Deepfake Detection that Generalizes Across Benchmarks (Supplementary Material)

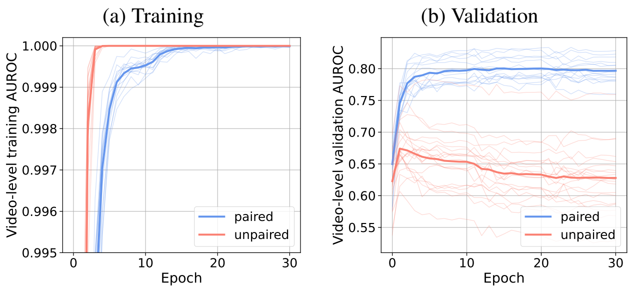


Figure S1. Video-level AUROC for (a) Training and (b) Validation, averaged over 20 randomly sampled paired and unpaired datasets from the FF++ training set. The image encoder is CLIP ViT-L/14.

S1. Quantitative Results for Paired vs. Unpaired Training

This section provides detailed quantitative results to support the analysis in Section 4.4 of the main paper, which argues for the importance of training on paired datasets. Although Fig. 3 in the main paper illustrates the training dynamics, Tab. S1 presents the final cross-dataset generalization performance. The results are averaged over 20 different paired and 20 unpaired training sessions.

The experiment provides strong empirical evidence for the hypothesis. The model trained on the Paired dataset consistently outperforms the one trained on the Unpaired dataset. It achieves a mean AUROC of 90.0%, a notable improvement by 4.7 pp. This performance gain is consistent across all benchmarks.

We verify that this finding generalizes to other backbones such as CLIP ViT-L/14 [8], presented in Fig. S1, where the training dynamic resembles Fig. 3 from the main paper.

In addition, we verify that this finding generalizes to other training datasets such as CDFv2 [6], shown in Fig. S3 and FAVC [5], shown in Fig. S2. We observe the same training dynamics but with faster overfitting in both cases, suggesting that on our validation set, models trained on CDFv2 or FAVC training sets have lower generalization performance; see Fig. S3 (b) and Fig. S2 (b).

S2. Detailed Ablation of Parameter-Efficient Fine-Tuning (PEFT) Methods

We provide a more comprehensive evaluation of parameter-efficient fine-tuning (PEFT) strategies in Tab. S2, which were discussed in Section 4.3 of the main paper. This table evaluates and compares the cross-dataset generalization performance of Full Fine-Tuning (FFT), a baseline (where only a linear classifier is trained), and three distinct PEFT methods: BitFit [1], LoRA [4], and LN-tuning [7]. Each PEFT method is evaluated in isolation as well as in combination with the other components of the proposed method: L2 normalization (+L2), uniformity and alignment losses (+UA). Every training run has the same training set and differs only in the ablated component.

As shown in the table, our findings reinforce the conclusions from the main text:

1. Full Fine-Tuning (FFT) results in poor generalization, achieving the lowest mean AUROC (56.8%) across all configurations. This supports the observation that FFT leads to rapid overfitting on the training data, as reported by, e.g., [10].

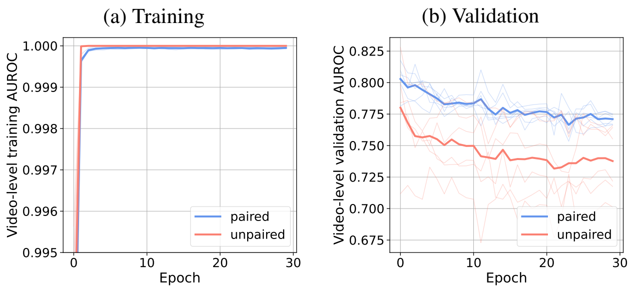


Figure S2. Video-level AUROC for (a) Training and (b) Validation, averaged over 6 randomly sampled paired and 6 unpaired datasets from the FAVC [5] dataset. The image encoder is CLIP ViT-L/14.

