中国科学院自动化研究所
Institute of Automation, Chinese Academy of Sciences

Recent Progress on Deepfake Detection

Prof. Zhen Lei, FIEEE

State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences

2025.8

Contents

1 Background in Deepfake Detection

2 WMamba: Wavelet-based Mamba for Deepfake Detection

3 FDJL: Forgery-Discomfort Joint Learning for Generalized Deepfake Detection

Background: Game




Real or Fake?

Background: Game

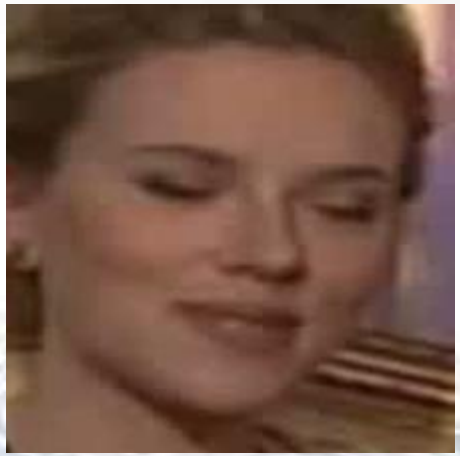
Fake

Fake

Fake

Fake

Real



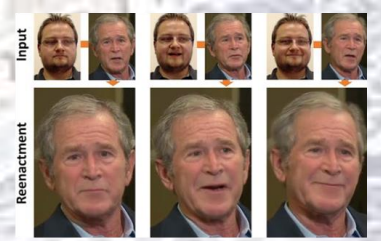
Real

Real

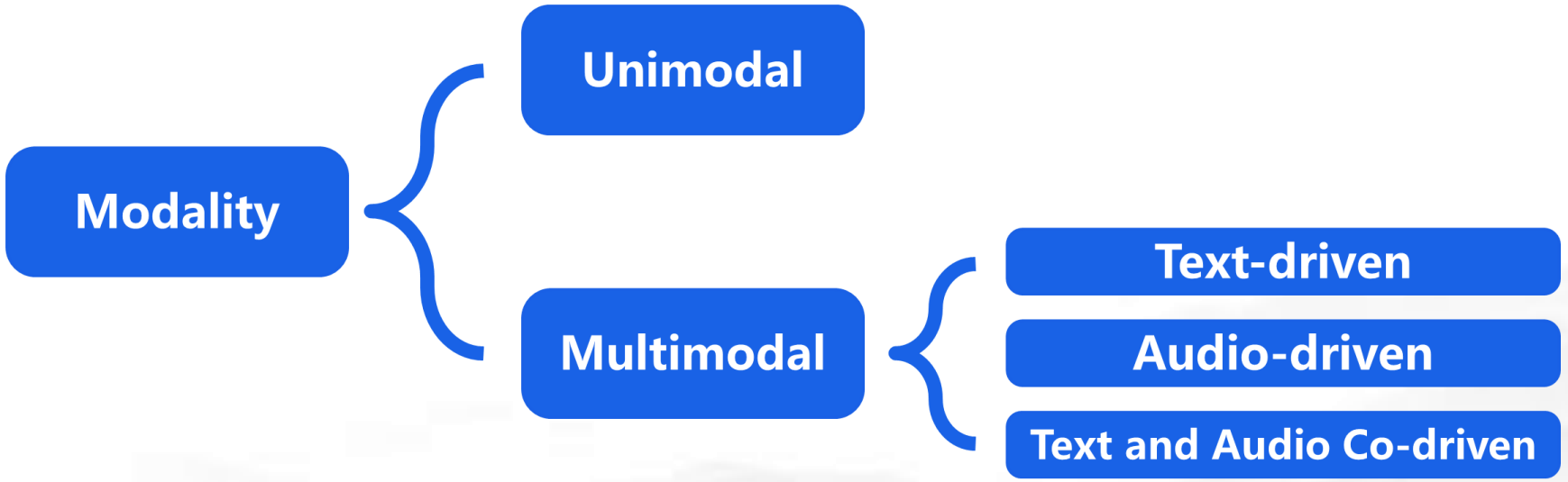
Real

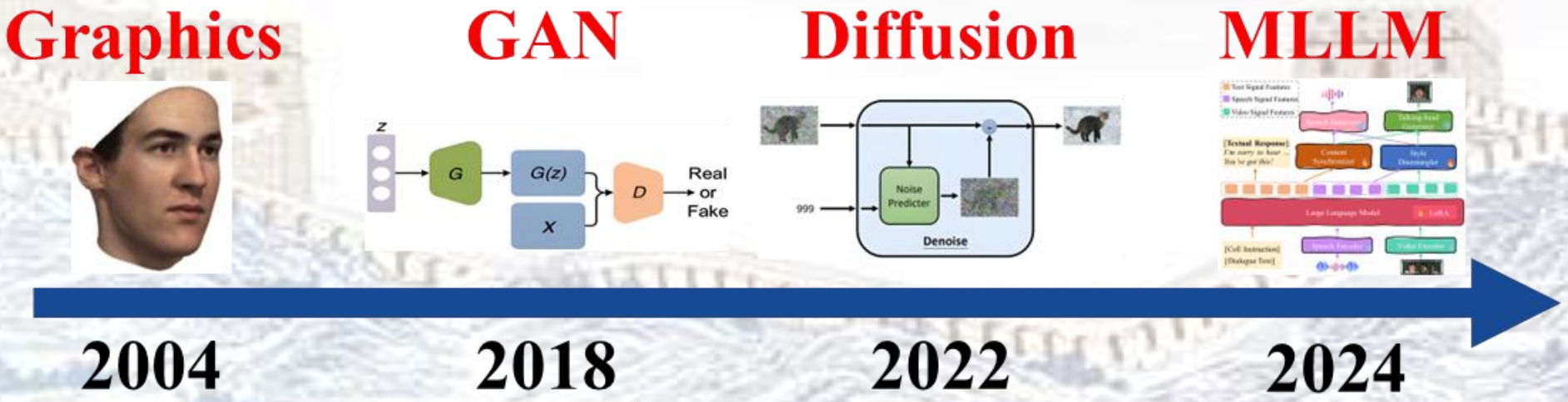
Background: Deepfake Generation

- **Deepfake Generation**
  - **Entire Face Synthesis**
    - Face Generation
    - Face Inpainting
    - Face Super-Resolution
  - **Face Swapping**
    - Face2Face
    - Face2Head
    - Expression Transfer
  - **Face Editing**
    - Attribute Editing
    - Expression Editing
    - Pose Editing
  - **Face Reenactment**
    - 2D Reenactment
    - 3D Reenactment
    - Audio Reenactment



Background: Deepfake Generation





Background: Threats and Solution

Explicit, AI-generated Taylor Swift images spread quickly on social media

By Samantha Murphy Kelly, CNN
4 min read · Published 3:19 PM EST, Thu January 25, 2024

Finance worker pays out $25 million after video call with deepfake ‘chief financial officer’

