Appendix for "Decoupling Bias, Aligning Distributions: Synergistic Fairness Optimization for Deepfake Detection"

A. Additional Experimental Settings

Table A.1 reports the total numbers of training, validation, and test samples for each dataset, together with the sensitive attributes considered in our experiments. Training and validation are conducted exclusively on FF++ dataset.

Table A.1. Sample usage for training and testing on the FF++, DFD, DFDC, and Celeb-DF datasets. '-' means not used.

Dataset	Samples			Sensitive Attributes		
	Train	Validation	Test	Gender	Race	Intersection
FF++	76139	25386	25401	Male, Female	Asian, Black, White, Others	Male-Asian, Male-Black, Male-White, Male-Others Female-Asian, Female-Black, Female-White, Female-Others
DFD	-	-	9385	Male, Female	Black, White, Others	Male-Black, Male-White, Male-Others Female-Black, Female-White, Female-Others
DFDC	-	-	22857	Male, Female	Asian, Black, White, Others	Male-Asian, Male-Black, Male-White, Male-Others Female-Asian, Female-Black, Female-White, Female-Others
Celeb-DF	-	-	28458	Male, Female	Black, White, Others	Male-Black, Male-White, Male-Others Female-Black, Female-White, Female-Others

B. Additional Robustness Results

As shown in Fig. B.1, all methods are evaluated for robustness on the DFD and DFDC datasets under four types of perturbations: IC (Image Compression), GN (Gaussian Noise), GB (Gaussian Blur), and BWN (Block-wise Noise). The results show that our proposed method is more robust than the other baselines, with its performance remaining almost unchanged under most perturbations and even improving under GN perturbation on the DFD dataset.

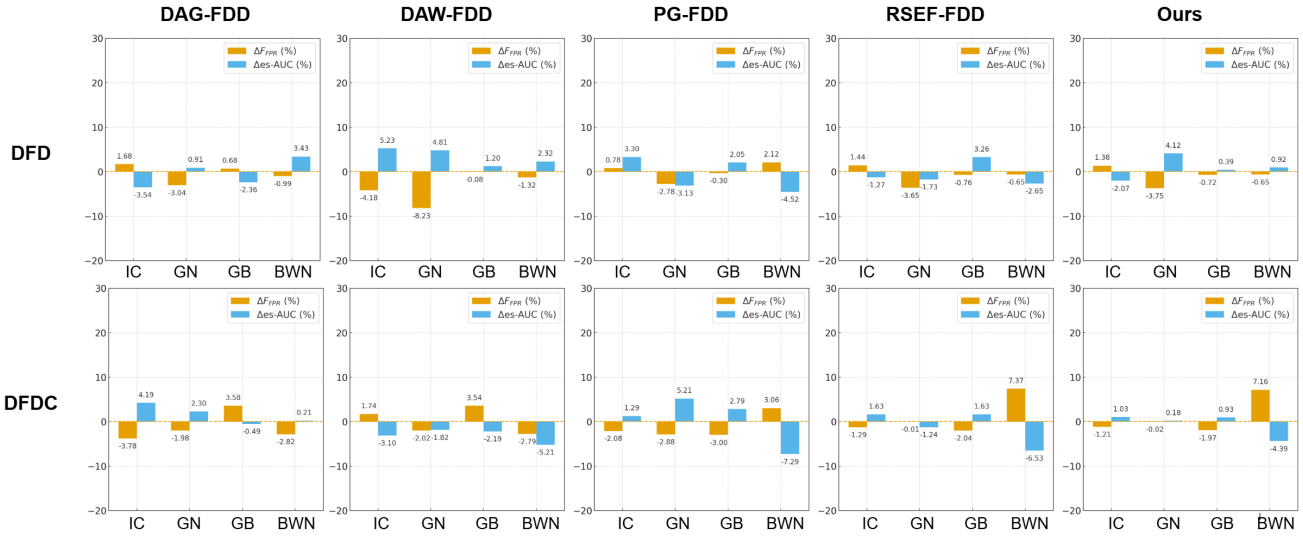

Figure B.1. Additional robustness evaluation on the DFD and DFDC datasets. All methods are tested under four different types of perturbations.

C. Additional Ablation Results

Fig. C.1 presents the analysis of how different decoupling iterations and decoupling ratios within the structural fairness decoupling module affect fairness performance. The results indicate that the optimal trade-off is achieved with a 2% decoupling ratio at the third iteration, which is consistent with the findings for the Xception backbone.

