DFD-HR: Generalizable Deepfake Detection via Hierarchical Routing Learning

Supplementary Material

A. Full Experimental Results

Comparison of different PEFT methods. We compare our DFD-HR with other popular parameter-efficient finetuning (PEFT) methods (Tab. A). The first observation is that Adapter [5] and LoRA [6] are better among PEFT variants. Then, for *LoRA Variants*, **a well-designed initialization** can improve the performance of traditional LoRA, *e.g.*, OLoRA [1] utilizes QR decomposition to initialize LoRA weights, which brings +1.2% AUC on CDF-v2. In addition, **a structure innovation** for LoRA can also improve performance, *e.g.*, rsLoRA scales each adapter during every forward pass by another designed scalar which stabilizes the adapters. Moreover, our proposed DFD-HR outperforms all other PEFT methods on the DFD benchmark, including ForAda [4], Effort [21] and MoE-FFD [10].

Additional ablation studies on DFD-HR. In Tab. B, we first conduct a careful ablation on our proposed HR module. The results show that **hierarchical routing is necessary for both split and original images**, which helps capture different semantic abstractions at different depths and concentrate on varied forgery cues while suppressing spurious tokens. Furthermore, **shared routers at both the layer and token levels enhance forgery learning** by facilitating information flow between global and local representations, *e.g.*, token judges guided by Spearman rank loss can also improve token selection of the original image. Other detailed ablations of "Token Selection" referred in the manuscript are provided in Tab. C. [CLS] and patch tokens can be treated equally for selection. Then, selected tokens should be post reweighted and token judges should differ across layers.

Additional analysis of Early Layer Pruning. Following Fig.5 (a) of the manuscript, we provide the pruning distribution of CDF-v2 and the other four forgery methods (Fig. A). Note that, to assess differences between instances, we compute the pruning distribution of original images. Conclusions are the same as the manuscript, indicating that our Early Layer Pruning captures distinct semantic abstractions which are asymmetric needed for real and fake samples (Fig.2 (b) in the manuscript). Moreover, we provide the visualization for fake images in CDF-v2 under Early Layer Pruning in Fig. B. The results show that early-pruning samples exhibit pronounced eye-region ghosting and inconsistencies, whereas forgery artifacts of late-pruning ones are not readily visible, validating that our method adaptively allocates inference depths according to forgery difficulty.

Additional analysis of Token-level Expert Routing. Following Fig.5 (b) in the manuscript, we present the complete weight distribution for our Expert Routing (Tab. D). Since token selection begins after layer 20, we extract the gating weights from layer 21 of the original images. The results show that Expert #1 and Expert #4 dominate the discrimination between real and fake samples. For different forgery methods, Expert #4 captures varied forgery cues and the other three experts adjust their weights to avoid overfitting

Table A. **Comparison of different parameter-efficient finetuning methods on video-level AUC (%).** Our settings are in blue.

Methods	CDF-v2	FaceShifter	Avg.
Baseline	91.8 +0.0	88.8 +0.0	90.3 +0.0
PEFT Variants:			
LNTuning [24]	85.3 -6.5	90.0 +1.2	87.7 -2.6
BiasTuning [2]	89.4 -2.4	89.3 +0.5	89.4 -0.9
VPT-deep [8]	88.0 -3.8	89.4 +0.6	88.7 -1.6
SideTuning [27]	75.8 -16.0	75.8 -13.0	75.8 -14.5
LinearProb	76.8 -15.0	81.1 -7.7	79.0 -11.3
Partial-4 [25]	88.4 -3.4	87.3 -1.5	87.9 -2.4
MLP-3	72.2 -19.6	73.8 -15.0	73.0 -17.3
Adapter [5]	92.8 +1.0	**90.2 +1.4**	**91.5 +1.2**
ForAda [4]	**95.7 +3.9**	82.0 -6.8	88.9 -1.4
LoRA Variants:			
LoRA [6]	91.4 -0.4	90.1 +1.3	90.8 +0.5
Gaussian LoRA	92.1 +0.3	**90.2 +1.4**	91.2 +0.9
PiSSA [17]	92.2 +0.4	**90.2 +1.4**	91.2 +0.9
OLoRA [1]	92.6 +0.8	90.0 +1.2	91.3 +1.0
rsLoRA [9]	92.6 +0.8	90.1 +1.3	91.4 +1.1
DoRA [14]	88.6 -3.2	88.8 +0.0	88.7 -1.6
LoHa [7]	90.5 -1.3	89.9 +1.1	90.2 -0.1
LoKr [23]	91.7 -0.1	**90.2 +1.4**	91.0 +0.7
AdaLora [28]	91.6 -0.2	90.1 +1.3	90.9 +0.6
IA3 [13]	80.9 +0.0	88.3 +0.0	84.6 +0.0
OFT [18]	91.1 -0.7	85.5 -3.3	88.3 -2.0
BOFT [15]	89.6 -2.2	87.2 -1.6	88.4 -1.9
Effort [21]	**95.6 +3.8**	87.7 -1.1	**91.7 +1.4**
Experts Variants:			
MoE-FFD [10]	91.3 -0.5	-	-
DFD-HR (Ours)	**96.0 +4.2**	**91.2 +2.4**	**93.6 +3.3**