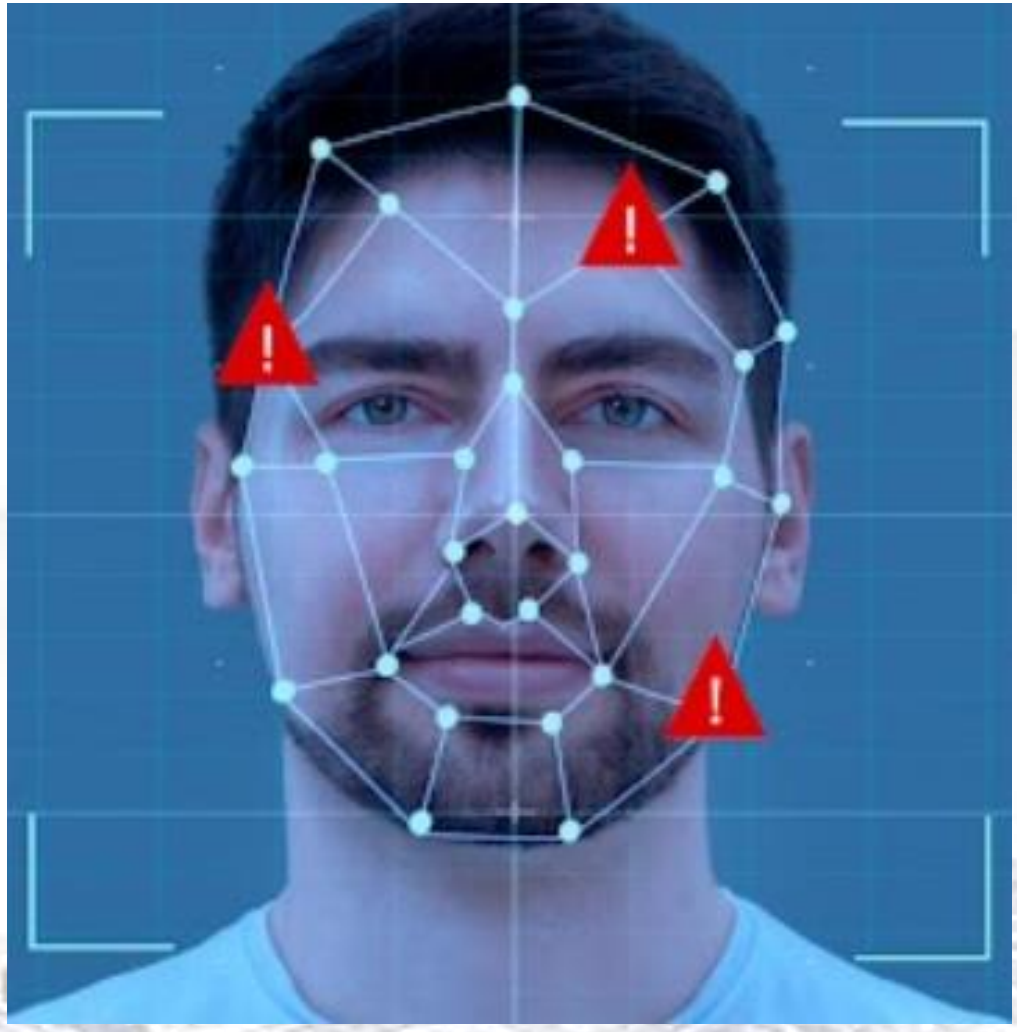
By Heather Chen and Kathleen Magramo, CNN
2 min read · Published 2:31 AM EST, Sun February 4, 2024

Deepfakes can cause long-lasting damage to children

Analysis by Kara Alaimo, CNN
5 min read · Updated 9:26 AM EST, Thu January 2, 2025

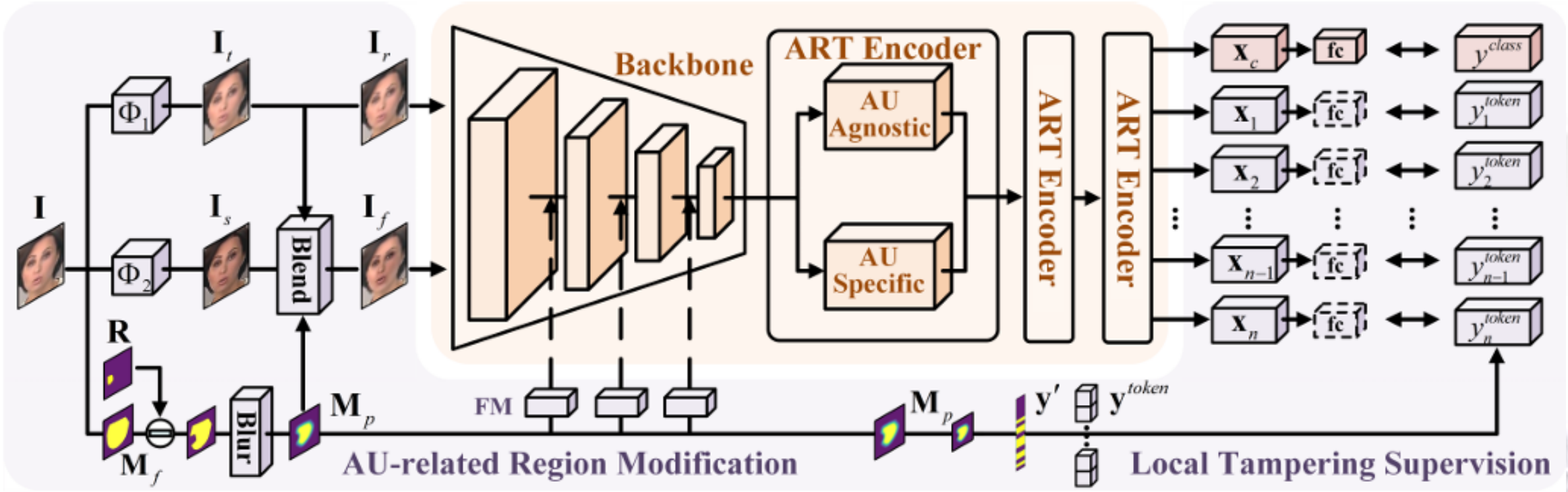
Victims of explicit deepfakes will now be able to take legal action against people who create them