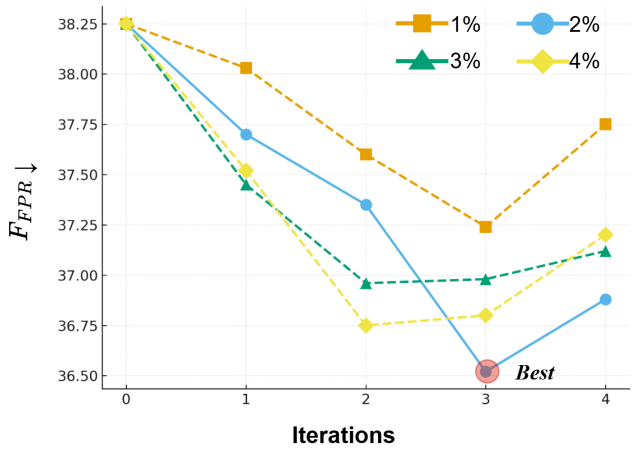

Figure C.1. Analysis of the impact of different decoupling iterations and decoupling ratios on the fairness performance (F_{FPR}) for ResNet-50 backnone.

D. End-to-end Training Algorithm

The following presents the pseudocode of our training optimization procedure, which integrates Structural Fairness Decoupling and Global Distribution Alignment and implements them throughout the end-to-end training process.

Algorithm 1: Training Optimization

Input: Training dataset $\mathcal{D}$ with sensitive attributes, pre-trained model, max_iterations, num_epoch, num_batch, learning rate η, fairness weight λ_{fair}, decoupling ratio pr_c, Sinkhorn regularization coefficient ε, and a set of subgroups $\mathcal{J}$.
Output: A deepfake detection model with improved fairness, parameterized by θ_l.
Initialization: θ_0, $l = 0$.
for $e = 1$ to max_iterations **do**
  For each channel k in the last convolutional layer, compute its fairness index F_k based on Eq. 2 using $\mathcal{D}$;
  Select pr_c percent of the channels with the smallest F_k as the decoupling index set $\mathcal{C}^{(e)}$;
  Apply channel decoupling to the last convolutional layer using $\mathcal{C}^{(e)}$;
  for epoch $= 1$ to num_epoch **do**
    for $b = 1$ to num_batch **do**
      Sample a mini-batch $\mathcal{D}_b$ from $\mathcal{D}$;
      Compute the classification loss $\mathcal{L}_{\text{cls}}$ on $\mathcal{D}_b$ based on Eq. 1;
      Sample an intersectional subgroup from $\mathcal{J}$ and obtain its predictions;
      Compute the fairness loss $\mathcal{L}_{\text{fair}}$ based on Eq. 6;
      Compute the total loss $\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{fair}}\mathcal{L}_{\text{fair}}$;
      Update parameters $\theta_{l+1} \leftarrow \theta_l - \eta\nabla_\theta\mathcal{L}$;
      $l \leftarrow l + 1$;
    end
  end
end
return θ_l

E. Experimental Results of Different Training Set

To verify that the proposed method also exhibits superior performance when trained on other datasets, we additionally train several methods on the DFDC dataset, using Xception and ResNet-50 as backbone networks, respectively. Since PG-FDD [24] requires a specific label for training, which is not provided in DFDC dataset, and RSEF-FDD [18] requires additional redundant samples that we are unable to construct from DFDC dataset, these two methods are not included in this comparison. The results for the Xception-based and ResNet-50-based models are reported in Tab. E.1 and Tab. E.2, respectively, and demonstrate that, when trained on DFDC dataset, our method still achieves superior fairness across the four test datasets while simultaneously attaining the highest detection performance.

Table E.1. Evaluation of methods with Xception backbone across intra-domain (DFDC) and cross-domain (FF++, Celeb-DF, DFD) datasets.

