SAFE: Semantic- and Frequency-Enhanced Curriculum for Cross-Domain Deepfake Detection

Yulin Yao*1, Kangfeng Zheng*1†, Bin Wu*1, Chunhua Wu*1, Jujie Wang*1, Jiaqi Gao*1, Minjiao Yang*1, Dan Luo*1

[1]School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, China
{yuliny, kfzheng, binwu, wuchunhua, wangjvjie, jessie_gao, yangminjiao, luodan}@bupt.edu.cn

Abstract

Driven by advances in GANs and diffusion models, deepfake content has reached an unprecedented level of photorealism, causing detectors to deteriorate once they leave their training domain. Most prior studies adopt CLIP as the backbone of an image-level binary classifier, yet overlook CLIP's core strength: text-to-image semantic alignment. Moreover, captions generated by CLIP-CAP lack sufficient high-level semantics to distinguish between authentic and manipulated faces. Deepfake generators often fail to maintain semantic coherence, resulting in contradictions that traditional visual models cannot capture. Existing approaches also intermingle all samples during training and thus lack a systematic, difficulty-aware curriculum. To bridge these gaps, we introduce Semantic- and Frequency-Enhanced (SAFE) deepfake detection, a two-component framework: 1) Semantic-enhanced multimodal alignment. Authenticity cues are injected into CLIP-CAP captions, and low-rank LoRA fine-tuning is applied to CLIP's visual branch, yielding dual supervision for text–image alignment and forgery discrimination. 2) Dual-score curriculum learning. Fourier Correlation Variance (FCV) measures local spectral consistency and, combined with the loss value, is transformed into a difficulty score that ranks training samples from easy to hard, reducing training time by 23.3% and enhancing generalization. SAFE attains state-of-the-art performance on several cross-dataset and cross-manipulation benchmarks. Ablation studies confirm that semantic enhancement, LoRA fine-tuning, and dual-score curriculum are complementary, jointly delivering substantial gains in open-set generalization.

Code — https://github.com/kingkongs7/SAFE

Introduction

Generative artificial intelligence (including generative adversarial networks, diffusion models, and an array of deepfake techniques) is advancing at an unprecedented pace, now able to create visual, auditory, and textual content that is virtually indistinguishable from authentic data. Although these methods offer substantial benefits for film production and artistic creation, their malicious use threatens personal privacy, erodes public trust, and increasingly outpaces conventional detection techniques (Li, Chang, and Lyu 2018; Li et al. 2020; Thies, Zollhöfer, and Nießner 2019). Consequently, deepfake detection has become a leading research focus at the intersection of multimedia security and computer vision. Most current studies frame the task as a binary "real versus fake" classification problem and report near-saturated accuracy under controlled conditions (Hasanaath et al. 2025; Yan et al. 2023b, 2024a). Such closed-set evaluations, however, fail to capture real-world scenarios in which previously unseen manipulation techniques continually emerge: regardless of whether they employ convolutional networks or Transformers, state-of-the-art detectors suffer substantial performance degradation when confronted with cross-domain shifts or entirely novel forgeries (Ojha, Li, and Lee 2023). Enhancing detectors' open-set generalization capability has therefore become the field's principal bottleneck (Yan et al. 2024b).

Since prior work demonstrated the strong deepfake detection capability of CLIP (Fu et al. 2025), most subsequent

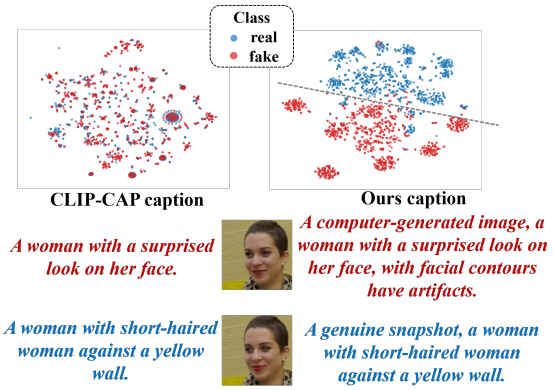

Figure 1: A t-SNE visualization of the latent embedding space shows that captions generated by our method form semantically distinct clusters, whereas CLIP-CAP captions remain intermingled.

*These authors contributed equally.
†Corresponding author: kfzheng@bupt.edu.cn

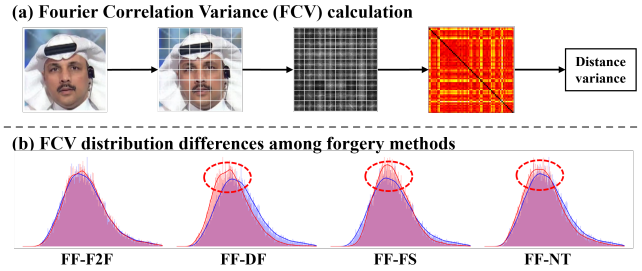

Figure 2: FCV and curriculum-learning analysis. (a) Pipeline for computing Fourier Correlation Variance (FCV). (b) Comparison of FCV distributions for forged and authentic images on FF++, with representative examples.

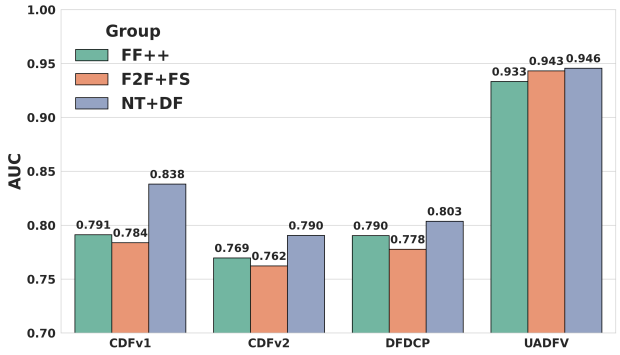

Figure 3: Grouped-test performance on FF++.

studies have adopted CLIP as the backbone network (Yan et al. 2024a; Fu et al. 2025; Haliassos et al. 2022; Wang et al. 2023; **?**). However, these works continue to treat the problem as an image-level binary classification task and overlook the fact that CLIP's strength originates from its alignment of semantic information between text and images. Although C2C-CLIP addressed this issue in a broader deepfake setting (Tan et al. 2025), it ultimately concluded that CLIP detects forgeries primarily through concept-level similarity. In this study, we use CLIP-CAP (Mokady, Hertz, and Bermano 2021) to generate textual descriptions of both authentic and manipulated face images. We find that the two classes receive nearly identical high-level semantics (see Figure 1), suggesting that this approach is ill-suited to the more specific task of facial deepfake detection.