By Clare Duffy, CNN
4 min read · Updated 10:29 PM EDT, Mon May 19, 2025



Problem: Forgery Feature Representation

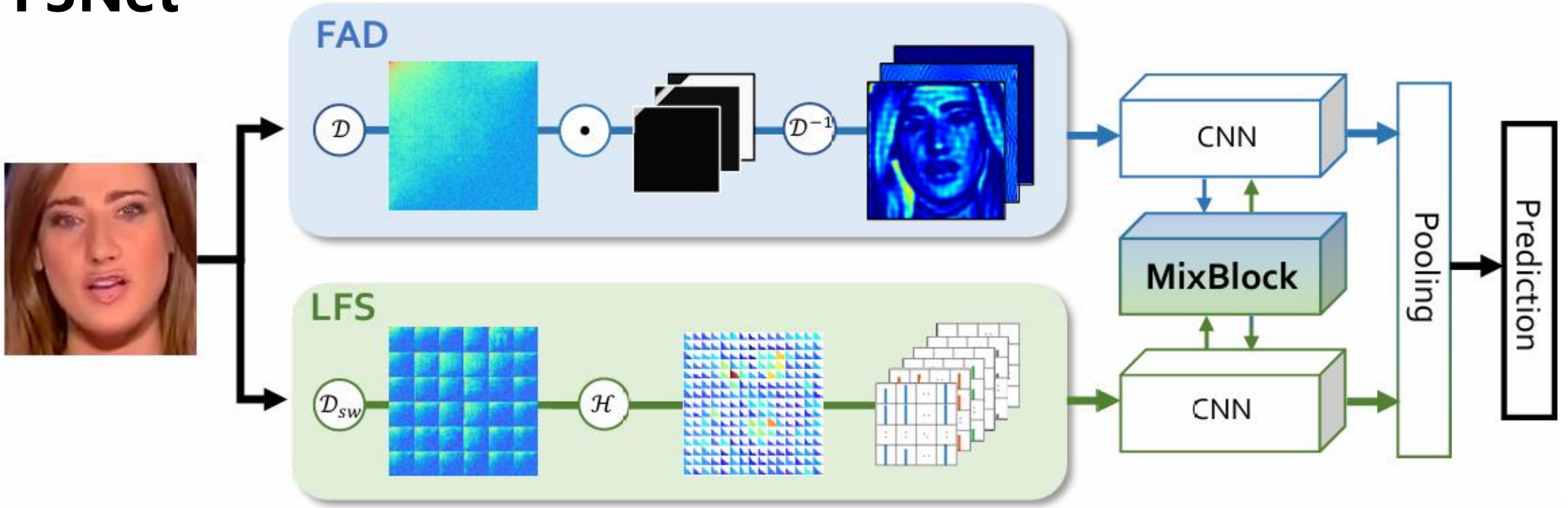
AUNet



Utilizing binary masks to represent forgery feature.

Problem: Forgery Feature Representation

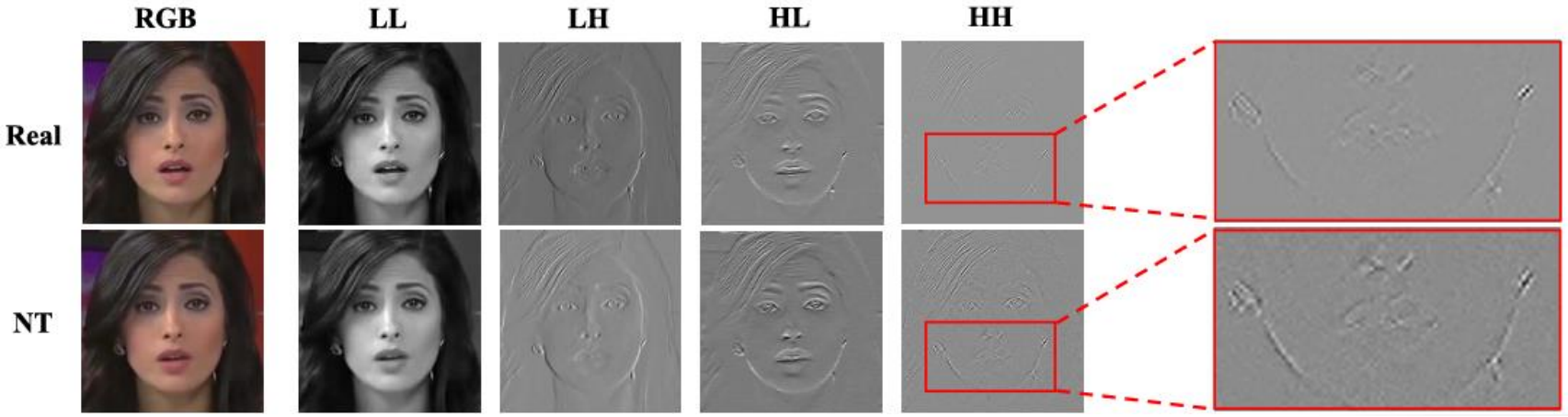
F3Net

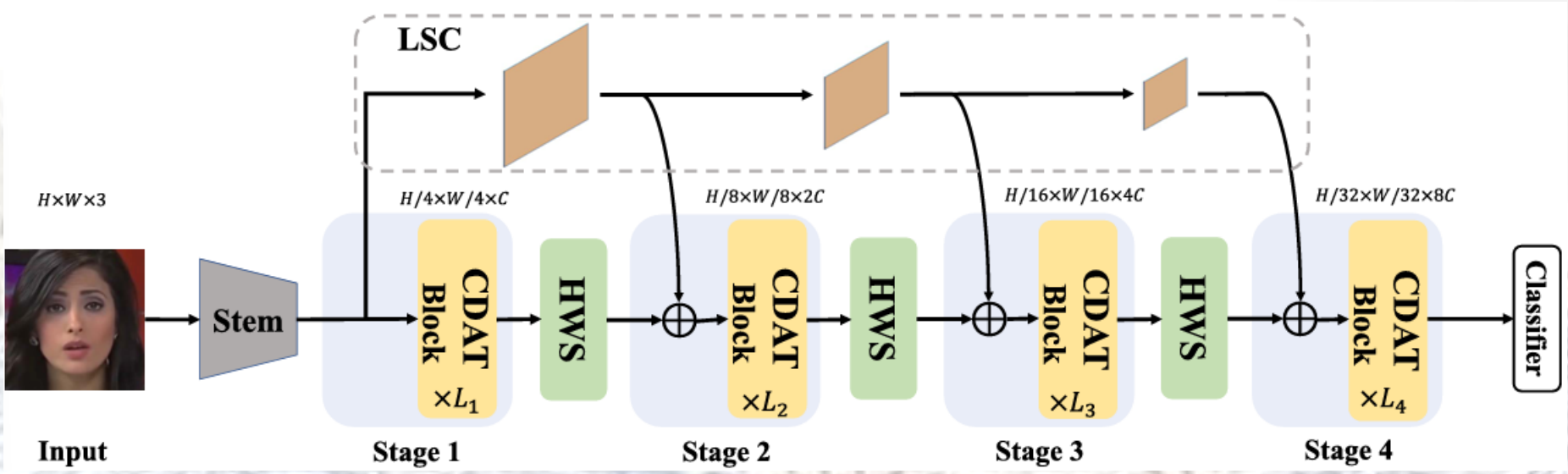


Utilizing Discrete Cosine Transform (DCT) and Local Frequency Statistics (LFS) for forgery feature representation.

Problem: Forgery Feature Representation

F2Trans





Utilizing wavelet transformation to represent forgery feature.