Datasets	Methods	Backbone	Fairness Metrics(%)									Detection Metrics(%)
			Gender			Race			Intersection			Overall
			$F_{FPR}\downarrow$	$F_{DP}\downarrow$	$es-AUC\uparrow$	$F_{FPR}\downarrow$	$F_{DP}\downarrow$	$es-AUC\uparrow$	$F_{FPR}\downarrow$	$F_{DP}\downarrow$	$es-AUC\uparrow$	$AUC\uparrow$
DFDC	Ori	Xception	1.13	7.54	85.90	13.95	23.59	70.76	68.76	40.34	52.91	87.48
	DAG-FDD WACV'24 [20]	Xception	2.27	6.69	82.27	18.43	25.47	64.28	68.90	39.55	47.44	85.31
	DAW-FDD WACV'24 [20]	Xception	1.74	4.47	78.35	35.46	32.22	72.19	72.45	35.20	53.75	78.81
	Fairadapter ICASSP'25 [10]	ViT-L/14	3.88	3.87	87.98	23.35	39.60	80.26	49.47	43.77	72.71	88.87
	Ours	Xception	**0.87**	**3.10**	**88.50**	**7.67**	**22.80**	**82.59**	**17.67**	**27.57**	**72.90**	**89.42**
FF++	Ori	Xception	16.14	16.35	55.29	22.42	16.94	**54.16**	81.51	33.20	**47.02**	56.06
	DAG-FDD WACV'24 [20]	Xception	18.00	23.36	56.06	17.45	11.17	50.60	69.19	29.73	46.13	56.82
	DAW-FDD WACV'24 [20]	Xception	9.67	6.94	52.78	29.98	13.43	46.21	99.66	30.44	36.95	54.78
	Fairadapter ICASSP'25 [10]	ViT-L/14	7.19	4.74	58.84	33.80	16.94	51.05	73.72	29.96	45.53	59.64
	Ours	Xception	**1.01**	**1.33**	**59.82**	**9.00**	**6.26**	51.34	**19.60**	**8.34**	42.82	**60.51**
Celeb-DF	Ori	Xception	9.13	15.43	62.33	11.32	16.95	57.44	14.71	16.19	57.27	66.09
	DAG-FDD WACV'24 [20]	Xception	**1.50**	7.26	51.87	17.23	18.58	46.63	20.68	18.02	44.94	56.43
	DAW-FDD WACV'24 [20]	Xception	19.77	25.13	53.36	24.58	23.85	42.21	23.83	25.55	42.93	58.82
	Fairadapter ICASSP'25 [10]	ViT-L/14	13.29	13.72	49.79	29.12	17.30	46.76	15.99	17.46	40.75	54.01
	Ours	Xception	3.07	**6.96**	**62.51**	**10.32**	**15.94**	**59.56**	**8.74**	**13.44**	**58.36**	**70.21**
DFD	Ori	Xception	7.81	11.44	61.91	13.28	6.19	62.11	48.35	17.02	45.47	67.51
	DAG-FDD WACV'24 [20]	Xception	**0.36**	11.35	65.12	13.02	10.39	63.57	27.30	24.24	57.27	68.20
	DAW-FDD WACV'24 [20]	Xception	14.89	13.39	50.10	14.67	15.89	47.04	45.36	37.80	44.58	53.71
	Fairadapter ICASSP'25 [10]	ViT-L/14	15.12	12.40	68.17	28.18	9.26	65.72	48.29	38.00	57.65	70.78
	Ours	Xception	6.25	**10.94**	**68.61**	**12.33**	**6.09**	**66.87**	**19.26**	**12.00**	**60.95**	**70.83**

Table E.2. Evaluation of methods with ResNet-50 backbone across intra-domain (DFDC) and cross-domain (FF++, Celeb-DF, DFD) datasets.

Datasets	Methods	Backbone	Fairness Metrics(%)									Detection Metrics(%)
			Gender			Race			Intersection			Overall
			$F_{FPR}\downarrow$	$F_{DP}\downarrow$	$es-AUC\uparrow$	$F_{FPR}\downarrow$	$F_{DP}\downarrow$	$es-AUC\uparrow$	$F_{FPR}\downarrow$	$F_{DP}\downarrow$	$es-AUC\uparrow$	$AUC\uparrow$
DFDC	Ori	ResNet-50	3.19	12.60	91.66	10.18	28.48	83.38	39.22	40.33	69.57	92.43
	DAG-FDD WACV'24 [20]	ResNet-50	2.30	14.24	91.76	9.06	**27.39**	81.53	27.43	38.95	66.70	92.47
	DAW-FDD WACV'24 [20]	ResNet-50	2.06	12.89	90.56	11.13	28.46	80.34	29.95	39.83	64.24	92.01
	Fairadapter ICASSP'25 [10]	ViT-L/14	3.88	10.87	87.98	23.35	39.60	80.26	49.47	43.77	**72.71**	88.87
	Ours	ResNet-50	**1.44**	**10.44**	**92.54**	**7.62**	28.23	**84.99**	**20.69**	**37.19**	72.12	**93.92**
FF++	Ori	ResNet-50	10.77	11.89	57.26	41.89	22.92	50.34	95.46	35.10	42.76	57.66
	DAG-FDD WACV'24 [20]	ResNet-50	8.68	9.44	58.72	21.09	19.06	50.24	57.90	22.92	43.89	59.22
	DAW-FDD WACV'24 [20]	ResNet-50	14.85	16.54	57.38	21.41	12.57	48.82	62.30	32.43	42.63	57.61
	Fairadapter ICASSP'25 [10]	ViT-L/14	8.59	**4.74**	58.84	33.80	16.94	51.05	73.72	29.96	45.53	59.64
	Ours	ResNet-50	**8.49**	13.02	**59.89**	**10.52**	**6.05**	**54.18**	**39.70**	**22.29**	**45.75**	**60.25**
Celeb-DF	Ori	ResNet-50	3.14	5.40	59.20	26.33	29.42	**62.44**	27.29	19.14	50.72	66.15
	DAG-FDD WACV'24 [20]	ResNet-50	**0.21**	5.27	60.49	26.52	27.08	55.19	31.88	29.02	53.40	63.78
	DAW-FDD WACV'24 [20]	ResNet-50	3.79	4.41	58.81	33.51	37.95	49.78	35.01	35.73	48.55	64.02
	Fairadapter ICASSP'25 [10]	ViT-L/14	13.29	13.72	49.79	29.12	**17.30**	46.76	28.99	**17.46**	40.75	54.01
	Ours	ResNet-50	2.32	**4.38**	**65.82**	**25.96**	39.51	56.70	**26.70**	34.79	**55.72**	**73.28**
DFD	Ori	ResNet-50	11.89	7.88	66.20	11.07	24.81	53.77	73.70	43.89	46.76	63.76
	DAG-FDD WACV'24 [20]	ResNet-50	5.22	6.85	57.48	6.69	17.95	50.66	34.71	26.31	46.82	58.75
	DAW-FDD WACV'24 [20]	ResNet-50	18.99	18.81	60.99	4.02	22.74	52.84	50.34	45.11	49.43	64.03
	Fairadapter ICASSP'25 [10]	ViT-L/14	15.12	12.40	**68.17**	28.18	19.26	54.72	48.29	38.00	**57.65**	60.78
	Ours	ResNet-50	**0.61**	**0.88**	63.24	**1.08**	**13.29**	**54.91**	**12.20**	**15.93**	45.08	**64.35**