Deepfake artifacts arise fundamentally from a misalignment of semantic-level authenticity: generative models can reproduce low-level attributes such as skin tone and texture, yet they often fail to maintain global semantic coherence. In face-swap videos, for example, facial geometry may conflict with the scene's natural illumination (Amerini et al. 2019). Purely visual detectors concentrate on pixel-level realism and therefore overlook such cross-level contradictions, resulting in poor generalization. This limitation also clarifies why CLIP's cross-modal alignment yields stronger performance in broader classification tasks. When human observers sense that *"something is wrong"* with a video, they typically frame an internal description (e.g., *"the skin texture looks odd"* or *"her expression seems unnatural"*) in effect performing an image–language consistency check. A model that jointly interprets images and text and judges their semantic coherence can therefore expose forgeries at the semantic level.

We operationalize this intuition by injecting discriminative prompt words into CLIP-CAP captions. As illustrated in Figure 1, each caption is augmented with generic descriptors that highlight artifacts frequently associated with deepfakes. For instance, the forged caption *"A woman with a surprised look on her face"* is extended to *"A computer-generated image: a woman with a surprised look on her face, with facial contours exhibiting artifacts,"* whereas the authentic caption *"A short-haired woman against a yellow wall"* becomes *"A genuine snapshot: a short-haired woman against a yellow wall."* These prompts are algorithm-agnostic and summarize inconsistencies most commonly noticed by human viewers.

FaceForensics++ (FF++) (Rossler et al. 2019) has become a benchmark for deepfake detection. It comprises authentic images alongside four manipulation types: DeepFakes (DF), Face2Face (F2F), NeuralTextures (NT), and FaceSwap (FS). Most studies merge all samples during training, implicitly assuming that a network learns equally from every instance. Only a handful have explored curriculum learning. CDFA first pre-trains on FF++ and then introduces out-of-domain images. While effective, this schedule is empirical and lacks theoretical justification (Lin et al. 2024). DFFC combines face-confidence scores with instantaneous resampling (Song, Lin, and Li 2024), yet the scoring rationale itself remains under-examined. In summary, the field still lacks a lightweight, physically interpretable difficulty metric that can drive a principled curriculum for deepfake detection.

Block-level Fourier analysis reveals that the variance of inter-block spectral similarity cleanly separates authentic images from forgeries produced by diverse manipulation pipelines. Building on this insight, we introduce Fourier Correlation Variance (FCV). Its computation pipeline appears in Figure 2a. As shown in Figure 2b, the FCV distributions for NT and DF lie significantly below those of genuine images, indicating a quantifiable hierarchy of forgery difficulty along the axis of local frequency consistency. We therefore split FF++ into a high-discrepancy subset (NT, DF) and a low-discrepancy subset (FS, F2F) for evaluation. Remarkably, a detector trained solely on the high-discrepancy subset outperforms one trained on the full dataset (Figure 3. This observation motivates our curriculum strategy: each sample receives a composite difficulty score comprising a static FCV term and a dynamic loss term. Training begins with high-score forgeries and incrementally introduces harder examples, enabling the network to consolidate robust coarse discrimination before adapting to fine-grained manipulations.

To address the dual bottlenecks of semantic misalignment and sample-difficulty imbalance, we consolidate the foregoing insights into an end-to-end Semantic- and Frequency-Enhanced (SAFE) deepfake detection framework comprising two components: 1) Semantic-enhanced multimodal

alignment. We enrich CLIP-CAP captions with authenticity/forgery prompts and generic artifact cues, and insert a 0.09M parameter low-rank LoRA adapter solely into CLIP's visual branch while freezing the remaining weights. A binary cross-entropy loss jointly supervises text–image consistency and forgery discrimination. 2) Dual-score curriculum learning. A static difficulty term derived from Fourier Correlation Variance (FCV) is combined with a dynamic loss term to rank samples. The network first encounters high-score (easy) forgeries and progressively incorporates more complex examples, thereby reducing training time and enhancing generalization. The guiding principle is straightforward: first align semantics, then advance from easy to hard in the frequency domain.

Our work has made the following key contributions:

- **SAFE framework.** We present SAFE, the first multimodal deepfake detector that injects authenticity prompts into CLIP-CAP captions and couples them with low-rank LoRA fine-tuning. By updating only 0.09M parameters, SAFE achieves state-of-the-art results on 11 public benchmarks.
- **Semantic-consistency gap.** We demonstrate that augmenting CLIP-CAP captions with generic authenticity/-forgery phrases and common artifact cues creates a quantifiable semantic gap between genuine and forged images in CLIP space.
- **Fourier Correlation Variance (FCV).** We introduce FCV, a frequency-domain difficulty metric that relies solely on block-level fast Fourier transforms to measure local spectral consistency, establishing, for the first time, a hierarchy of forgery difficulty within FF++.
- **Dual-score curriculum.** We devise a curriculum that orders samples from easy to hard according to FCV and loss value. Training first on high-discrepancy forgeries and then on lower-discrepancy ones reduces training time by 23.3% while preserving accuracy.