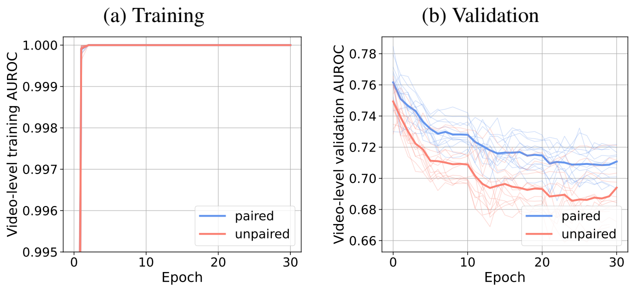


Figure S3. Video-level AUROC for (a) Training and (b) Validation, averaged over 15 randomly sampled paired and 15 unpaired datasets from the CDFv2 [6] training set. The image encoder is CLIP ViT-L/14.

Table S1. Cross-dataset video-level AUROC (%) results for models trained on paired and unpaired datasets.

Training dataset	UADFV	DFD	DFDC	FSh	CDFv2	FFIW	KoDF	FAVC	DFDM	PGF	IDF	DSv1.1	DSv2	CDFv3	Mean
Unpaired	97.5±0.3	92.4±0.7	79.7±0.8	85.3±2.4	81.0±2.8	87.4±1.7	78.5±4.0	89.8±2.7	94.6±2.9	83.9±0.6	95.7±0.6	78.4±0.0	70.2±2.4	80.2±1.5	85.3
Paired	**98.1±0.3**	**95.3±0.8**	**80.4±0.2**	**86.3±1.3**	**91.5±1.3**	**90.9±1.0**	**84.8±1.6**	**95.8±0.3**	**97.8±1.0**	**91.5±2.7**	**97.0±1.2**	**87.7±1.3**	**74.6±0.2**	**88.3±2.1**	**90.0**

2. All three PEFT methods substantially outperform both the FFT and the baseline, demonstrating the effectiveness of parameter-efficient adaptation for this task.
3. Although all PEFT methods initially perform well, their synergy with the proposed components varies. We observe that LN-Tuning, together with L2 normalization and UA losses, achieves the highest mean AUROC (92.1%); see the last row.

This detailed comparison confirms that LN-tuning is the most effective PEFT for the problem. Striking the right balance between expressiveness and regularization, enabling synergistic improvements when combined with L2 normalization, metric learning losses, achieves the state-of-the-art generalization.

S3. Detailed performance metrics: AUROC, AP and EER, TPR@FPR=1%,5%

We include additional performance metrics such as average precision (AP) in Tab. S5, equal error rate (EER) in Tab. S4, TPR@FPR=5% in Tab. S6, and TPR@FPR=1% in Tab. S7 for GenD, ForAda [2] and Effort [10] computed on all 14 cross-dataset test benchmarks presented in the paper. For the GenD, we show the mean and standard deviation calculated in five different training seeds. Extending the Tab. 4 of the main paper, we include standard deviations to Tab. S3.

S4. Ablation study of L2 normalizaion and uniformity-alignment without LN-tuning

LN-tuning [3, 7, 9] has the most noticeable influence on the performance, allowing for the reshaping of the feature space of the classification token. This makes it more suitable for solving the deepfake detection problem using a linear classification layer. Considering that uniformity and alignment (UA) loss operates in the feature space of L2-normalized classification token features, disabling LN tuning blocks all gradient propagation from the classifier backward, rendering UA useless. This explains why rows 2 and 3 of the Tab. S8 lead to the same results. However, only by making the feature space hyperspherical, performance can be improved substantially for some datasets, such as IDF (15.3 pp), DFDM (6.9 pp), leading to an improvement in 8 of 14 datasets.

S5. Visual examples for image degradations

We include visualizations of different levels of image degradation, corresponding to Fig. 5 of the main paper: Resizing (Fig. S4), Gaussian blurring (Fig. S5), and JPEG compression (Fig. S6).

S6. Visual examples for common failure cases

We visually present common failure cases in Fig. S7. We ordered approximately 60K testing videos from the most to the least misclassified video according to the softmax score of the output. We manually investigated the top 300 videos with the highest error. We observe a few distinct modes of failures. Photos a-h show that the network misclassified black people. Photos f-m suggest that eyeglasses may cause misclassifications. Photos n-r again signify the ethnic bias for asian people. Photos p-u can be failures due to low quality, image intensity, contrast, or monochrome processing. An interesting case is v-y, showing that the ground-truth class represents a real category, but the photos have been visibly altered through the addition of cartoon-like effects.