F. Limitation and Future Work

One limitation of our method lies in its reliance on datasets with accurate demographic annotations. Currently, fairness-related annotations in existing datasets are quite limited; therefore, our fairness analysis can only focus on gender and race. In future

work, we plan to enrich existing datasets with more fairness-related labels, and even construct our own dataset to enable more comprehensive and in-depth research.

G. Effect of Trade-off Hyperparameter λ

To verify the effect of the trade-off hyperparameter in Eq. 7, we conducted sensitivity analysis on the FF++ dataset. Fig. G.1 shows the fairness metrics and detection metric AUC for different λ values. The experimental results indicate that when λ is set to 0.005, the model achieves optimal fairness performance while maintaining fair AUC scores. It is noteworthy that the analysis reveals a trade-off between fairness and AUC scores: as λ increases from 0.003 to 0.005, AUC decreases, but fairness improves. To more clearly illustrate the relationship between each fairness metric and AUC, we present these dynamic changes separately in Fig. G.2, which displays the trend of increased AUC corresponding to reduced fairness.

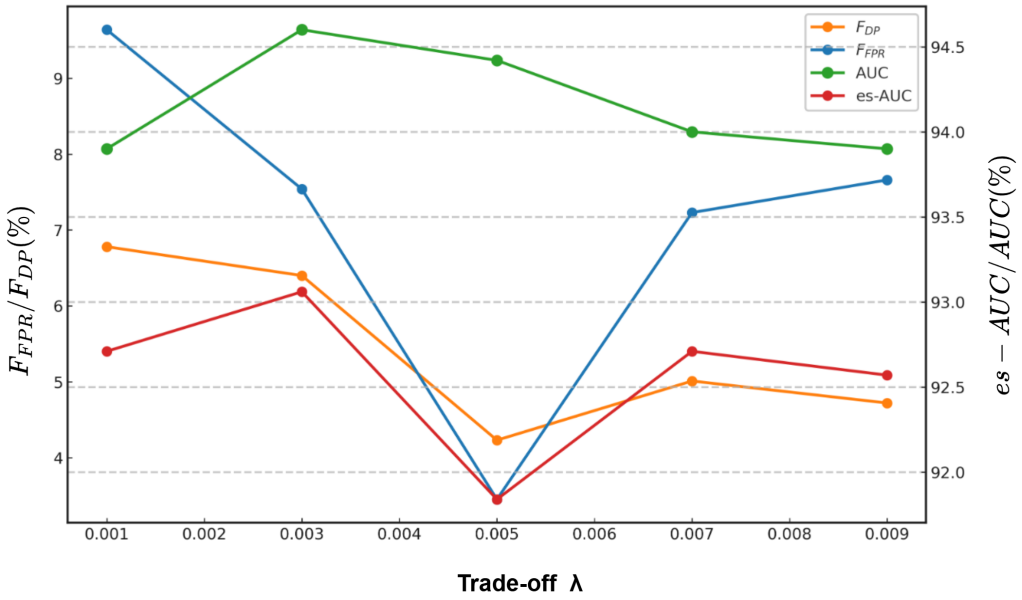

Figure G.1. Sensitivity analysis of parameter λ on the trade-off between fairness and detection accuracy for the gender attribute on FF++.

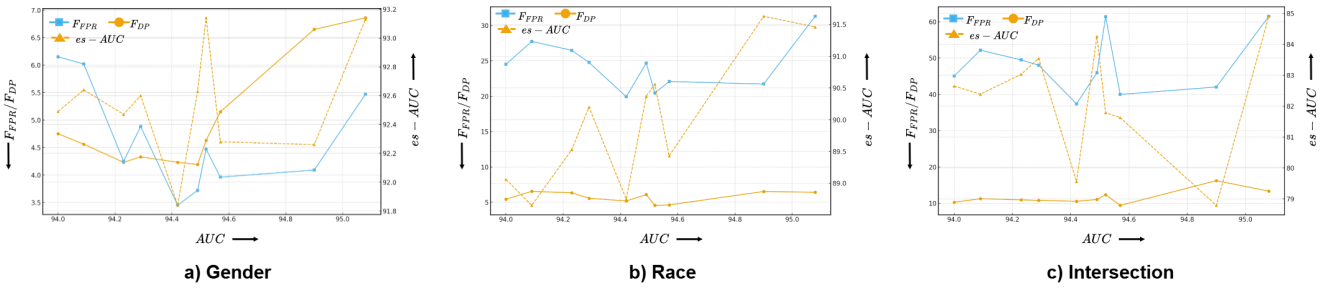

Figure G.2. Trends in Fairness Metrics vs. AUC Score. From left to right, the chart shows how fairness metrics for gender, race, and intersection attribute change with AUC, illustrating the trade-off between accuracy and fairness.