Related Work

Traditional deepfake detection methods. Most work treats deepfake detection as an image-level binary-classification task and advances along three principal axes: 1) Fine-grained texture and geometric cues. Early studies exploit subtle inconsistencies in eye geometry and reflections, such as abnormal blinking (Liy and InIctuOculi 2018), head-pose mismatches (Yang, Li, and Lyu 2019), optical flow anomalies (Amerini et al. 2019), and compression artifacts (Li and Lyu 2018). 2) Frequency-domain and multi-resolution analysis. FreqDebias mitigates overfitting through frequency-domain augmentation and consistency regularization (Kashiani, Talemi, and Afghah 2025), whereas FSBI couples wavelet transforms with forged-data augmentation to improve cross-domain generalization (Hasanaath et al. 2025). Other approaches fuse spatial and spectral features to enhance further robustness (Qian et al. 2020; Wu et al. 2020; Luo et al. 2021). 3) Architectural and training paradigm innovation. Following CLIP's strong baseline performance (Fu et al. 2025), many detectors adopt CLIP-based backbones (Liu et al. 2024a; Dong et al. 2023; Chen et al. 2024; Liu et al. 2024b). Parameter-efficient fine-tuning (PEFT) has also gained traction: Effort leverages singular value decomposition (SVD) to preserve pre-training knowledge while capturing diverse forgery patterns (Yan et al. 2025), and subsequent work explores mixture-of-experts (MoE) designs (Kong et al. 2024; Liu 2024) and low-rank adaptation (LoRA) modules (Kong, Li, and Wang 2023).

Multimodal deepfake detection. Multimodal vision-language models (MLLMs) have achieved state-of-the-art performance across numerous benchmarks (Li et al. 2023a,b, 2022). As deepfake content expands from purely visual clips to fully multimodal videos, researchers increasingly exploit cross-modal inconsistencies to enhance detection. For example, the CAD framework fuses visual and audio cues via semantic alignment and cross-modal distillation (Du et al. 2025). At the same time, M2F2-Det employs forgery-prompt learning to embed manipulation patterns in CLIP's textual space (Guo et al. 2025). More broadly, C2C-CLIP aligns textual semantics with imagery (Tan et al. 2025); however, its generated verbal descriptions are not tightly coupled with the forged images, rendering the representations separable in some latent spaces and limiting interpretability.

Curriculum learning for deepfake detection. Curriculum learning emulates human cognition: models are exposed to training data in a progressively harder sequence so that salient cues are learned first and subtler artifacts later (Bengio et al. 2009). This idea has informed work across multiple domains (Duan et al. 2020; Ranjan and Hansen 2017; Kocmi and Bojar 2017). Within deepfake detection, DFFC assigns each sample a dynamic forgery-difficulty score and employs a pacing function to schedule presentation during training (Song, Lin, and Li 2024). CDFA reports that withholding forgery augmentation in the early stage accelerates convergence because an untrained model cannot yet recognize raw forged inputs; introducing p-fake[1] too early is counterproductive. Broader forgery-augmentation strategies can also enrich learned representations (Lin et al. 2024), though the underlying mechanisms remain underexplained. Beyond classical curricula, other difficulty-adaptive regimes have been explored: Alt-Freezing alternately freezes network layers to mitigate overfitting to specific artifacts (Wang et al. 2023); continual-learning approaches preserve historical distributions to resist catastrophic forgetting as new forgeries emerge (Pan et al. 2023); and one-class detectors trained solely on (augmented) real data improve sensitivity to unseen manipulations (Soltandoost et al. 2025).

Method

Figure 4 presents the overall architecture of SAFE, which integrates two complementary components: 1) Semantic-enhanced multimodal alignment. We first use CLIP-CAP to generate an automatic image description and then append authenticity cues (see Appendix B) so that genuine and ma-

[1]The augmented sample (labeled as fake) is so-called pseudo fake (p-fake) sample to distinguish them from the original fake (o-fake) sample of the training data.

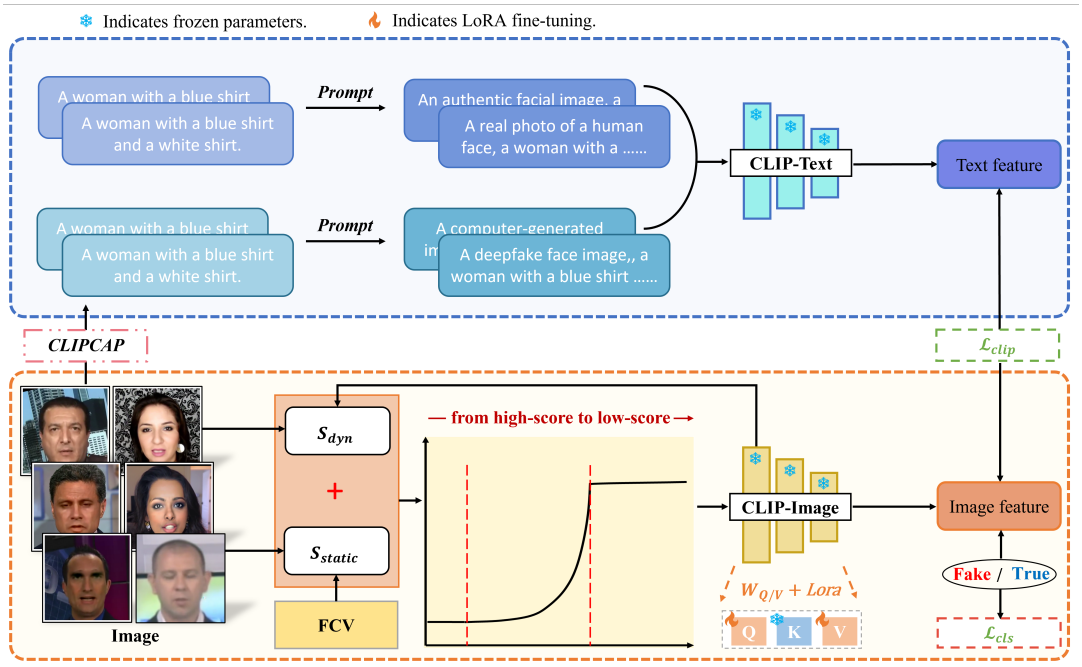

Figure 4: Overall architecture of the proposed SAFE framework.

nipulated samples become separable in the textual space. A low-rank LoRA adapter ($r = 1, \alpha = 4$) is inserted exclusively into the visual branch of CLIP. Training jointly minimizes a modality-alignment loss and a forgery-classification loss. 2) Dual-score curriculum learning. Block-level Fourier Correlation Variance (FCV) provides a static difficulty score that reflects the salience of each forgery. The classification loss from the previous epoch supplies a dynamic score. Their sum forms a composite difficulty metric. During training, a scheduler presents high-score (easy) samples first and gradually introduces low-score (hard) samples, enabling the model to progress from coarse to fine discrimination.