Table S2. Experimental results comparing various PEFT strategies for CLIP ViT-L/14 backbone. Each row represents a different model configuration across all test datasets. The best video-level AUROC results are shown in bold. The last row is the proposed GenD.

Method	2019	2019	2019	2020	2020	2021	2021	2021	2022	2024	2024	2024	2025	**Mean**
	UADFV	**DFD**	**DFDC**	**FSh**	**CDFv2**	**FFIW**	**KoDF**	**FAVC**	**DFDM**	**PGF**	**IDF**	**DSv1.1**	**DSv2**	
FFT	62.4	49.5	56.0	58.8	65.5	53.4	52.8	60.8	59.7	57.5	68.3	56.7	50.8	57.9
Baseline	95.8	86.7	74.0	77.9	78.8	84.6	82.2	80.8	80.7	61.1	68.2	61.6	56.6	76.1
BitFit	99.6	96.9	84.6	86.6	95.3	91.2	88.9	95.4	99.6	86.5	94.2	86.6	76.0	90.9
BitFit+L2	99.4	96.1	82.6	85.3	92.7	90.3	85.0	96.2	98.8	91.8	**98.1**	88.9	76.8	90.9
BitFit+L2+UA	99.8	96.9	85.9	85.2	**95.4**	89.4	87.6	96.1	**99.9**	**92.1**	97.0	90.2	76.3	91.7
LoRA	98.8	97.7	84.4	89.7	94.1	**93.7**	88.2	**96.3**	99.4	90.7	96.6	89.7	75.9	91.9
LoRA+L2	98.6	95.5	81.9	86.2	89.3	90.8	87.8	95.5	99.4	90.6	95.3	89.4	75.9	90.5
LoRA+L2+UA	99.5	96.1	86.2	**92.0**	92.3	92.7	**90.8**	94.8	99.5	88.1	96.4	86.7	73.0	91.4
LN	99.3	96.6	84.0	86.7	93.1	91.6	87.4	94.8	99.1	87.8	93.3	89.1	77.6	90.8
LN+L2	**99.8**	97.3	**86.6**	88.9	94.5	91.6	84.5	96.0	99.4	89.6	95.2	89.8	**79.8**	91.8
LN+L2+UA	99.6	**97.8**	86.5	87.3	94.3	90.7	87.0	96.1	99.6	90.7	98.0	**90.6**	78.6	**92.1**

Table S3. Cross-dataset video-level AUROC (%) for reproduced methods. The highest score in each column is in bold. Results for GenD are the averages over five training seeds.

Method	2019	2019	2019	2020	2020	2021	2021	2021	2022	2024	2024	2024	2025	2025	**Mean**
	UADFV	**DFD**	**DFDC**	**FSh**	**CDFv2**	**FFIW**	**KoDF**	**FAVC**	**DFDM**	**PGF**	**IDF**	**DSv1.1**	**DSv2**	**CDFv3**	
ForAda [2]	**99.4**	**97.2**	**87.3**	82.0	**95.7**	90.6	88.2	93.1	97.1	86.6	90.8	81.8	72.8	75.6	88.4
Effort [10]	97.4	95.2	85.4	**91.2**	93.2	92.5	88.1	92.4	98.2	84.9	96.0	82.1	64.4	78.7	88.5
GenD (CLIP)	99.2±0.1	96.4±0.5	86.4±0.4	86.6±0.5	94.6±0.9	91.5±1.4	84.9±0.6	96.0±0.8	99.6±0.1	89.6±0.6	97.8±0.5	**90.1±0.8**	77.7±0.7	85.9±0.7	91.2
GenD (PE)	97.7±0.2	96.8±0.5	82.2±0.5	87.6±1.1	95.0±0.8	**93.7±0.5**	85.1±1.2	97.3±0.6	98.3±0.5	92.3±0.5	97.9±0.5	87.8±1.6	78.6±1.6	**89.5±0.6**	91.4
GenD (DINO)	98.6±0.1	96.2±0.4	85.6±0.5	88.8±1.3	92.5±0.9	92.9±1.2	**89.7±0.7**	**98.4±0.5**	**99.8±0.1**	**92.4±0.4**	**98.2±0.5**	86.9±0.8	**79.4±0.6**	83.5±1.0	**91.6**

Table S4. Cross-dataset video-level EER (%) for reproduced methods. The lowest score in each column is in bold. Results for GenD are the averages over five training seeds.