H. The T-SNE Visualization of Demographic Features

We present in Fig. H.1 the t-SNE visualizations of demographic features extracted on FF++ by the vanilla Xception model and by our method, respectively. In the visualization, the different intersectional demographic groups extracted by the unconstrained Xception model form clearly separated clusters, whereas the features extracted by our method show these intersectional groups intermingled in the feature space. This indicates that our model has discovered common fingerprints

across demographics, leading to a more fair representation. The t-SNE results further reveal that most subgroups in FF++ belong to the Male-White and Female-White categories, highlighting a strong dataset bias that makes fair detection particularly challenging and underscoring the necessity of our dual-mechanism synergistic optimization framework to improve fairness for minority subgroups.

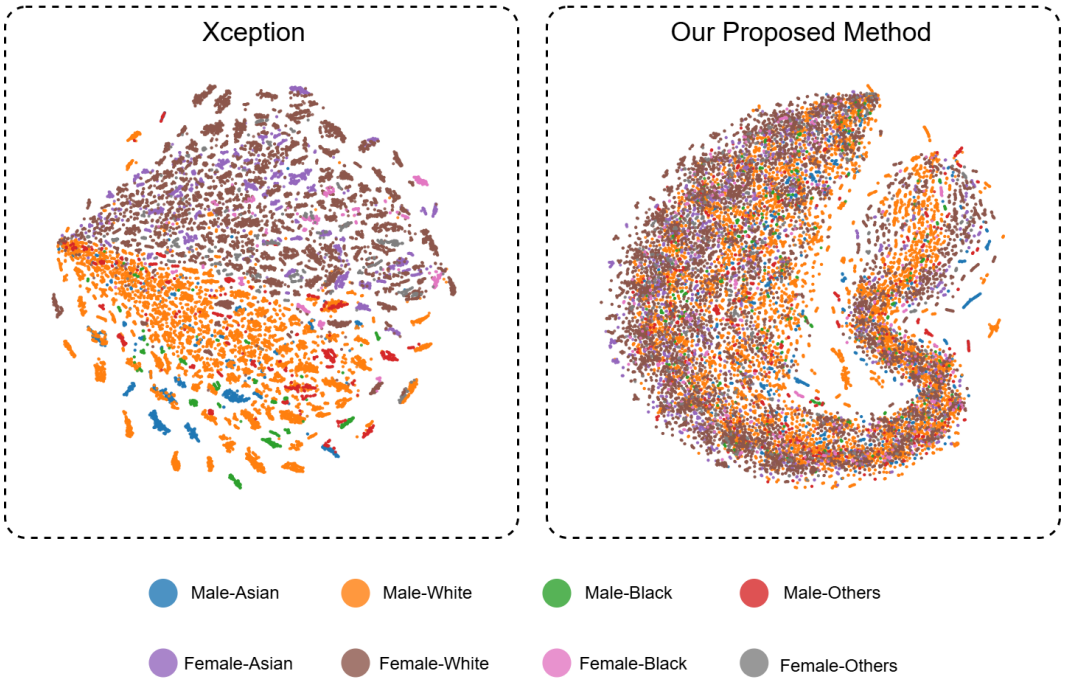

Figure H.1. The T-SNE visualization of demographic features extracted from Xception and our method on FF++ dataset.

I. Additional Experimental Results

For a fairer comparison, we conducted additional experiments on a recent public fair forgery detection benchmark [26] and included more comparison methods under our experimental settings. The results in Tables I and I show that our method still outperforms all compared methods on this benchmark. The results on all four datasets, with more comparison methods included, consistently show that our method outperforms the other compared methods.

Table I.1. Additional experimental results on the FairFD benchmark. Best results are in **bold** and second best are underlined.