Semantic-enhanced multimodal alignment

For an input image x_i, we first feed the natural-language description generated by CLIP-CAP into the prompt template (Appendix B) to obtain an augmented caption $\tilde{c}$. CLIP's text encoder then encodes this caption to produce a text embedding $\mathbf{t}_i \in \mathbb{R}^d$. In parallel, the image x_i is processed by CLIP's visual encoder, yielding the corresponding image embedding $\mathbf{v}_i \in \mathbb{R}^d$.

The visual encoder uses ViT-L/14, with all original weights W_0 kept frozen. We insert LoRA adapters only into the query and value projections of every self-attention layer; keys, MLPs, LayerNorms and positional embeddings remain frozen. For a projection matrix $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, the adapted weight is

$$W_{\text{adapt}} = W_0 + \alpha B A^\top$$

where $A, B \in \mathbb{R}^{d \times r}$ with $r = 1$ and scaling factor $\alpha = 4$. During training, only A and B are updated, so the trainable parameter count is roughly 0.09M of the visual backbone, substantially lowering both training and deployment costs.

Modality-alignment loss. We adopt the symmetric InfoNCE objective from the original CLIP. For a mini-batch of B image–text pairs, we compute a similarity matrix $S \in \mathbb{R}^{B \times B}$ whose entries are:

$$S_{ij} = \frac{\mathbf{v}_i^\intercal \mathbf{t}_j}{\|\mathbf{v}_i\| \|\mathbf{t}_j\|}$$

The image–text alignment loss is defined as:

$$\mathcal{L}_{clip} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(S_{ii})}{\sum_{j=1}^{B} \exp(S_{ij})}$$

A binary linear classifier $\mathbf{W}\text{cls} \in \mathbb{R}^{2 \times d}$ is appended on top of the visual feature $\mathbf{v}i$:

To complement this objective, we append a binary linear classifier $\mathbf{W}\text{cls} \in \mathbb{R}^{2 \times d}$ to each visual embedding $\mathbf{v}i$ and obtain the class probabilities $\mathbf{p}_i = \text{softmax}(\mathbf{W}_{\text{cls}} \mathbf{v}_i)$, $y_i \in \{0, 1\}$. The classifier is trained with the cross-entropy loss $\mathcal{L}_{\text{cls}}$.

Empirically, alignment alone fails to capture fine-grained forgery artifacts, whereas stand-alone classification can disrupt cross-modal consistency. We therefore adopt a dual-branch objective:

$$\mathcal{L} = \frac{1}{\lambda} \mathcal{L}_{\text{clip}} + (1 - \frac{1}{\lambda}) \mathcal{L}_{\text{cls}}$$

Furthermore, show in next section that $\lambda = 20$ delivers the best overall performance.

Methods	Reference	#Param	Cross-dataset evaluation			Cross-manipulation evaluation						
			CDFv2	DFDCP	DFDC	UniFace	BleFace	e4s	FaceDan	FSGAN	InSwap	SimSwap
SPSL	CVPR-2021	21M	0.799	0.770	0.724	0.747	0.748	0.514	0.666	0.812	0.643	0.665
RECEE	CVPR-2022	48M	0.823	0.734	0.696	0.898	0.832	0.683	0.848	0.949	0.848	0.768
CORE	CVPRW-2022	22M	0.809	0.720	0.721	0.871	0.843	0.679	0.774	0.958	0.855	0.724
SBI	CVPR-2022	18M	0.886	0.848	0.717	0.724	0.891	0.750	0.594	0.803	0.712	0.701
UCF	ICCV-2023	47M	0.837	0.770	0.742	0.831	0.827	0.731	0.862	0.937	0.809	0.647
IID	CVPR-2023	66M	0.939	0.689	0.700	0.839	0.789	0.766	0.844	0.927	0.789	0.644
LSDA	CVPR-2024	133M	0.881	0.812	0.701	0.872	0.875	0.694	0.721	0.939	0.855	0.793
CDFA	ECCV-2024	87M	0.938	0.881	0.830	0.762	0.756	0.631	0.803	0.942	0.772	0.757
ProDet	NeurIPS-2024	96M	0.938	0.828	0.707	0.908	0.929	0.771	0.747	0.928	0.837	0.844
Effort	ICML-2025	0.19M	0.956	0.909	0.843	0.962	0.873	0.983	0.926	0.957	0.936	0.926
SAFE (Ours)	–	**0.09M**	**0.969**	**0.942**	**0.882**	**0.983**	**0.954**	**0.998**	**0.971**	**0.991**	**0.968**	**0.956**

Table 1: Video-level benchmark results for both cross-dataset and cross-manipulation evaluations. By the DeepfakeBench protocol (Yan et al. 2023b), every detector is trained on FaceForensics++ (FF++). Performance figures are taken from the original publications or, where unavailable, from the Effort study (Yan et al. 2025).

Dual-score curriculum learning

We evaluated three candidate scoring metrics (see Appendix A) and ultimately selected Fourier Correlation Variance (FCV) as the static difficulty score S_{static} because it provides the strongest generalization. The dynamic score is the classification loss from the previous forward pass $S_{dyn} = \mathcal{L}_{\text{cls}}(x, \theta_t)$. After min–max normalization, the two scores are combined into a composite measure $S_{\text{total}}(x_i) = (1 - \widehat{S}_{\text{static}}) + (1 - \widehat{S}_{\text{dyn}})$. So that samples that are easy in both senses receive higher total scores.