Method	2019	2019	2019	2020	2020	2021	2021	2021	2022	2024	2024	2024	2025	2025	**Mean**
	UADFV	**DFD**	**DFDC**	**FSh**	**CDFv2**	**FFIW**	**KoDF**	**FAVC**	**DFDM**	**PGF**	**IDF**	**DSv1.1**	**DSv2**	**CDFv3**	
ForAda [2]	4.1	**9.0**	**20.7**	24.3	**11.2**	17.1	20.0	15.2	8.5	22.0	18.4	25.6	30.5	32.0	18.5
Effort [10]	6.1	11.3	22.9	**16.4**	14.7	16.1	18.4	16.2	7.1	23.2	11.0	24.6	40.0	29.8	18.4
GenD (CLIP)	**3.7±1.7**	10.4±0.8	21.8±0.3	21.7±0.8	13.2±1.4	16.8±1.8	22.7±0.8	11.2±1.0	2.7±0.6	17.9±1.0	7.7±1.2	**16.1±0.8**	28.7±0.8	22.1±0.7	15.5
GenD (PE)	6.5±0.9	9.3±0.8	24.8±0.7	20.7±1.0	12.0±1.3	**13.1±0.7**	22.6±1.2	9.2±1.4	7.2±1.2	**14.5±0.7**	7.9±1.0	20.5±2.8	29.8±1.3	**17.4±0.7**	15.4
GenD (DINO)	4.9±1.1	10.7±0.6	22.6±0.5	19.1±1.1	14.4±1.2	14.6±1.4	**17.4±0.7**	**6.2±0.9**	**1.8±0.4**	15.6±0.7	**6.8±1.2**	20.5±1.1	**28.2±1.0**	22.9±1.2	**14.7**

Table S5. Cross-dataset video-level AP (%) for reproduced methods. The highest score in each column is in bold. Results for GenD are the averages over five training seeds.

Method	2019	2019	2019	2020	2020	2021	2021	2021	2022	2024	2024	2024	2025	2025	**Mean**
	UADFV	**DFD**	**DFDC**	**FSh**	**CDFv2**	**FFIW**	**KoDF**	**FAVC**	**DFDM**	**PGF**	**IDF**	**DSv1.1**	**DSv2**	**CDFv3**	
ForAda [2]	**99.4**	**90.0**	**86.9**	81.7	**95.3**	90.3	79.5	66.0	95.4	64.7	85.7	80.4	70.5	53.6	81.4
Effort [10]	97.2	82.0	84.6	**90.4**	91.9	92.6	80.6	61.7	97.0	66.3	93.1	80.4	63.1	54.2	81.1
GenD (CLIP)	99.3±0.1	88.8±1.5	86.1±0.5	86.2±0.4	94.3±0.9	91.5±1.4	74.6±0.5	76.7±3.8	99.5±0.2	72.7±0.8	96.2±0.7	**89.2±0.7**	76.0±0.8	58.5±0.8	85.0
GenD (PE)	97.8±0.2	89.1±1.5	81.5±0.4	87.8±0.9	94.7±0.8	**93.5±0.5**	75.4±1.7	81.5±1.8	97.5±0.7	79.7±1.4	96.3±0.7	87.6±1.6	78.8±1.6	**59.8±0.6**	85.8
GenD (DINO)	98.7±0.1	88.5±0.9	85.1±0.5	88.9±1.2	92.0±0.9	92.7±1.2	**82.5±1.1**	**87.8±2.7**	**99.7±0.1**	**80.6±1.5**	**96.7±1.3**	86.5±0.8	**78.9±0.6**	55.8±0.9	**86.8**

Table S6. Cross-dataset video-level TPR@FPR=5% for reproduced methods. The highest score in each column is in bold. Results for GenD are the averages over five training seeds.