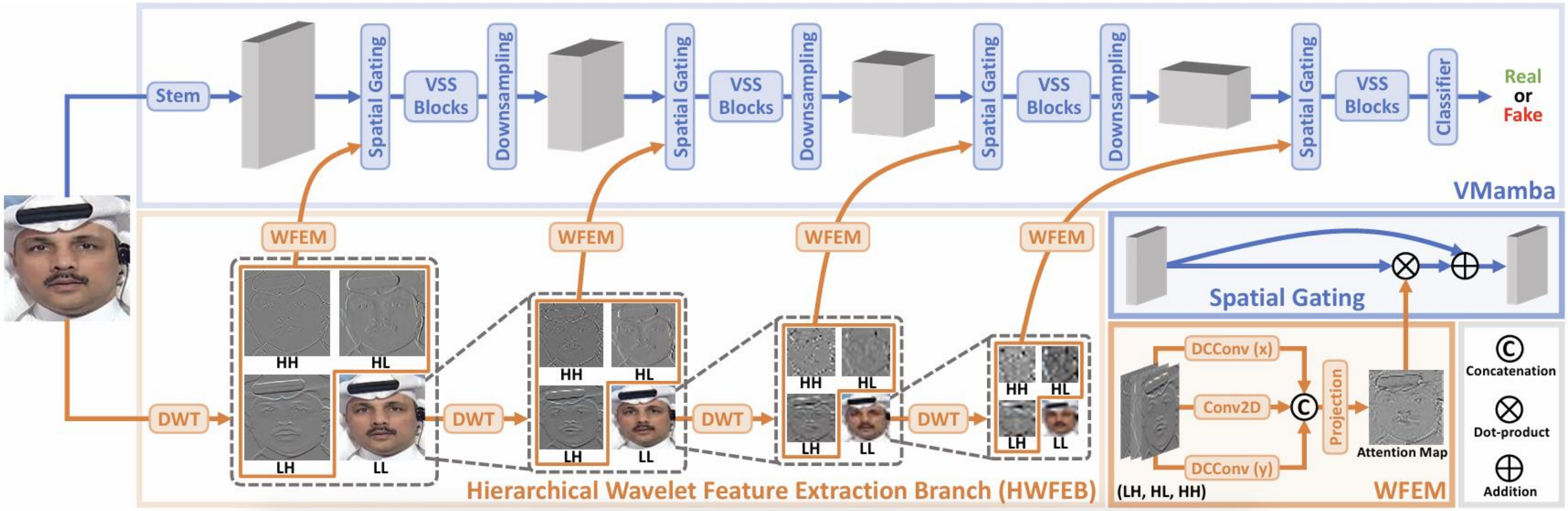
Contents

1 Background in Deepfake Detection

2 WMamba: Wavelet-based Mamba for Deepfake Detection

3 FDJL: Forgery-Discomfort Joint Learning for Generalized Deepfake Detection

WMamba: Architecture



WMamba: Dynamic Contour Convolution

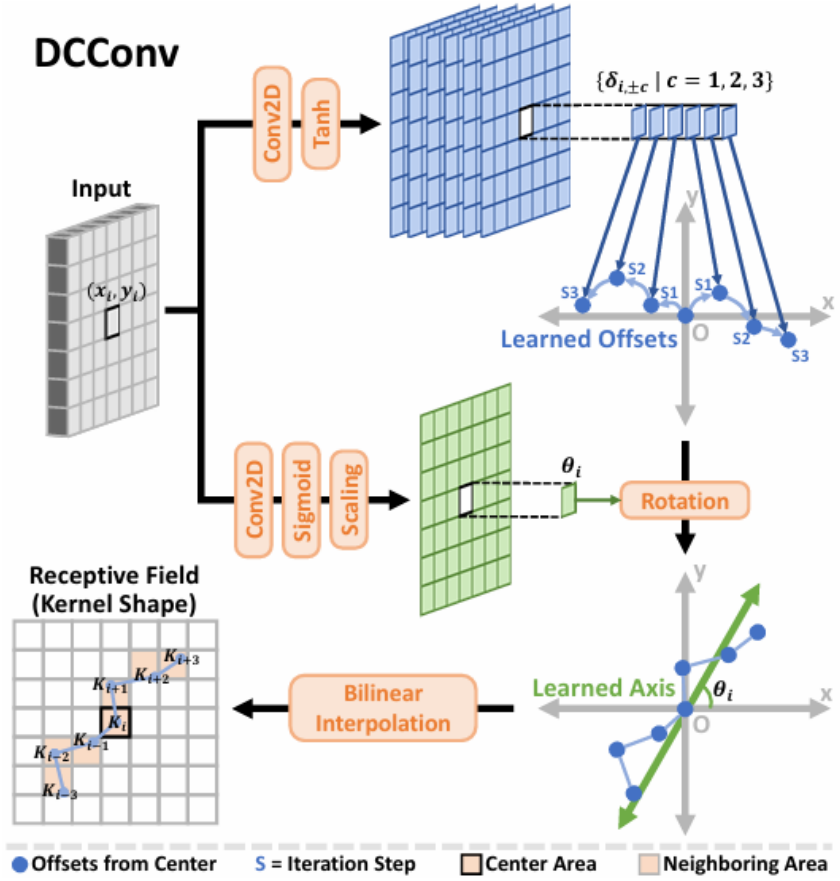
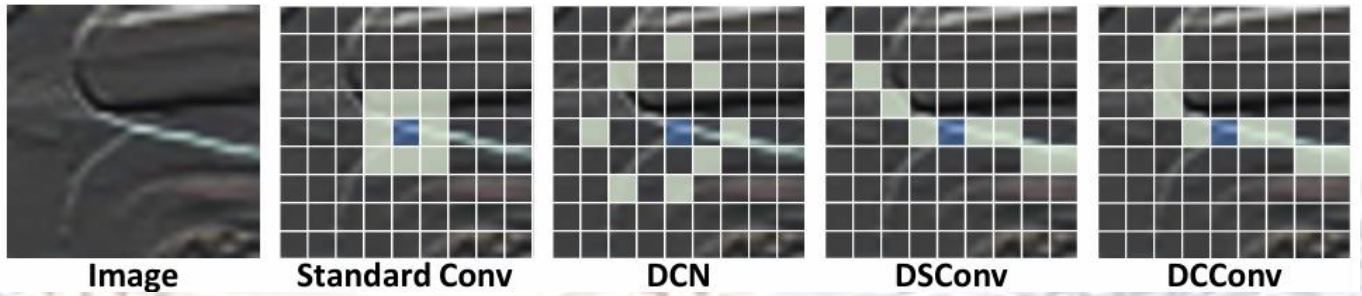


Figure 6: Schematic diagram of the proposed DCConv, initialized along the x-axis. Our method predicts both the offsets and axis orientations simultaneously, allowing for the effective representation of slender structures aligned in arbitrary directions. For illustration, the kernel length k is set to 7.

$$K_{i\pm c} = (x_i, y_i) + (\pm c, \sum_{j=0}^{c} \delta_{i,\pm j}) \cdot \begin{bmatrix} \cos\theta_i & \sin\theta_i \\ -\sin\theta_i & \cos\theta_i \end{bmatrix}$$

$$K_{i\pm c} = (x_i, y_i) + (\sum_{j=0}^{c} \delta_{i,\pm j}, \pm c) \cdot \begin{bmatrix} \cos\theta_i & \sin\theta_i \\ -\sin\theta_i & \cos\theta_i \end{bmatrix}.$$