For S_{static}, given an input image x_i, we partition it into non-overlapping 16×16 pixel blocks $\{B_1, B_2, \ldots, B_{M\times N}\}$. For each block B_j, we compute the magnitude of its 2-D discrete Fourier transform,

$$F_j(u,v) = \mathcal{F}\{B_j\}(u,v) = \left| \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} B_j(x,y) e^{-i2\pi\left(\frac{ux}{W} + \frac{vy}{H}\right)} \right|$$

where $W = H = 16$. For every pair of blocks (B_j, B_k), we calculate the Pearson correlation coefficient between their magnitude spectra,

$$\rho_{jk} = \frac{\sum_{u,v}(F_j(u,v) - \mu_j)(F_k(u,v) - \mu_k)}{\sqrt{\sum_{u,v}(F_j(u,v) - \mu_j)^2 \sum_{u,v}(F_k(u,v) - \mu_k)^2}}$$

where μ_j and μ_k denote the mean magnitudes of F_j and F_k, respectively. The FCV of image x_i is defined as the variance of all block-pair correlations,

$$\text{FCV}(x_i) = \text{Var}(\{\rho_{jk} \mid 1 \le j < k \le MN\})$$

We set $S_{\text{static}}(x_i) = \text{FCV}(x_i)$. Samples that are easy under both the static and dynamic criteria receive higher composite scores:

$$S_{\text{total}}(x_i) = (1 - S_{\text{static}}(x_i)) + (1 - S_{\text{dyn}}(x_i))$$

Let e denote the current training epoch and E the total number of epochs. Two turning points, T_0 and T_1, partition the curriculum into three stages (Table 2):

Stages	Sampling rules
Stage I ($e \le T_0$)	All authentic samples ∪ 30% forged samples.
Stage II ($T_0 < e \le T_1$)	The sampling ratio increases with epoch: $p(e) = p_0 + (1 - p_0)\, f\left(\frac{e - T_0}{T_1 - T_0}\right)$.
Stage III ($e > T_1$)	All samples.

Table 2: Curriculum sampling schedule.

We compared several pacing functions (linear, convex concave, cosine, sigmoid, and step). The convex quadratic schedule achieved the best trade-off between cross-set AUC and convergence speed (see next section for details).

Experiment

Dataset. We adopt two widely used evaluation protocols in deepfake research: cross-dataset and cross-manipulation testing, while strictly following the benchmark splits and preprocessing pipelines of DeepfakeBench (Yan et al. 2023b) and DF40 (Yan et al. 2024b). 1) Cross-dataset evaluation. The detector is trained on FaceForensics++ (FF++) (Rossler et al. 2019) and evaluated on Celeb-DF v1 (CDF-v1) and v2 (CDF-v2) (Li et al. 2020), the DeepFake Detection Challenge dataset (DFDC) (Dolhansky et al. 2020), and the DFDC Preview set (DFDCP) (Dolhansky et al. 2019). 2) Cross-manipulation evaluation. We use DF40, whose forged videos are generated with novel manipulation techniques while remaining within the FF++ domain, thereby isolating the effect of unseen forgeries (Yan et al. 2024b).

Implementation details. We employ CLIP ViT-L/14 (Radford et al. 2021) as the backbone and retain all hyperparameter settings from DeepfakeBench to ensure fair comparison. The network is trained for 10 epochs using Adam with a learning rate of 5e-5 and a batch size of 32. In the semantic-enhanced multimodal alignment module, the loss-balancing weight is fixed at $\lambda = 20$. For the dual-score curriculum learning, the turning points are set to $T_0 = 1$ and $T_1 = 5$. Performance is reported in terms of video-level Area Under the Curve (AUC), the standard metric in prior work; frame-

Methods	Reference	CDFv1	CDFv2	DFDCP	DFDC
UCF	ICCV-2023	0.779	0.752	0.759	0.719
IID	CVPR-2023	–	0.838	0.812	–
LSDA	CVPR-2024	0.867	0.830	0.815	0.736
Forensic-Adapter	CVPR-2025	–	0.837	0.799	0.775
FreqDebias	CVPR-2025	0.875	0.836	0.824	0.741
UDD	AAAI-2025	–	0.869	0.856	0.758
SAFE (Ours)	–	**0.931**	**0.904**	**0.911**	**0.850**

Table 3: Frame-level benchmark results for cross-dataset evaluations.

Caption	CDFv2	DFDCP	InSwap	SimSwap	Avg.
ForR	0.964	0.928	0.942	0.950	0.946
CLIPCAP	0.964	0.931	0.948	0.952	0.949
Ours	**0.969**	**0.942**	**0.968**	**0.956**	**0.959**

Table 4: Impact of caption semantic granularity on multi-modal alignment (video-level AUC).

level AUC is provided as a supplementary reference.

Comparison with existing methods

We conduct comprehensive experiments at both video-level and frame-level granularity.

Video-level evaluation. We benchmark SAFE against recent state-of-the-art detectors, including SPSL (Liu et al. 2021), RECEE (Cao et al. 2022), CORE (Ni et al. 2022), SBI (Shiohara and Yamasaki 2022), UCF (Yan et al. 2023a), IID (Huang et al. 2023), LSDA (Yan et al. 2024a), and the latest ProDet (Cheng et al. 2024) and Effort (Yan et al. 2025). Table 1 reports video-level AUC on 10 datasets. SAFE achieves the best score on every dataset while fine-tuning only 0.09M parameters, which is orders of magnitude fewer than all baselines. Thanks to the curriculum schedule, forged samples are introduced gradually (up to epoch 5), cutting total training time by 23.3% relative to full fine-tuning.