Method	2019	2019	2019	2020	2020	2021	2021	2021	2022	2024	2024	2024	2025	2025	**Mean**
	UADFV	**DFD**	**DFDC**	**FSh**	**CDFv2**	**FFIW**	**KoDF**	**FAVC**	**DFDM**	**PGF**	**IDF**	**DSv1.1**	**DSv2**	**CDFv3**	
ForAda [2]	**95.9**	87.3	55.7	46.4	**76.2**	60.1	72.2	73.8	86.7	51.5	61.5	45.5	19.4	33.9	61.9
Effort [10]	93.9	84.5	**57.0**	**74.3**	73.8	**72.3**	71.4	73.8	91.7	51.3	81.5	45.8	16.9	42.0	66.4
GenD (CLIP)	94.9±1.4	80.9±1.1	50.6±0.8	53.6±4.0	68.8±0.8	62.6±0.8	69.0±1.0	81.5±1.8	98.6±0.4	49.9±1.0	88.5±2.2	**59.5±1.5**	31.5±0.6	55.0±0.3	67.5
GenD (PE)	91.4±2.2	**87.5±1.9**	32.7±1.7	51.9±5.2	72.1±4.7	67.7±2.6	67.4±1.6	86.7±2.6	90.2±3.2	60.2±2.6	89.2±2.3	41.3±4.6	**42.7±4.4**	**64.3±2.0**	67.5
GenD (DINO)	94.7±2.7	84.0±2.2	54.2±4.0	57.4±9.1	57.6±5.2	68.8±5.9	**73.4±0.9**	**92.1±2.5**	**99.3±0.3**	**61.9±2.9**	**91.5±3.0**	46.3±5.0	41.1±4.9	48.3±1.2	**69.3**

Table S7. Cross-dataset video-level TPR@FPR=1% for reproduced methods. The highest score in each column is in bold. Results for GenD are the averages over five training seeds.

Method	2019	2019	2019	2020	2020	2021	2021	2021	2022	2024	2024	2024	2025	2025	**Mean**
	UADFV	**DFD**	**DFDC**	**FSh**	**CDFv2**	**FFIW**	**KoDF**	**FAVC**	**DFDM**	**PGF**	**IDF**	**DSv1.1**	**DSv2**	**CDFv3**	
ForAda [2]	**93.9**	**77.1**	31.4	28.6	**55.3**	31.4	**67.2**	59.7	75.8	29.5	45.9	30.9	6.5	18.2	46.5
Effort [10]	89.8	75.8	**44.7**	15.7	47.1	**55.8**	54.6	61.5	79.7	20.5	64.8	**34.7**	3.8	27.9	48.3
GenD (CLIP)	92.9±1.4	72.3±2.0	30.4±0.1	**34.6±5.6**	37.9±10.4	38.0±2.6	60.7±0.7	66.9±0.9	92.3±0.2	19.2±4.4	78.8±3.0	28.4±0.5	10.9±4.0	23.8±0.7	49.1
GenD (PE)	86.1±5.8	74.6±4.4	17.6±2.3	23.9±5.0	33.2±7.8	29.9±5.8	62.7±1.7	72.0±5.6	78.3±2.5	26.8±2.5	79.4±2.9	16.0±3.8	**29.6±3.1**	**30.4±3.9**	47.2
GenD (DINO)	82.0±6.0	69.8±4.3	34.8±3.2	31.4±7.9	21.8±5.8	41.9±7.6	66.1±1.2	**76.2±7.2**	**96.3±1.6**	**33.6±4.3**	**82.0±7.8**	25.5±2.7	14.6±5.3	16.4±5.4	**49.5**

Table S8. Ablation study of L2 normalization and uniformity-alignment (UA) without LN-tuning. LP is the linear probing.

Method	2019	2019	2019	2020	2020	2021	2021	2021	2022	2024	2024	2024	2025	2025	**Mean**
	UADFV	**DFD**	**DFDC**	**FSh**	**CDFv2**	**FFIW**	**KoDF**	**FAVC**	**DFDM**	**PGF**	**IDF**	**DSv1.1**	**DSv2**	**CDFv3**	
CLIP+LP	94.6±0.7	89.2±1.2	**75.3±0.3**	**77.6±0.9**	74.6±1.2	80.7±1.9	**81.8±1.3**	**83.0±1.3**	77.8±1.6	62.7±0.4	68.7±3.9	**64.5±1.0**	**57.8±1.0**	75.6±1.3	76.0
CLIP+LP+L2	**97.3±1.3**	**89.2±0.7**	73.7±1.9	77.4±2.3	**76.5±0.7**	**84.2±0.6**	77.1±3.0	80.6±1.5	**84.7±2.2**	**68.1±0.5**	**84.0±5.4**	62.3±3.2	57.1±0.9	**77.2±0.7**	**77.8**
CLIP+LP+L2+UA	**97.3±1.3**	**89.2±0.7**	73.7±1.9	77.4±2.3	**76.5±0.7**	**84.2±0.6**	77.1±3.0	80.6±1.5	**84.7±2.2**	**68.1±0.5**	**84.0±5.4**	62.3±3.2	57.1±0.9	**77.2±0.7**	**77.8**