WMamba: VMamba

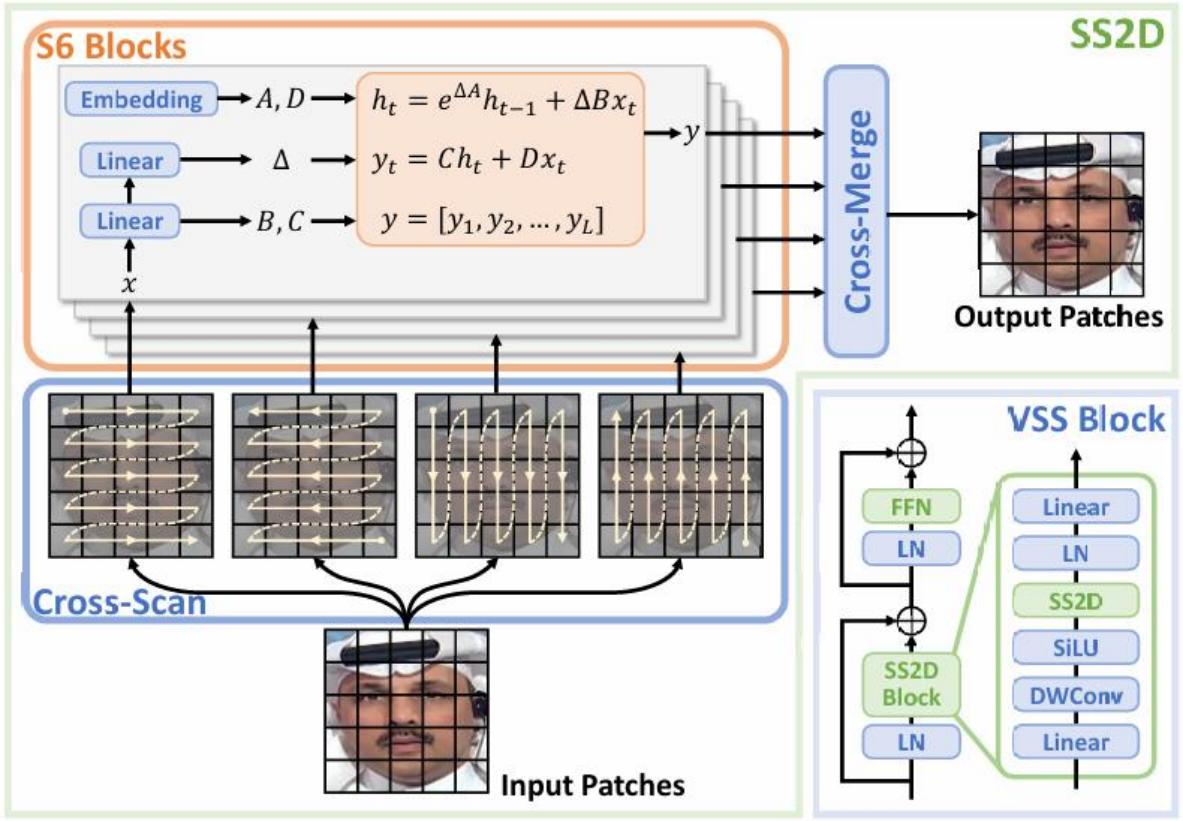
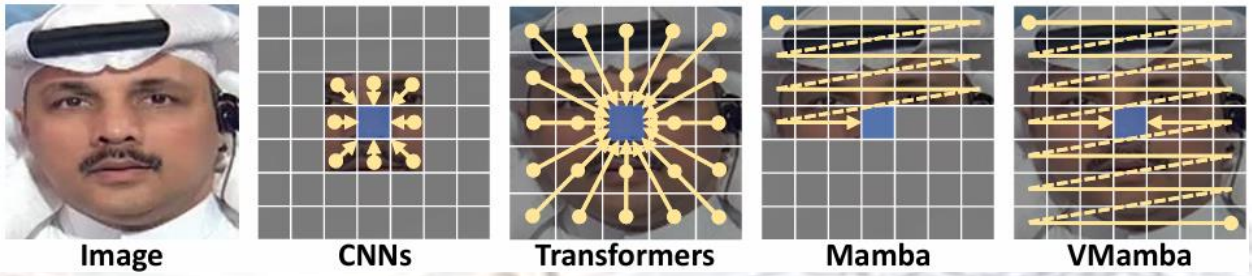


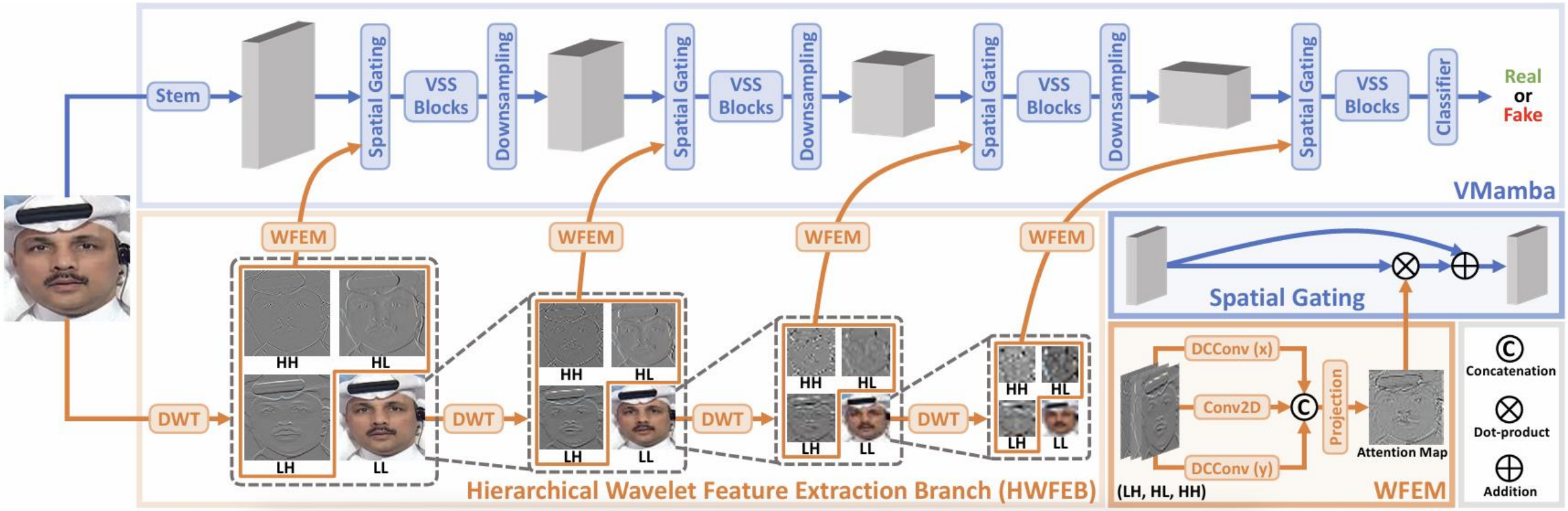
Figure 5: Schematic diagram of the VSS block, with the SS2D mechanism at its core. This mechanism flattens input image patches along four principle directions, facilitating comprehensive global perception. L denotes the number of patches.

$$h_t = \overline{A} h_{t-1} + \overline{B} x_t,$$
$$y_t = C h_t + D x_t.$$

$$\overline{A} = e^{\Delta A},$$
$$\overline{B} = (\Delta A)^{-1}(e^{\Delta A} - E) \cdot \Delta B \approx \Delta B.$$