Frame-level evaluation. For frame-level testing, we compare SAFE with recent state-of-the-art detectors, including UCF (Yan et al. 2023a), IID (Huang et al. 2023), LSDA (Yan et al. 2024a), Forensic Adapter (Cui et al. 2024), FreqDebias (Kashiani, Talemi, and Afghah 2025), and UDD (Fu et al. 2025). As summarized in Table 3, SAFE achieves the best AUC on all four cross-dataset benchmarks. On the challenging DFDC set, it improves upon the previous best by 12%, surpassing even most video-level detectors.

Ablation study and analysis

Caption ablation study. We examine how caption semantic granularity affects detection by comparing three settings: 1) ForR: a single token ("fake" or "real"). 2) CLIP-CAP: the unmodified natural description generated by CLIP-CAP. 3) Ours: the CLIP-CAP caption augmented with authenticity/forgery cues. Table 4 reports the results. Using only the minimal fake/real token (ForR) performs on par with the LoRA baseline, indicating that a bare class label adds little beyond the classification head. Replacing this token with the raw CLIP-CAP description yields a modest improvement,

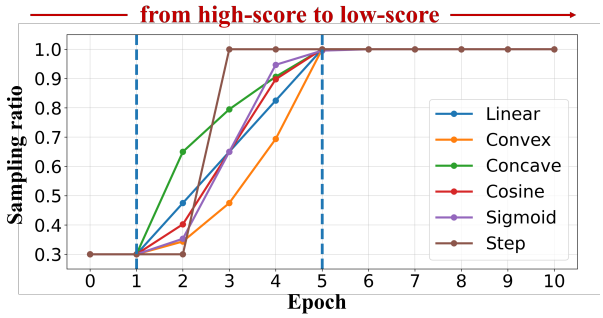

Figure 5: Comparison of curriculum pacing functions (linear, convex, concave, cosine, sigmoid, and step). Blue dashed lines denote stage boundaries $T_0 = 1$ and $T_1 = 5$. A smaller area under the curve means fewer samples are used early in training.

λ	CDFv2	DFDCP	Avg.
10	0.965	0.939	0.952
20	**0.969**	**0.942**	**0.956**
40	0.962	0.937	0.950

Table 5: Weight-balancing factor (λ) ablation (video-level AUC).

indicating that richer, generic semantics provide an additional supervisory signal. Our augmented captions achieve the highest accuracy, increasing average AUC by 1.3% over ForR. This confirms that explicitly encoding authenticity-related cues in the text space helps the model discriminate highly convincing forgeries.

Weight-balancing factor (λ) ablation study. We evaluate three values of the loss-balancing coefficient, $\lambda \in \{10, 20, 40\}$. As reported in Table 5, $\lambda = 20$ delivers the best AUC on both CDF-v2 and DFDCP, with an average score of 0.956. Increasing λ to 40 weakens the alignment term, causing the model to rely too heavily on the classification branch, whereas reducing it to 10 over-emphasizes semantic consistency and curtails the classifier's expressive capacity.

Curriculum schedule ablation study. Using two anchor points $(T_0, p_0) = (1, 0.3)$ and $(T_1, 1.0) = (5, 1.0)$, we evaluate 6 pacing functions as shown in Figure 5: linear, convex, concave, cosine, sigmoid, and step. Thus, the first epoch utilizes 30% of the training data, and by epoch 5, the entire dataset is in use. Table 6 presents AUC and training time on three cross-manipulation sets. The convex schedule achieves the highest mean AUC while reducing training time by 23.3%, thanks to its slow-then-fast progression: the model stabilizes on easier samples before tackling harder ones.

Dual-score curriculum ablation study. To evaluate the generality of our curriculum module, we integrate it into four widely used deepfake baselines: ResNet-34 (He et al. 2016), Xception (Rossler et al. 2019), EfficientNet-B4 (Tan and Le 2019), and ViT-B/16 (CLIP) (Radford et al. 2021).

Schedule	Training time	BleFace	InSwap	FaceDan	Avg.
Linear	↓ 17.5%	0.935	0.965	0.970	0.953
Convex	↓ **23.3%**	**0.954**	0.968	0.971	**0.964**
Concave	↓ 11.7%	0.945	0.964	0.971	0.958
Cosine	↓ 17.5%	0.944	0.965	**0.974**	0.959
Sigmoid	↓ 17.5%	0.938	**0.969**	**0.974**	0.956
Step	↓ 21.0%	0.928	0.964	0.967	0.948

Table 6: Performance comparison of 6 curriculum-pacing functions (video-level AUC).

Method	Config.	CDFv1	CDFv2	DFDCP	DFDC	UADFV	Avg.
ResNet34	–	0.777	0.741	0.711	0.703	0.902	0.767
	+Ours	**0.839**	**0.778**	**0.736**	**0.737**	**0.945**	**0.807**
Xception	–	0.779	0.736	0.737	0.707	0.937	0.779
	+Ours	**0.821**	**0.768**	**0.768**	**0.765**	**0.960**	**0.816**
Efficient-B4	–	0.751	0.741	0.711	0.701	0.915	0.764
	+Ours	**0.854**	**0.803**	**0.760**	**0.712**	**0.958**	**0.817**
ViT-B16	–	0.832	0.824	0.812	0.792	0.965	0.845
	+Ours	**0.897**	**0.827**	**0.829**	**0.811**	**0.979**	**0.869**

Table 7: Effectiveness of the dual-score curriculum on additional models (frame-level AUC).

r_0	CDFv2	SimSwap	Avg.
0.2	0.968	0.949	0.958
0.3	**0.969**	**0.956**	**0.963**
0.4	0.964	0.952	0.958
0.5	0.958	0.946	0.952

Table 8: Effect of the initial forged-sample ratio r_0 under the convex curriculum (video-level AUC).

Component			AUC		Avg.
LoRA	Comp. I	Comp. II	CDFv2	DFDCP	
✘	✘	✘	0.886	0.855	0.871
✔	✘	✘	0.953	0.928	0.941
✔	✔	✘	0.955	0.932	0.944
✔	✘	✔	0.962	0.931	0.947
✔	✔	✔	**0.969**	**0.942**	**0.956**

Table 9: Ablation study of SAFE's core components, including LoRA fine-tuning, Component I (semantic alignment), and Component II (dual-score curriculum), with video-level AUC results.