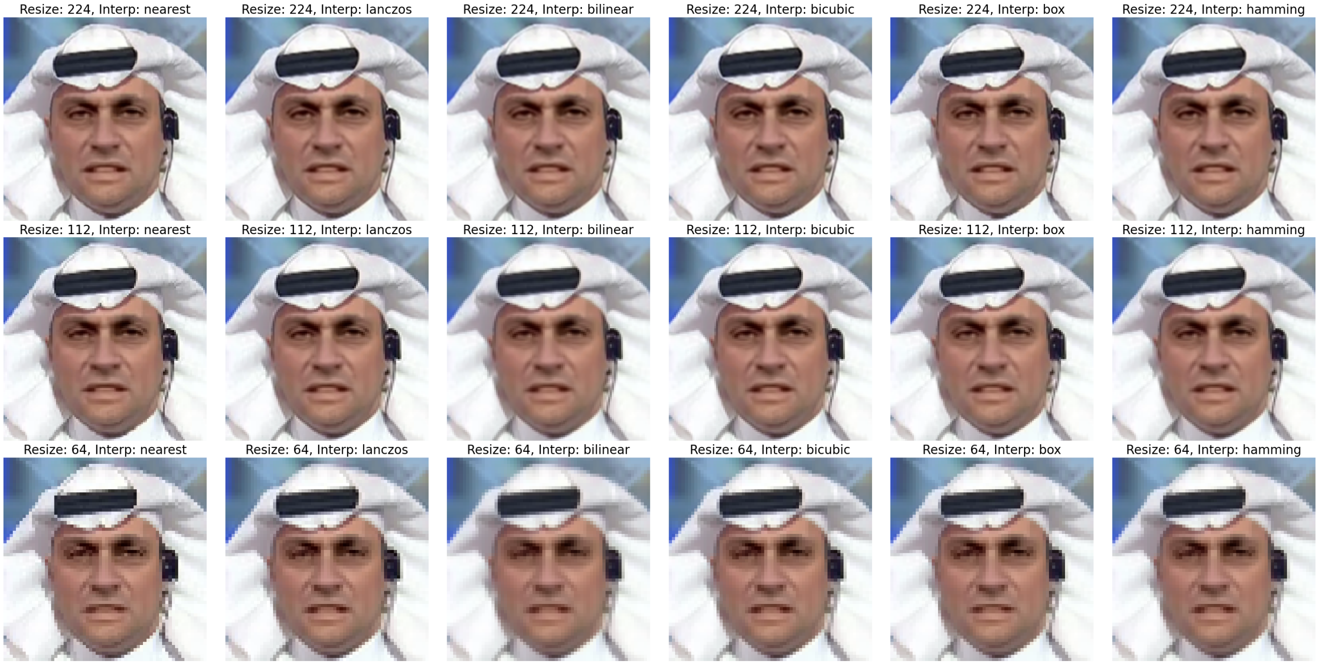

Resize: 224, Interp: nearest
Resize: 224, Interp: lanczos
Resize: 224, Interp: bilinear
Resize: 224, Interp: bicubic
Resize: 224, Interp: box
Resize: 224, Interp: hamming
Resize: 112, Interp: nearest
Resize: 112, Interp: lanczos
Resize: 112, Interp: bilinear
Resize: 112, Interp: bicubic
Resize: 112, Interp: box
Resize: 112, Interp: hamming
Resize: 64, Interp: nearest
Resize: 64, Interp: lanczos
Resize: 64, Interp: bilinear
Resize: 64, Interp: bicubic
Resize: 64, Interp: box
Resize: 64, Interp: hamming

Figure S4. Visual examples for various resizing levels and interpolations used in robustness to image degradation experiments.