WMamba: Architecture



WMamba: Results

Training set: FF++

Cross-Dataset

Method	Venue	Input Type	Training Set		Test Set AUC (%)			
			Real	Fake	CDF	DFDC	DFDCP	FFIW
F^3-Net* [43]	ECCV 2020	Frame	✓	✓	77.92	67.35	73.54	70.11
LTW* [49]	AAAI 2021	Frame	✓	✓	77.14	69.00	74.58	76.63
PCL+I2G [65]	ICCV 2021	Frame	✓		90.03	67.52	74.37	-
DCL [50]	AAAI 2022	Frame	✓	✓	82.30	-	76.71	71.14
SBI [46]	CVPR 2022	Frame	✓		93.18	72.42	86.15	84.83
F^2Trans [36]	TIFS 2023	Frame	✓	✓	89.87	-	76.15	-
SeeABLE [25]	ICCV 2023	Frame	✓		87.30	75.90	86.30	-
AUNet [1]	CVPR 2023	Frame	✓		92.77	73.82	86.16	81.45
LAA-Net [37]	CVPR 2024	Frame	✓		95.40	-	86.94	-
RAE [54]	ECCV 2024	Frame	✓		<u>95.50</u>	80.20	<u>89.50</u>	-
FreqBlender [67]	NeurIPS 2024	Frame	✓		94.59	74.59	87.56	<u>86.14</u>
UDD [10]	AAAI 2025	Frame	✓	✓	93.10	<u>81.20</u>	88.10	-
LESB [48]	WACVW 2025	Frame	✓		93.13	71.98	-	83.01
FTCN* [66]	ICCV 2021	Video	✓	✓	86.90	71.00	74.00	74.47
RealForensics [16]	CVPR 2022	Video	✓	✓	86.90	-	75.90	-
TALL [56]	ICCV 2023	Video	✓	✓	90.79	-	76.78	-
TALL++ [57]	IJCV 2024	Video	✓	✓	91.96	-	78.51	-
NACO [63]	ECCV 2024	Video	✓	✓	89.50	-	76.70	-
WMamba (Ours)	-	Frame	✓		**96.29**	**82.97**	**89.62**	**86.59**

Cross-Manipulation

Method	Test Set AUC (%)				
	DF	F2F	FS	NT	FF++
PCL+I2G [65]	**100**	98.97	99.86	97.63	99.11
SBI [46]	<u>99.99</u>	<u>99.88</u>	<u>99.91</u>	<u>98.79</u>	<u>99.64</u>
SeeABLE [25]	99.20	98.80	99.10	96.90	98.50
AUNet [1]	99.98	99.60	99.89	98.38	99.46
RAE [54]	99.60	99.10	99.20	97.60	98.90
WMamba	**100**	**99.98**	**99.94**	**98.88**	**99.70**

WMamba: Results

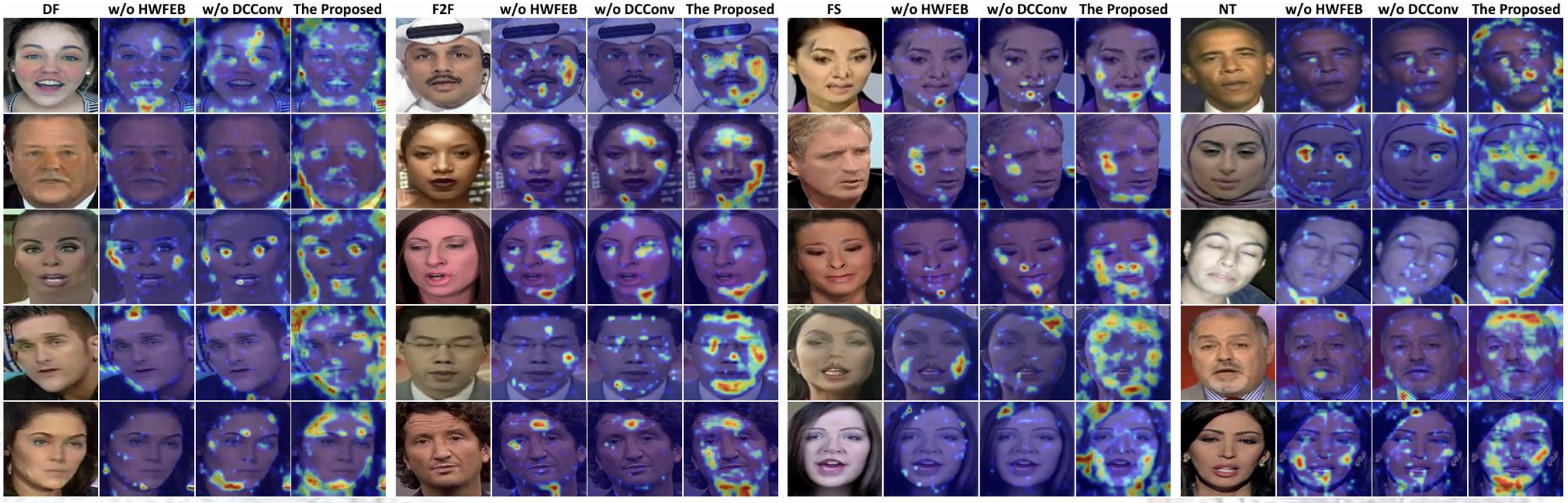


Figure 2: Saliency maps for a variety of fake samples from the FF++ dataset, where redder regions indicate areas of higher model attention. The HWFEB and DCConv effectively direct our model's attention toward critical facial contours.