Table B. **Ablation on HR module.** Our settings are in blue.

Settings	CDF-v2	FaceShifter	Avg.
Insertion position of HR:			
Split	95.0 -1.0	90.4 -0.8	92.7 -0.9
Split+Origin	**96.0 -0.0**	**91.2 -0.0**	**93.6 -0.0**
Weights between branches:			
w.o. shared	93.4 -2.6	90.2 -1.0	91.8 -1.8
w.i. shared	**96.0 -0.0**	**91.2 -0.0**	**93.6 -0.0**

case	CDF.	Fsh.	Avg.
Patch	95.7	91.1	93.4
CLS+Patch	**96.0**	**91.2**	**93.6**

(a) **Token Selection position**. No need to retain [CLS] token.

case	CDF.	Fsh.	Avg.
Pre-weight	95.4	91.0	93.2
Post-weight	**96.0**	**91.2**	**93.6**

(b) **Reweighting position**. Reweighting the selected tokens after the whole block is better.

case	CDF.	Fsh.	Avg.
w.i. shared	93.7	90.2	92.0
w.o. shared	**96.0**	**91.2**	**93.6**

(c) **Token judge between layers**. Token judges should also differ across layers.

Table C. **DFD-HR detailed ablation experiments** with CLIP ViT-L/14 trained on FF++_c23 [19]. We report video-level AUC (%) on CDF-v2 and Faceshifter. If not specified, default settings are marked in blue .

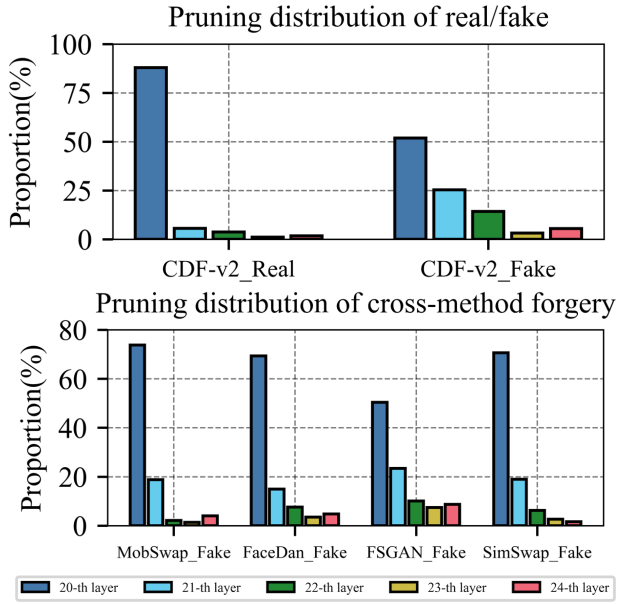

Figure A. **Additional analysis of Early Layer Pruning.** Figure indicates that real samples are predominantly terminated at shallow layers while fake ones tend to propagate to deeper layers and differ across methods for varied forgery learning.

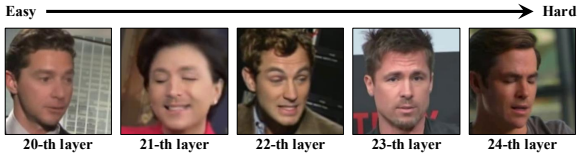

Figure B. **Visualization of samples under Early Layer Pruning.** Original fake images in CDF-v2 [12] skip at different early layers, illustrating a progression from simpler to more challenging cases.

specific forgery artifacts, thus proving the effectiveness.