Table 7 reports results on 5 datasets (CDF-v1, CDF-v2, DFDCP, DFDC, and UADFV). All baselines gain on average. For example, on CDF-v1, ResNet-34 and EfficientNet-B4 improve by 8.0% and 13.7%, respectively, while Xception posts a mean increase of 4.5% across all datasets. These findings demonstrate that the dual-score curriculum learning enhances performance well beyond the SAFE detector itself.

Initial sampling ratio (r_0) ablation study. We varied the initial fraction of forged samples in Stage I from 0.2 to 0.5; Table 8 summarizes the results. With $r_0 = 0.3$, the model attains AUCs of 0.969 on CDF-v2 and 0.956 on SimSwap. A lower value ($r_0 = 0.2$) yields too few forged examples initially, resulting in under-convergence on SimSwap. Conversely, higher values ($r_0 = 0.4$ or 0.5) expose the network to complex forgeries prematurely, causing gradient oscillations and lower accuracy.

Component ablation study. Table 9 analyzes 3 key componets of SAFE: LoRA fine-tuning, semantic alignment, and dual-score curriculum learning. Starting from a baseline without any enhancements, adding LoRA alone significantly boosts performance, demonstrating that updating just 0.09M parameters is sufficient to sensitise the CLIP visual branch to deepfake artifacts. Introducing either the semantic alignment loss or the curriculum scheduler on top of LoRA yields further gains. Enabling all componets delivers the best AUC scores on both datasets, confirming their complementarity: LoRA supplies parameter flexibility, semantic alignment highlights cross-modal inconsistencies, and curriculum learning smooths the difficulty distribution.

Conclusion

In open-set settings, how can we, with minimal parameters and shorter training, stably capture "semantic inconsistency between genuine and forged content" while preserving cross-domain generalisation? SAFE unifies *semantic consistency* and *frequency-based difficulty*. On the semantic side, CLIP-CAP generates captions that are augmented with authenticity/forgery prompts; together with a LoRA adapter of only 0.09M parameters on CLIP's visual branch, the model jointly aligns text-image semantics and performs binary discrimination. On the learning side, FCV measures local spectral consistency and organises a curriculum that proceeds from high to low discrepancies—*learn salient cues first, then finer artefacts*. This simultaneously clarifies what to learn (semantic contradictions), what to learn first (high-discrepancy forgeries), and how large a model to learn with (LoRA), bringing the solution to an engineering practical scale.

Why SAFE is effective? *Signal level*: genuine and forged samples exhibit a measurable gap in high-level semantics; prompt injection amplifies this gap in CLIP space, while FCV converts "deviation from the real distribution" into an ordered difficulty gradient, ensuring that early batches carry stronger discriminative information. *Generalisation level*: in FF++, NT/DF naturally lie farther from the real distribution; starting with these high-discrepancy subsets inject more essential forgery cues into the curriculum, improving cross-domain robustness.

Future work. We will focus on an *integrated optimisation of learnable prompts and adaptive curricula*. On the semantic side, replace manual prompts with learnable soft prompts trained end-to-end with CLIP's text encoder. On the learning side, explore stronger static priors and combine them with dynamic terms to form a parameterised curriculum function. Beyond this, we aim to sustain generalisation at low cost in broader deepfake scenarios—extending from faces to speech, text-guided video editing, and multimodal synthetic content, and developing a unified measure of semantic consistency across modalities.

References

Amerini, I.; Galteri, L.; Caldelli, R.; and Del Bimbo, A. 2019. Deepfake video detection through optical flow based cnn. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 0–0.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48.

Cao, J.; Ma, C.; Yao, T.; Chen, S.; Ding, S.; and Yang, X. 2022. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4113–4122.

Chen, S.; Yao, T.; Liu, H.; Sun, X.; Ding, S.; Ji, R.; et al. 2024. Diffusionfake: Enhancing generalization in deepfake detection via guided stable diffusion. *Advances in Neural Information Processing Systems*, 37: 101474–101497.

Cheng, J.; Yan, Z.; Zhang, Y.; Luo, Y.; Wang, Z.; and Li, C. 2024. Can We Leave Deepfake Data Behind in Training Deepfake Detector? *arXiv preprint arXiv:2408.17052*.

Cui, X.; Li, Y.; Luo, A.; Zhou, J.; and Dong, J. 2024. Forensics Adapter: Adapting CLIP for Generalizable Face Forgery Detection. *arXiv preprint arXiv:2411.19715*.

Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; and Ferrer, C. C. 2020. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*.

Dolhansky, B.; Howes, R.; Pflaum, B.; Baram, N.; and Ferrer, C. C. 2019. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*.

Dong, S.; Wang, J.; Ji, R.; Liang, J.; Fan, H.; and Ge, Z. 2023. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3994–4004.

Du, Y.; Wang, Z.; Luo, Y.; Piao, C.; Yan, Z.; Li, H.; and Yuan, L. 2025. CAD: A General Multimodal Framework for Video Deepfake Detection via Cross-Modal Alignment and Distillation. *arXiv preprint arXiv:2505.15233*.

Duan, Y.; Zhu, H.; Wang, H.; Yi, L.; Nevatia, R.; and Guibas, L. J. 2020. Curriculum deepsdf. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, 51–67. Springer.

Fu, X.; Yan, Z.; Yao, T.; Chen, S.; and Li, X. 2025. Exploring Unbiased Deepfake Detection via Token-Level Shuffling and Mixing. *arXiv preprint arXiv:2501.04376*.

Guo, X.; Song, X.; Zhang, Y.; Liu, X.; and Liu, X. 2025. Rethinking Vision-Language Model in Face Forensics: Multi-Modal Interpretable Forged Face Detector. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 105–116.

Haliassos, A.; Mira, R.; Petridis, S.; and Pantic, M. 2022. Leveraging real talking faces via self-supervision for robust forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14950–14962.