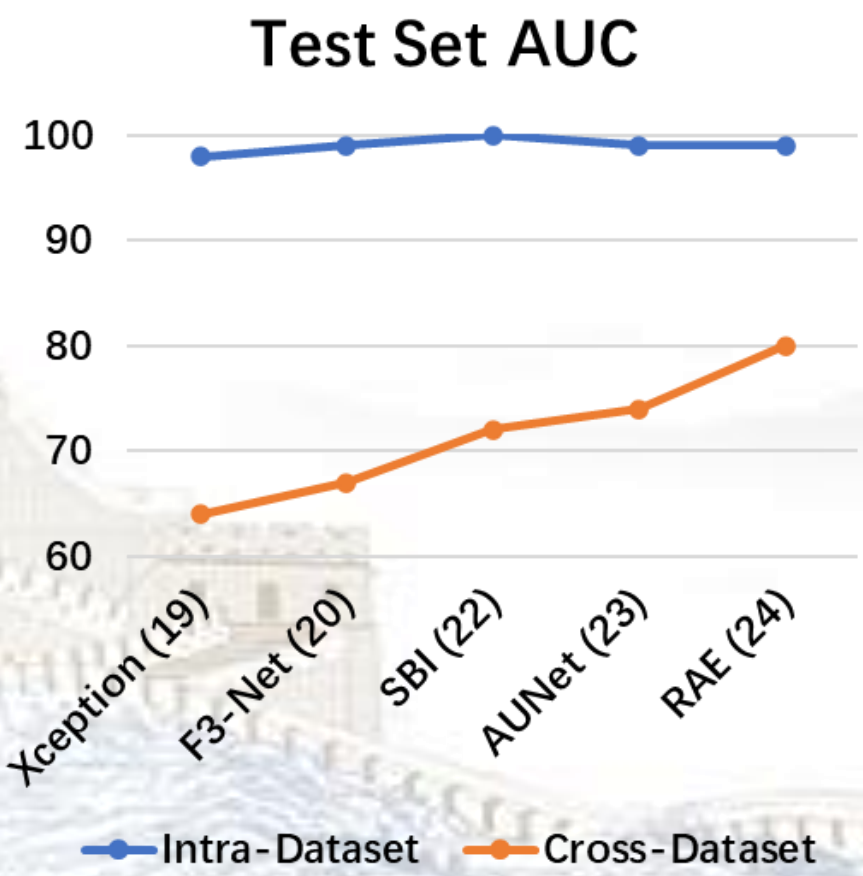
Contents

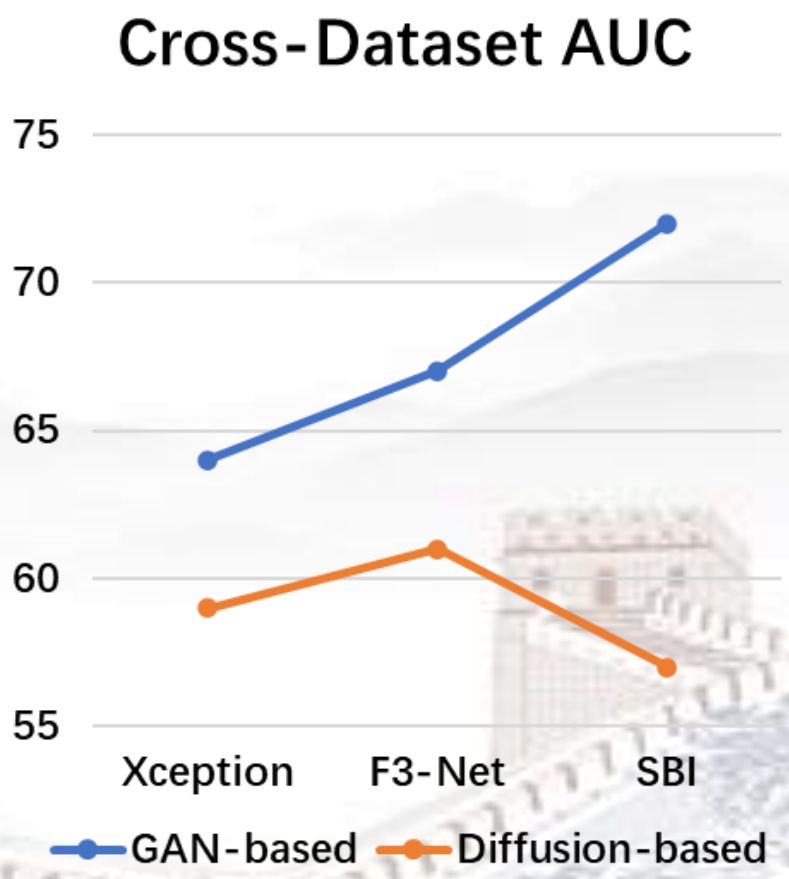
1 Background in Deepfake Detection

2 WMamba: Wavelet-based Mamba for Deepfake Detection

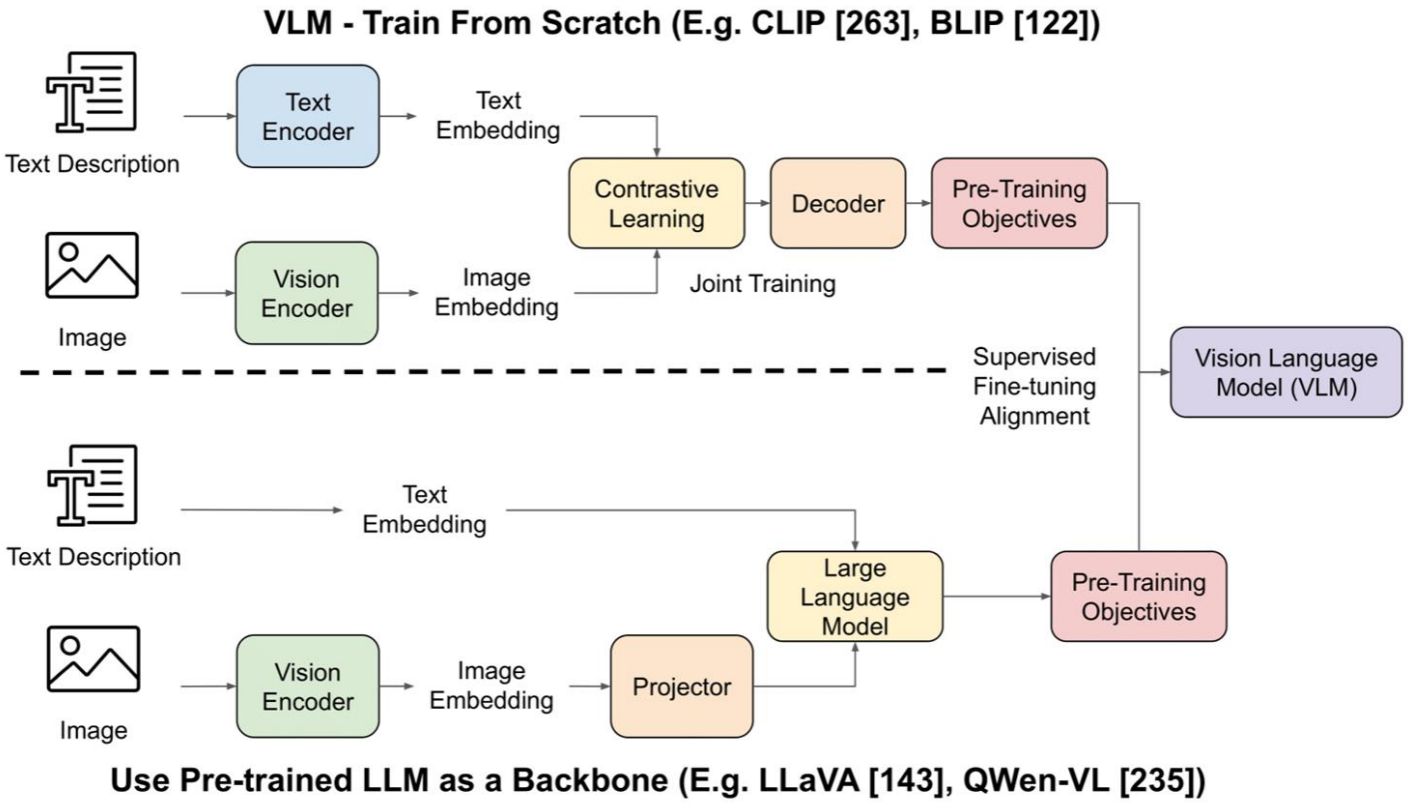
3 FDJL: Forgery-Discomfort Joint Learning for Generalized Deepfake Detection