Computing Cost Estimation. We consider the computation of multi-head attention (MHA) and feed-forward network (FFN) module in the FLOPs estimation. For one transformer layer, assume n is the token number, d is the hidden state size, m is the intermediate size of FFN, the total FLOPs can be estimated by $4nd^2 + 2n^2d + 2ndm$. Assume that DFD-HR selects tokens from n to $\hat{n} = (1 - R\%) \cdot n$ after layer K and one sample early prunes at layer T. The theoretical FLOPs of original Transformer blocks is $K \times (4nd^2 + 2n^2d + 2ndm) + (T - K) \times (4\hat{n}d^2 + 2\hat{n}^2d + 2\hat{n}dm)$. Furthermore, integrating our Mixture-of-Experts (MoE) module incurs additional FLOPs.

Table D. **Gate Weight Distribution (%) of Expert Routing.** $(\cdot)$ is the rank of weights among experts. The **largest** results are indicated in bold and the second-largest results are underlined.

Dataset	Expert #1	Expert #2	Expert #3	Expert #4
CDF-v2$_{real}$	**26.09**(1)	24.22(4)	24.73(3)	24.96(2)
CDF-v2$_{fake}$	25.02(2)	23.84(4)	24.93(3)	**26.21**(1)
FaceShifter$_{real}$	**26.21**(1)	23.77(4)	24.80(3)	25.23(2)
FaceShifter$_{fake}$	25.61(2)	23.27(4)	25.18(3)	**25.94**(1)
Different Forgery Methods:				
UniFace$_{fake}$	25.34(2)	23.28(4)	25.31(3)	**26.07**(1)
BleFace$_{fake}$	25.66(2)	23.36(4)	25.10(3)	**25.87**(1)
MobSwap$_{fake}$	25.32(2)	23.13(4)	25.23(3)	**26.31**(1)
e4s$_{fake}$	25.10(3)	22.93(4)	25.55(2)	**26.42**(1)
FaceDan$_{fake}$	25.31(3)	23.08(4)	25.32(2)	**26.29**(1)
FSGAN$_{fake}$	25.13(3)	23.15(4)	25.39(2)	**26.33**(1)
InSwap$_{fake}$	25.41(2)	23.18(4)	25.28(3)	**26.13**(1)
SimSwap$_{fake}$	25.47(2)	23.32(4)	25.10(3)	**26.10**(1)

B. Detailed Experimental Settings

Comparison of methods for Token Selection. In Tab.5 of the manuscript, under the "Early Layer Pruning" design, we compare our "Token Selection" method with alternative designs. **PCA** denotes the popular feature selection method used in LongCLIP [26]. Here, we apply it to the network's output hidden states and retain the top 32 principal components. **FastV** [3], commonly used in MLLMs, selects tokens based on attention weights. For fair comparison, we use the selection metric for the last 4 layers. **Sparse Attention** [20] is a sparse manner searching for non-semantic features used in image manipulation localization. For fair comparison, we apply it to the last 4 layers with sparse size of 8. Our "Token Selection" method guided by the Spearman rank loss performs best among these designs.

Comparison of methods for multi-scale complementarity. In Tab.6 of the manuscript, under our baseline design, we compare our "Multi-Scale Fusion" with alternative designs. **"Split Ensemble"** is a commonly used deep-learning technique that aggregates the prediction scores from split and original images during both training and inference. **"Multi CLS Pooling"** and **RINE** [11] both leverage the [CLS] tokens from multiple layers to capture multi-scale information. "Multi CLS Pooling" is trained with a learning rate of $1e-6$, consistent with our baseline, while RINE keeps the backbone frozen, following its original configuration. For fair comparison, $\mathbf{D}^3$ [22] uses patch shuffling 4 times and the original image to focus on local forgery. Be-

sides, **MRA** [16] employs the ConvNext network for multi-scale fusion, using high-resolution images resized to 384 pixels. Our approach employs single shared backbone to learn common forgery artifacts from both split and original images, thereby better supporting token selection.

Comparison of different parameter-efficient finetuning methods. In Tab. A, we compare our DFD-HR with other popular PEFT methods. For fair comparison, all methods are finetuned using learning rate of $1e-4$. For **LNTuning** [24], all pre and post LayerNorms are finetuned. **BiasTuning** [2] finetunes only the bias terms of the pre-trained backbone. **VPT-deep** [8] inserts 5 prepend random initialized learnable tokens into the input of all Transformer layers. **SideTuning** [27] trains a "side" AlexNet network and linear interpolates between pretrained features and side-tuned features before being fed into the head. **Partial-k** [25] finetunes the last k layers of the backbone while freezing the others. **MLP-k** utilizes a multilayer perceptron (MLP) with k layers as classification head. **Adapter** [5] inserts new MLP modules with residual connection inside Transformer layers. For *LoRA Variants*, we set the rank to 16 and introduce additional learnable parameters into every linear layer. All other settings follow default of the *PEFT* library.