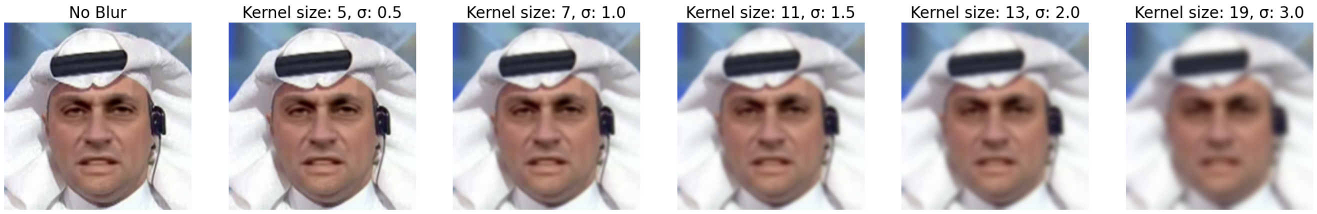


Figure S5. Visual examples for different kernel sizes and sigmas for Gaussian blurring in robustness to image degradation experiments.

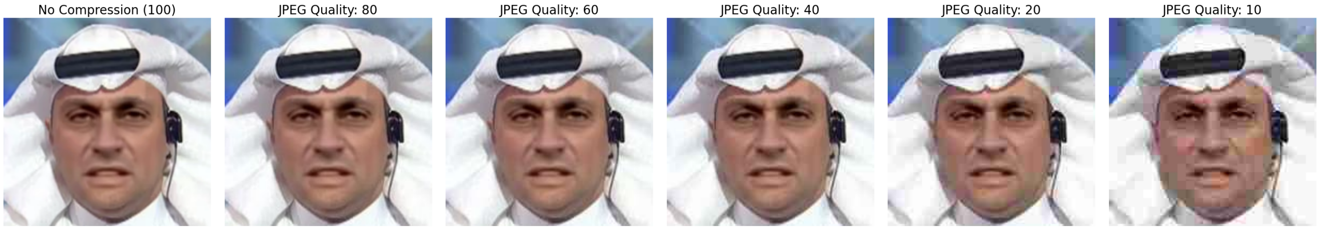
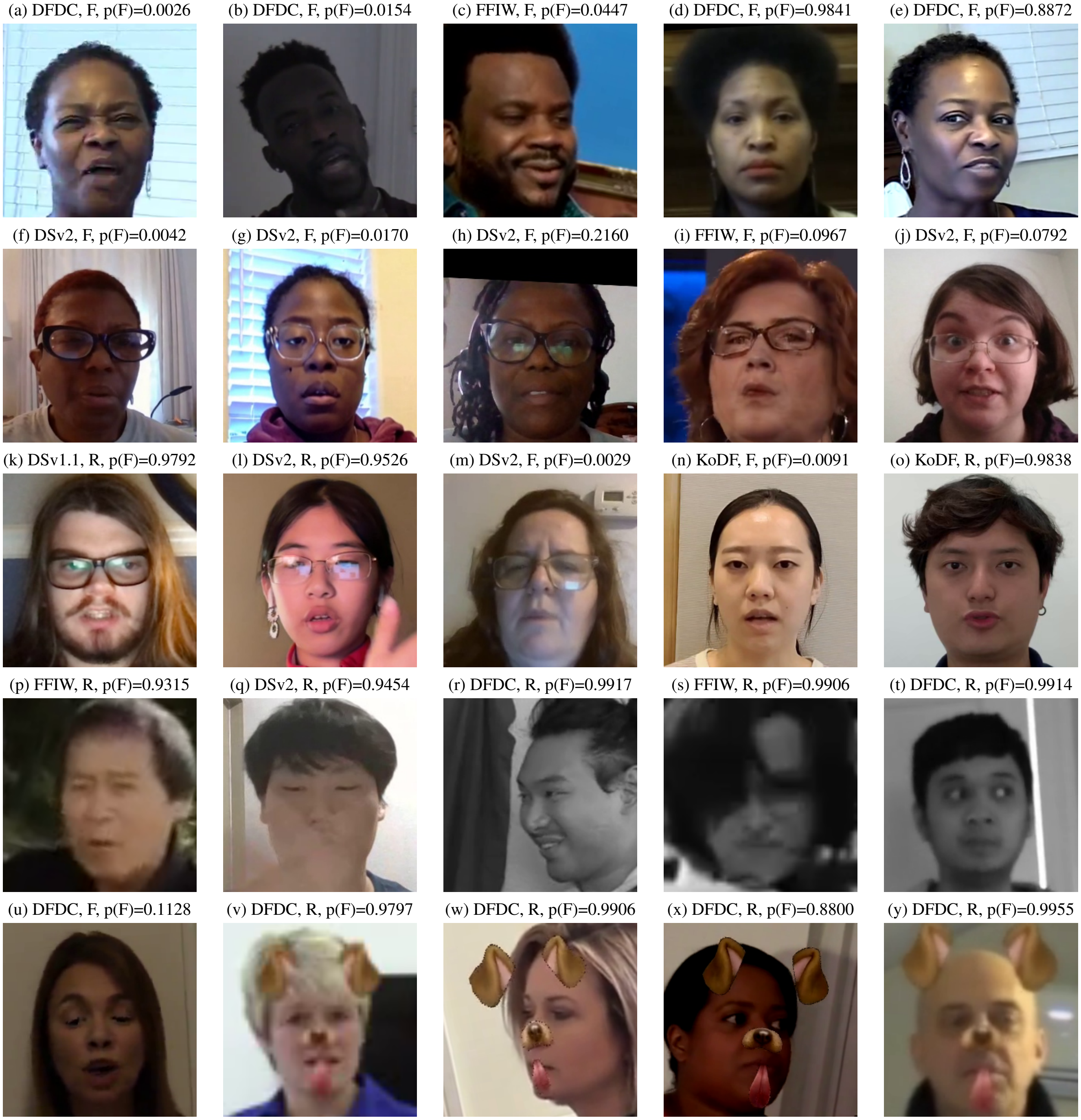