Fairness Metric		Spatial-based						Frequency-based			Fairness-enhanced				
		Xception	RECCE	UCF	Capsule	FFD	CORE	F3Net	SPSL	SRM	DAG	DAW	PFGDFD	RSEF-FDD	Ours
Naive Metric	DPD↓	0.1810	0.1338	0.1765	0.0969	0.1099	0.0951	0.0674	0.0203	0.0990	0.1723	0.0513	0.0805	0.0378	**0.0159**
	DEOdds↓	0.1666	0.1264	0.1495	0.0902	0.1005	0.0798	0.0763	0.0304	0.0714	0.2288	0.0593	0.1396	0.0693	**0.0118**
	DEO↓	0.2088	0.1548	0.2014	0.1118	0.1242	0.1084	0.0801	0.0215	0.1090	0.2105	0.0611	0.1032	0.0513	**0.0129**
	STD↓	0.0647	0.0474	0.0631	0.0343	0.0398	0.0342	0.0265	**0.0080**	0.0355	0.0636	0.0195	0.0328	0.0134	0.0148
Approach Averaged	AADPD↓	0.2024	0.1572	0.2175	0.1323	0.1552	0.1147	0.1158	0.0556	0.1413	0.2201	0.0735	0.1393	0.0707	**0.0438**
	AADEOdds↓	0.1669	0.1302	0.1630	0.1034	0.1196	0.0858	0.0961	0.0481	0.0925	0.2324	0.0662	0.1560	0.0755	**0.0306**
	AADEO↓	0.2095	0.1626	0.2284	0.1381	0.1623	0.1205	0.1197	0.0571	0.1511	0.2177	0.0749	0.1360	0.0738	**0.0433**
	AASTD↓	0.0750	0.0578	0.0809	0.0493	0.0576	0.0449	0.0448	0.0219	0.0531	0.0834	0.0283	0.0530	0.0273	**0.0154**
Utility Regularized	URDPD↓	0.1357	0.1118	0.1523	0.0808	0.1037	0.0803	0.0806	0.0320	0.0904	0.1474	0.0555	0.0881	0.0562	**0.0255**
	URDEOdds↓	0.1057	0.0852	0.1069	0.0639	0.0763	0.0567	0.0625	0.0299	0.0584	0.1445	0.0440	0.0986	0.0495	**0.0210**
	URDEO↓	0.1417	0.1171	0.1614	0.0842	0.1092	0.0850	0.0842	0.0324	0.0968	0.1480	0.0578	0.0860	0.0595	**0.0264**
	URSTD↓	0.0501	0.0410	0.0565	0.0301	0.0384	0.0313	0.0312	0.0126	0.0339	0.0559	0.0214	0.0335	0.0216	**0.0119**
Utility	AUC↑	0.6911	0.6897	0.7214	0.6815	**0.7304**	0.6864	0.6564	0.6763	0.7102	0.6672	0.6604	0.6302	0.6413	0.7026

Table I.2. Additional experimental results on the FairFD benchmark, where "-BPFA" denotes models processed using the benchmark's BPFA method. Best results are in **bold** and second best are underlined.

Fairness Metric		Spatial-based						Frequency-based			Fairness-enhanced				
		Xception-BPFA	RECCE-BPFA	UCF-BPFA	Capsule-BPFA	FFD-BPFA	CORE-BPFA	F3Net-BPFA	SPSL-BPFA	SRM-BPFA	DAG-BPFA	DAW-BPFA	PFGDFD-BPFA	RSEF-FDD	Ours
Naive Metric	DPD↓	0.1644	0.1328	0.1705	0.0951	0.1096	0.0946	0.0674	0.0181	0.0993	0.1286	0.0510	0.0594	0.0378	**0.0159**
	DEOdds↓	0.1532	0.1327	0.1571	0.1011	0.0999	0.0837	0.0762	0.0209	0.0732	0.1807	0.0553	0.1337	0.0693	**0.0118**
	DEO↓	0.1899	0.1550	0.1967	0.1119	0.1237	0.1085	0.0801	0.0200	0.1090	0.1587	0.0602	0.0796	0.0513	**0.0129**
	STD↓	0.0590	0.0471	0.0608	0.0337	0.0397	0.0341	0.0264	**0.0072**	0.0356	0.0457	0.0198	0.0238	0.0134	0.0148
Approach Averaged	AADPD↓	0.1854	0.1584	0.2159	0.1342	0.1546	0.1155	0.1155	0.0473	0.1417	0.1648	0.0774	0.1079	0.0707	**0.0438**
	AADEOdds↓	0.1540	0.1366	0.1711	0.1143	0.1189	0.0898	0.0959	0.0357	0.0943	0.1820	0.0651	0.1442	0.0755	**0.0306**
	AADEO↓	0.1917	0.1628	0.2249	0.1382	0.1617	0.1206	0.1195	0.0496	0.1511	0.1614	0.0799	0.1006	0.0738	**0.0433**
	AASTD↓	0.0693	0.0582	0.0803	0.0501	0.0574	0.0450	0.0447	0.0182	0.0532	0.0613	0.0298	0.0411	0.0273	**0.0154**
Utility Regularized	URDPD↓	0.1218	0.1125	0.1507	0.0820	0.1032	0.0807	0.0804	0.0265	0.0906	0.1107	0.0565	0.0644	0.0562	**0.0255**
	URDEOdds↓	0.0964	0.0888	0.1109	0.0708	0.0758	0.0589	0.0623	0.0218	0.0595	0.1131	0.0430	0.0969	0.0495	**0.0210**
	URDEO↓	0.1268	0.1173	0.1587	0.0843	0.1087	0.0851	0.0840	0.0275	0.0968	0.1102	0.0592	0.0578	0.0595	**0.0264**
	URSTD↓	0.0454	0.0413	0.0559	0.0306	0.0382	0.0313	0.0311	**0.0102**	0.0339	0.0412	0.0218	0.0245	0.0216	0.0119
Utility	AUC↑	0.6952	0.6877	0.7226	0.6794	**0.7305**	0.6874	0.6577	0.6862	0.7100	0.6512	0.6668	0.6445	0.6413	0.7026

Table I.3. Additional experimental results on the FF++ dataset with more included methods. Best results are in **bold** and second best are underlined.