Detailed ablations of proposed designs. In Tab.8 of the manuscript, we carefully ablate our proposed designs. **For Multi-Scale Fusion**, "SA Pooling" denotes using Self-Attention pooling operation to aggregate local information of split images. **For Early Layer Pruning**, "*w.o.* Gumbel" uses soft routers during training while hard routers during inference. Then, "*w.i.* Shared" uses the shared layer judges across layers. "*w.i.* GAP" introduces global average pooling of all tokens to layer judges instead of the [CLS] token. **For Token Selection**, "Learn.+Weight" denotes the selection metric that uses the average ensemble of learnable scores from token judges and cosine similarity with global query from the original image.

References

[1] Kerim Büyükakyüz. Olora: Orthonormal low-rank adaptation of large language models. *arXiv preprint arXiv:2406.01775*, 2024. 1

[2] Han Cai, Chuang Gan, Ligeng Zhu, and Song Han. Tinytl: Reduce memory, not parameters for efficient on-device learning. *Advances in Neural Information Processing Systems*, 33:11285–11297, 2020. 1, 3

[3] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024. 2

[4] 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 (CVPR)*, pages 19207–19217, 2025. 1

[5] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. 1, 3

[6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 1

[7] Nam Hyeon-Woo, Moon Ye-Bin, and Tae-Hyun Oh. Fedpara: Low-rank hadamard product for communication-efficient federated learning. *arXiv preprint arXiv:2108.06098*, 2021. 1

[8] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pages 709–727. Springer, 2022. 1, 3

[9] Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with lora. *arXiv preprint arXiv:2312.03732*, 2023. 1

[10] Chenqi Kong, Anwei Luo, Peijun Bao, Yi Yu, Haoliang Li, Zengwei Zheng, Shiqi Wang, and Alex C Kot. Moe-ffd: Mixture of experts for generalized and parameter-efficient face forgery detection. *arXiv preprint arXiv:2404.08452*, 2024. 1

[11] Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on Computer Vision*, pages 394–411. Springer, 2024. 2

[12] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A new dataset for deepfake forensics. In *CVPR*, 2020. 2

[13] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022. 1

[14] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024. 1

[15] Weiyang Liu, Zeju Qiu, Yao Feng, Yuliang Xiu, Yuxuan Xue, Longhui Yu, Haiwen Feng, Zhen Liu, Juyeon Heo, Songyou Peng, et al. Parameter-efficient orthogonal finetuning via butterfly factorization. *arXiv preprint arXiv:2311.06243*, 2023. 1

[16] Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *arXiv preprint arXiv:2403.03003*, 2024. 3

[17] Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37:121038–121072, 2024. 1

[18] Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard

Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. *Advances in Neural Information Processing Systems*, 36:79320–79362, 2023. 1

[19] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, pages 1–11, 2019. 2

[20] Lei Su, Xiaochen Ma, Xuekang Zhu, Chaoqun Niu, Zeyu Lei, and Ji-Zhe Zhou. Can we get rid of handcrafted feature extractors? sparsevit: Nonsemantics-centered, parameter-efficient image manipulation localization through spare-coding transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7024–7032, 2025. 2

[21] Zhiyuan Yan, Jiangming Wang, Zhendong Wang, Peng Jin, Ke-Yue Zhang, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Effort: Efficient orthogonal modeling for generalizable ai-generated image detection. In *ICML*, 2025. 1

[22] Yongqi Yang, Zhihao Qian, Ye Zhu, Olga Russakovsky, and Yu Wu. Dˆ 3: Scaling up deepfake detection by learning from discrepancy. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23850–23859, 2025. 2

[23] Shih-Ying Yeh, Yu-Guan Hsieh, Zhidong Gao, Bernard BW Yang, Giyeong Oh, and Yanmin Gong. Navigating text-to-image customization: From lycoris fine-tuning to model evaluation. In *The Twelfth International Conference on Learning Representations*, 2023. 1

[24] Andrii Yermakov, Jan Cech, and Jiri Matas. Unlocking the hidden potential of clip in generalizable deepfake detection. *arXiv preprint arXiv:2503.19683*, 2025. 1, 3

[25] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014. 1, 3

[26] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European conference on computer vision*, pages 310–325. Springer, 2024. 2

[27] Jeffrey O Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. Side-tuning: a baseline for network adaptation via additive side networks. In *European conference on computer vision*, pages 698–714. Springer, 2020. 1, 3

[28] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023. 1