Figure S6. Visual examples for various JPEG compression levels in robustness to image degradation experiments.



(a) DFDC, F, p(F)=0.0026

(b) DFDC, F, p(F)=0.0154

(c) FFIW, F, p(F)=0.0447

(d) DFDC, F, p(F)=0.9841

(e) DFDC, F, p(F)=0.8872

(f) DSv2, F, p(F)=0.0042

(g) DSv2, F, p(F)=0.0170

(h) DSv2, F, p(F)=0.2160

(i) FFIW, F, p(F)=0.0967

(j) DSv2, F, p(F)=0.0792

(k) DSv1.1, R, p(F)=0.9792

(l) DSv2, R, p(F)=0.9526

(m) DSv2, F, p(F)=0.0029

(n) KoDF, F, p(F)=0.0091

(o) KoDF, R, p(F)=0.9838

(p) FFIW, R, p(F)=0.9315

(q) DSv2, R, p(F)=0.9454

(r) DFDC, R, p(F)=0.9917

(s) FFIW, R, p(F)=0.9906

(t) DFDC, R, p(F)=0.9914

(u) DFDC, F, p(F)=0.1128

(v) DFDC, R, p(F)=0.9797

(w) DFDC, R, p(F)=0.9906

(x) DFDC, R, p(F)=0.8800

(y) DFDC, R, p(F)=0.9955

Figure S7. Examples of common failure cases. The first word denotes the dataset. The second word is the ground truth class, R for real and F for fake. p(F) means the predicted probability of a frame being of a fake class.

References

[1] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Dublin, Ireland, 2022. Association for Computational Linguistics. 1

[2] Xinjie Cui, Yuezun Li, Ao Luo, Jiaran Zhou, and Junyu Dong. Forensics adapter: Adapting CLIP for generalizable face forgery detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19207–19217, 2025. 2, 4

[3] Angeliki Giannou, Shashank Rajput, and Dimitris Papailiopoulos. The expressive power of tuning only the normalization layers. *arXiv preprint arXiv:2302.07937*, 2023. 2

[4] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 1

[5] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*, 2021. 1

[6] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3207–3216, 2020. 1

[7] Wang Qi, Yu-Ping Ruan, Yuan Zuo, and Taihao Li. Parameter-efficient tuning on layer normalization for pretrained language models. *arXiv preprint arXiv:2211.08682*, 2022. 1, 2

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PmLR, 2021. 1

[9] Taha ValizadehAslani and Hualou Liang. Layernorm: A key component in parameter-efficient fine-tuning. *arXiv preprint arXiv:2403.20284*, 2024. 2

[10] Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *Proceedings of the International Conference on Machine Learning*, 2025. 1, 2, 4