Fairness Metric		Spatial-based						Frequency-based			Fairness-enhanced					
		Xception	RECCE	UCF	Capsule	FFD	CORE	F3Net	SPSL	SRM	DAG	DAW	PFGDFD	Fairadapter	RSEF-FDD	Ours
Gender	$F_{FPR}\downarrow$	4.10	0.64	8.87	34.54	3.07	1.51	2.05	11.51	10.06	1.82	0.78	0.62	4.16	0.57	**0.53**
	$F_{DP}\downarrow$	5.72	3.88	6.44	18.65	3.88	3.89	4.60	3.72	4.24	4.65	9.52	4.74	12.21	8.55	**3.61**
	$es-AUC\uparrow$	91.93	85.76	86.38	74.85	85.84	85.88	88.23	87.21	84.99	94.87	95.76	96.32	67.85	94.91	**96.45**
Race	$F_{FPR}\downarrow$	19.76	26.05	32.02	44.47	23.22	10.81	16.20	35.03	26.13	5.48	**5.43**	11.13	43.22	8.39	9.29
	$F_{DP}\downarrow$	4.74	4.66	4.55	8.34	4.86	4.42	5.33	5.13	4.79	9.20	14.69	4.78	20.39	5.28	**4.35**
	$es-AUC\uparrow$	82.85	80.63	84.53	64.57	80.82	82.62	83.75	81.06	81.98	93.43	94.15	94.52	56.03	93.60	**94.86**
Intersection	$F_{FPR}\downarrow$	36.03	61.92	73.40	141.99	50.39	26.62	35.04	37.73	82.75	24.08	14.36	**9.19**	86.91	23.64	20.18
	$F_{DP}\downarrow$	14.64	10.05	10.07	26.62	10.78	9.98	10.07	10.35	**6.68**	26.25	26.13	13.39	42.44	21.74	9.47
	$es-AUC\uparrow$	74.43	69.75	70.96	52.46	72.29	70.86	75.19	68.91	68.15	85.80	86.74	86.83	45.98	85.77	**86.91**
Overall	$AUC\uparrow$	92.69	88.50	89.33	79.15	87.07	88.31	90.45	89.87	88.46	96.72	97.46	97.66	71.50	97.09	**97.71**

Table I.4. Additional experimental results on the DFD dataset with more included methods. Best results are in **bold** and second best are underlined.

Fairness Metric		Spatial-based						Frequency-based			Fairness-enhanced					
		Xception	RECCE	UCF	Capsule	FFD	CORE	F3Net	SPSL	SRM	DAG	DAW	PFGDFD	Fairadapter	RSEF-FDD	Ours
Gender	$F_{FPR}\downarrow$	9.45	5.89	6.06	12.86	10.92	5.71	5.47	5.76	16.14	5.41	5.30	5.05	6.32	18.50	**4.72**
	$F_{DP}\downarrow$	6.63	6.40	6.02	5.69	7.76	6.91	6.63	6.84	9.23	5.60	10.63	**1.86**	7.55	6.11	5.49
	$es-AUC\uparrow$	72.68	75.41	75.47	75.03	72.76	72.34	72.17	75.26	73.26	75.87	71.68	76.37	63.28	75.97	**78.98**
Race	$F_{FPR}\downarrow$	7.75	4.21	2.78	4.23	2.79	9.36	5.59	18.43	14.20	2.56	4.32	5.79	11.29	19.45	**2.09**
	$F_{DP}\downarrow$	22.31	5.62	7.14	**0.09**	6.87	5.02	11.14	5.81	2.70	21.34	20.16	19.31	12.28	19.40	18.66
	$es-AUC\uparrow$	69.07	73.18	74.11	71.82	72.03	76.83	74.45	71.91	75.06	70.14	63.24	74.81	56.28	72.65	**77.44**
Intersection	$F_{FPR}\downarrow$	35.06	28.46	28.50	29.00	29.75	28.33	27.73	65.35	36.50	33.00	33.81	28.00	29.56	35.34	**27.32**
	$F_{DP}\downarrow$	25.53	23.33	28.28	28.37	26.32	22.83	28.27	23.33	25.66	23.02	24.72	23.93	33.11	**10.97**	22.81
	$es-AUC\uparrow$	59.20	67.25	68.89	60.86	66.01	62.22	62.82	67.12	69.37	61.09	56.01	66.32	48.63	67.24	**69.73**
Overall	$AUC\uparrow$	74.32	79.86	80.93	76.48	75.26	80.97	80.77	80.51	81.22	76.29	73.71	80.70	68.12	80.54	**81.46**

Table I.5. Additional experimental results on the DFDC dataset with more included methods. Best results are in **bold** and second best are underlined.