Hasanaath, A. A.; Luqman, H.; Katib, R.; and Anwar, S. 2025. FSBI: Deepfake detection with frequency enhanced self-blended images. *Image and Vision Computing*, 105418.

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*, 770–778.

Huang, B.; Wang, Z.; Yang, J.; Ai, J.; Zou, Q.; Wang, Q.; and Ye, D. 2023. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4490–4499.

Kashiani, H.; Talemi, N. A.; and Afghah, F. 2025. FreqDebias: Towards Generalizable Deepfake Detection via Consistency-Driven Frequency Debiasing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8775–8785.

Kocmi, T.; and Bojar, O. 2017. Curriculum learning and minibatch bucketing in neural machine translation. *arXiv preprint arXiv:1707.09533*.

Kong, C.; Li, H.; and Wang, S. 2023. Enhancing general face forgery detection via vision transformer with low-rank adaptation. In *2023 IEEE 6th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, 102–107. IEEE.

Kong, C.; Luo, A.; Bao, P.; Yu, Y.; Li, H.; Zheng, Z.; Wang, S.; and Kot, A. C. 2024. Moe-ffd: Mixture of experts for generalized and parameter-efficient face forgery detection. *arXiv preprint arXiv:2404.08452*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; and Qiao, Y. 2023b. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Li, Y.; Chang, M.-C.; and Lyu, S. 2018. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International workshop on information forensics and security (WIFS)*, 1–7. Ieee.

Li, Y.; and Lyu, S. 2018. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*.

Li, Y.; Yang, X.; Sun, P.; Qi, H.; and Lyu, S. 2020. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3207–3216.

Lin, Y.; Song, W.; Li, B.; Li, Y.; Ni, J.; Chen, H.; and Li, Q. 2024. Fake it till you make it: Curricular dynamic forgery augmentations towards general deepfake detection. In *European Conference on Computer Vision*, 104–122. Springer.

Liu, A. 2024. Ca-moeit: Generalizable face anti-spoofing via dual cross-attention and semi-fixed mixture-of-expert. *International Journal of Computer Vision*, 132(11): 5439–5452.

Liu, A.; Ma, H.; Zheng, J.; Yuan, H.; Yu, X.; Liang, Y.; Escalera, S.; Wan, J.; and Lei, Z. 2024a. Fm-clip: Flexible modal clip for face anti-spoofing. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8228–8237.

Liu, A.; Xue, S.; Gan, J.; Wan, J.; Liang, Y.; Deng, J.; Escalera, S.; and Lei, Z. 2024b. Cfpl-fas: Class free prompt learning for generalizable face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 222–232.

Liu, H.; Li, X.; Zhou, W.; Chen, Y.; He, Y.; Xue, H.; Zhang, W.; and Yu, N. 2021. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 772–781.

Liy, C. M.; and InIctuOculi, L. 2018. Exposingaicreated fakevideosbydetectingeyeblinking. In *2018IEEEInterG national Workshop on Information Forensics and Security (WIFS). IEEE*.

Luo, Y.; Zhang, Y.; Yan, J.; and Liu, W. 2021. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16317–16326.

Mokady, R.; Hertz, A.; and Bermano, A. H. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Ni, Y.; Meng, D.; Yu, C.; Quan, C.; Ren, D.; and Zhao, Y. 2022. Core: Consistent representation learning for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12–21.

Ojha, U.; Li, Y.; and Lee, Y. J. 2023. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24480–24489.

Pan, K.; Yin, Y.; Wei, Y.; Lin, F.; Ba, Z.; Liu, Z.; Wang, Z.; Cavallaro, L.; and Ren, K. 2023. Dfil: Deepfake incremental learning by exploiting domain-invariant forgery clues. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8035–8046.

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

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

Ranjan, S.; and Hansen, J. H. 2017. Curriculum learning based approaches for noise robust speaker recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(1): 197–210.

Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1–11.

Shiohara, K.; and Yamasaki, T. 2022. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18720–18729.

Soltandoost, E.; Plesh, R.; Schuckers, S.; Peer, P.; and Štruc, V. 2025. Extracting local information from global representations for interpretable deepfake detection. In *Proceedings of the Winter Conference on Applications of Computer Vision*, 1629–1639.

Song, W.; Lin, Y.; and Li, B. 2024. Towards General Deepfake Detection with Dynamic Curriculum. *arXiv preprint arXiv:2410.11162*.

Tan, C.; Tao, R.; Liu, H.; Gu, G.; Wu, B.; Zhao, Y.; and Wei, Y. 2025. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7184–7192.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.

Thies, J.; Zollhöfer, M.; and Nießner, M. 2019. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4): 1–12.

Wang, Z.; Bao, J.; Zhou, W.; Wang, W.; and Li, H. 2023. Altfreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4129–4138.

Wu, X.; Xie, Z.; Gao, Y.; and Xiao, Y. 2020. Sstnet: Detecting manipulated faces through spatial, steganalysis and temporal features. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2952–2956. IEEE.

Yan, Z.; Luo, Y.; Lyu, S.; Liu, Q.; and Wu, B. 2024a. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8984–8994.

Yan, Z.; Wang, J.; Jin, P.; Zhang, K.-Y.; Liu, C.; Chen, S.; Yao, T.; Ding, S.; Wu, B.; and Yuan, L. 2025. Orthogonal Subspace Decomposition for Generalizable AI-Generated Image Detection. In *Forty-second International Conference on Machine Learning*.

Yan, Z.; Yao, T.; Chen, S.; Zhao, Y.; Fu, X.; Zhu, J.; Luo, D.; Wang, C.; Ding, S.; Wu, Y.; et al. 2024b. Df40: Toward next-generation deepfake detection. *arXiv preprint arXiv:2406.13495*.

Yan, Z.; Zhang, Y.; Fan, Y.; and Wu, B. 2023a. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22412–22423.

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

Yang, X.; Li, Y.; and Lyu, S. 2019. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 8261–8265. IEEE.