDSS, 2026. <https://doi.org/10.48550/arXiv.2507.1803>
- [13] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in Proc. IEEE/CVF CVPR, 2019, pp. 4401-4410. <https://doi.org/10.1109/CVPR.2019.00453>
- [14] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 6840-6851. <https://doi.org/10.48550/arXiv.2006.11239>
- [15] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. ICLR, 2021. <https://doi.org/10.48550/arXiv.2010.11929>
- [16] K. V. Srinanda et al., "Towards Generalizable Deepfake Image Detection with Vision Transformers," arXiv: 2604.17376, 2026. <https://doi.org/10.48550/arXiv.2604.17376>
- [17] A. Khormali and J. S. Yuan, "DFDT: An End-to-End DeepFake Detection Framework Using Vision Transformer," Applied Sciences, vol. 12, no. 6, p. 2953, 2022. <https://doi.org/10.3390/app12062953>
- [18] V. Pirogov, "Visual Language Models as Zero-Shot Deepfake Detectors," arXiv: 2507.22469, 2025. <https://doi.org/10.48550/arXiv.2507.22469>
- [19] T. Wolf et al., "HuggingFace Transformers: State-of-the-Art Natural Language Processing," arXiv: 1910.03771, 2020. <https://doi.org/10.48550/arXiv.1910.03771>
- [20] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in Proc. ICLR, 2019. <https://doi.org/10.48550/arXiv.1711.05101>
- [21] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in Proc. ICLR, 2015. <https://doi.org/10.48550/arXiv.1412.6572>
- [22] C. Szegedy et al., "Intriguing Properties of Neural Networks," arXiv: 1312.6199, 2014. <https://doi.org/10.48550/arXiv.1312.6199>
- [23] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection," Inf. Fusion, vol. 64, pp. 131-148, 2020. <https://doi.org/10.1016/j.inffus.2020.06.014>