Fairness Metric		Spatial-based						Frequency-based			Fairness-enhanced					
		Xception	RECCE	UCF	Capsule	FFD	CORE	F3Net	SPSL	SRM	DAG	DAW	PFGDFD	Fairadapter	RSEF-FDD	Ours
Gender	$F_{FPR} \downarrow$	8.67	6.70	7.94	7.56	6.34	5.76	4.60	3.04	10.20	5.29	3.60	2.35	2.39	2.09	**1.76**
	$F_{DP} \downarrow$	6.70	5.81	9.68	5.21	5.41	6.00	4.19	4.38	7.73	6.68	3.67	**2.57**	4.57	3.83	3.67
	$es-AUC \uparrow$	54.56	56.98	55.15	56.98	50.58	51.22	57.64	57.54	51.61	57.41	55.88	57.71	55.88	57.90	**57.94**
Race	$F_{FPR} \downarrow$	19.41	19.02	14.64	6.97	12.22	11.58	16.64	20.21	**5.76**	9.26	22.36	10.10	32.49	12.24	17.93
	$F_{DP} \downarrow$	7.99	8.36	8.67	7.77	11.80	7.69	14.42	15.93	11.06	12.51	9.72	11.49	26.77	7.74	**7.58**
	$es-AUC \uparrow$	47.02	46.05	51.94	42.83	49.25	51.23	45.36	46.07	49.77	45.91	47.77	50.53	52.27	50.27	**52.33**
Intersection	$F_{FPR} \downarrow$	53.42	60.30	53.98	58.34	49.31	44.75	41.88	42.98	49.97	45.30	46.80	**32.16**	74.10	38.89	37.37
	$F_{DP} \downarrow$	17.34	12.60	18.79	22.24	20.63	14.79	19.52	21.95	22.44	12.92	12.69	17.71	31.53	12.18	**11.92**
	$es-AUC \uparrow$	36.96	34.48	37.32	30.05	36.38	37.35	32.49	36.91	37.23	36.42	36.63	37.42	31.39	37.72	**39.03**
Overall	$AUC \uparrow$	56.13	61.35	60.96	57.39	60.65	61.84	59.73	60.55	61.58	59.13	56.90	59.64	59.25	59.08	**61.86**

Table I.6. Additional experimental results on the Celeb-DF dataset with more included methods. Best results are in **bold** and second best are underlined.

Fairness Metric		Spatial-based						Frequency-based			Fairness-enhanced					
		Xception	RECCE	UCF	Capsule	FFD	CORE	F3Net	SPSL	SRM	DAG	DAW	PFGDFD	Fairadapter	RSEF-FDD	Ours
Gender	$F_{FPR} \downarrow$	6.93	13.27	5.70	20.74	9.19	5.05	9.24	15.52	**1.18**	8.70	8.71	7.95	8.59	1.94	6.41
	$F_{DP} \downarrow$	20.23	14.69	27.82	12.55	12.84	15.52	13.62	22.16	10.90	14.25	12.73	16.29	26.84	35.63	**10.81**
	$es-AUC \uparrow$	60.94	61.27	65.23	68.92	65.01	67.51	63.77	64.51	61.66	66.23	68.31	69.47	56.84	68.92	**69.68**
Race	$F_{FPR} \downarrow$	23.44	28.41	36.72	34.69	17.55	17.78	16.80	17.49	31.58	19.44	**14.76**	23.99	46.14	17.40	16.60
	$F_{DP} \downarrow$	17.89	26.26	21.27	34.26	22.72	20.97	19.84	20.01	10.42	13.26	7.80	12.47	36.35	27.48	**7.29**
	$es-AUC \uparrow$	61.21	59.40	66.41	60.61	66.12	69.86	61.35	63.63	51.73	62.61	66.47	67.78	58.34	67.80	**69.96**
Intersection	$F_{FPR} \downarrow$	27.38	31.69	51.71	62.46	30.52	30.04	27.90	33.32	23.81	23.60	24.75	23.71	47.03	24.17	**23.32**
	$F_{DP} \downarrow$	18.72	39.04	26.00	52.04	38.39	27.40	29.26	30.17	16.75	13.80	13.78	13.30	35.69	27.73	**12.01**
	$es-AUC \uparrow$	61.22	62.24	**63.52**	61.67	62.63	62.87	62.44	61.31	57.34	63.08	59.24	62.63	54.77	62.56	63.27
Overall	$AUC \uparrow$	69.18	73.96	73.21	72.47	74.22	70.03	76.70	75.58	73.21	72.29	74.09	74.24	64.94	74.36	**76.75**