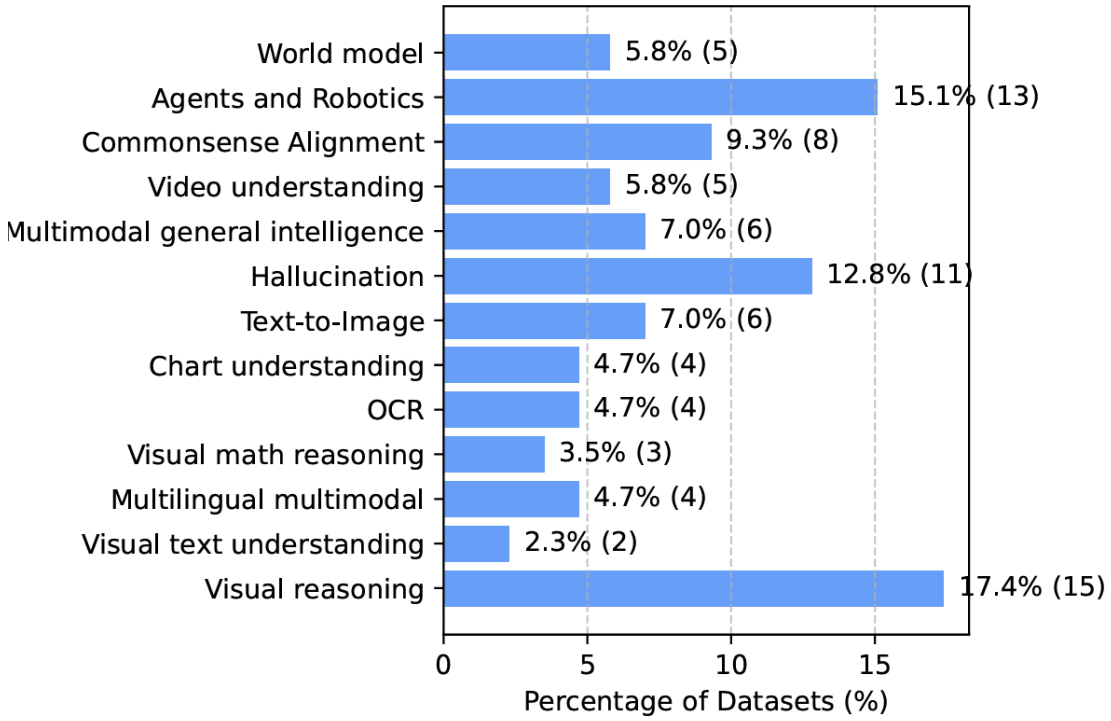
Background: Generalization



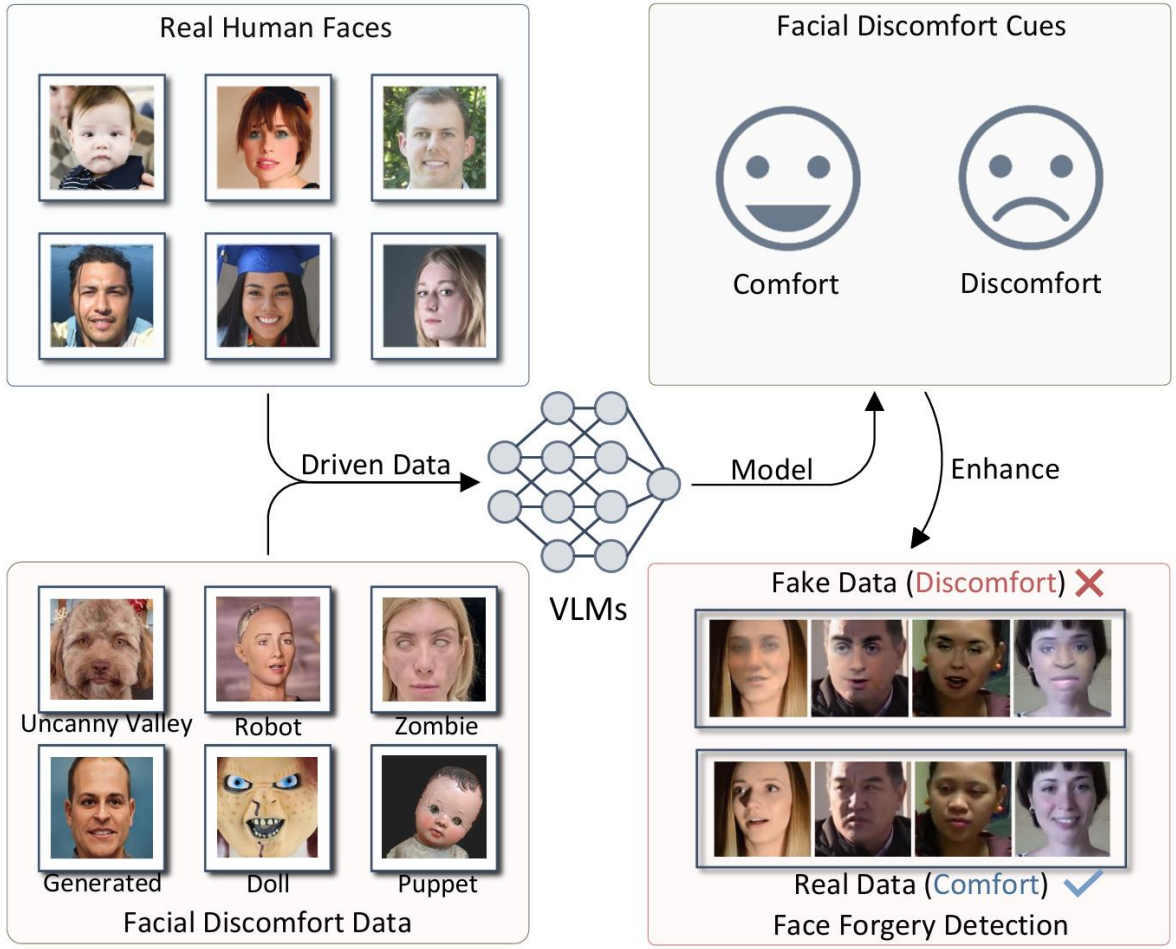


FDJL: Motivation

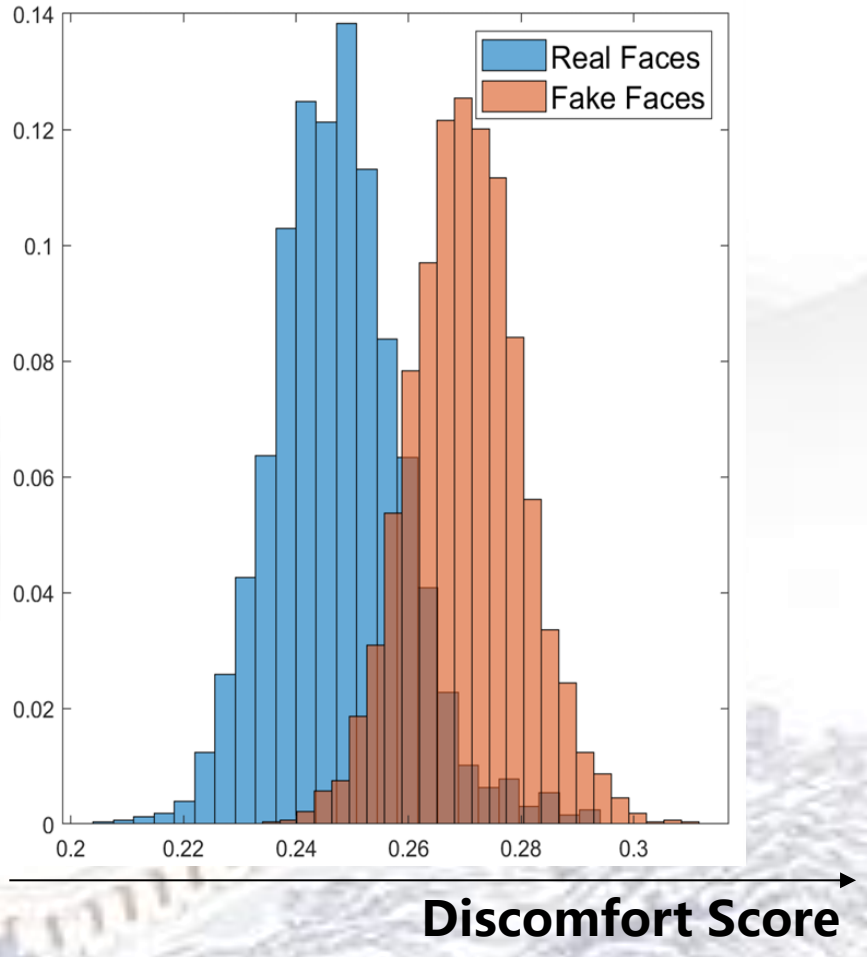




FDJL: Motivation

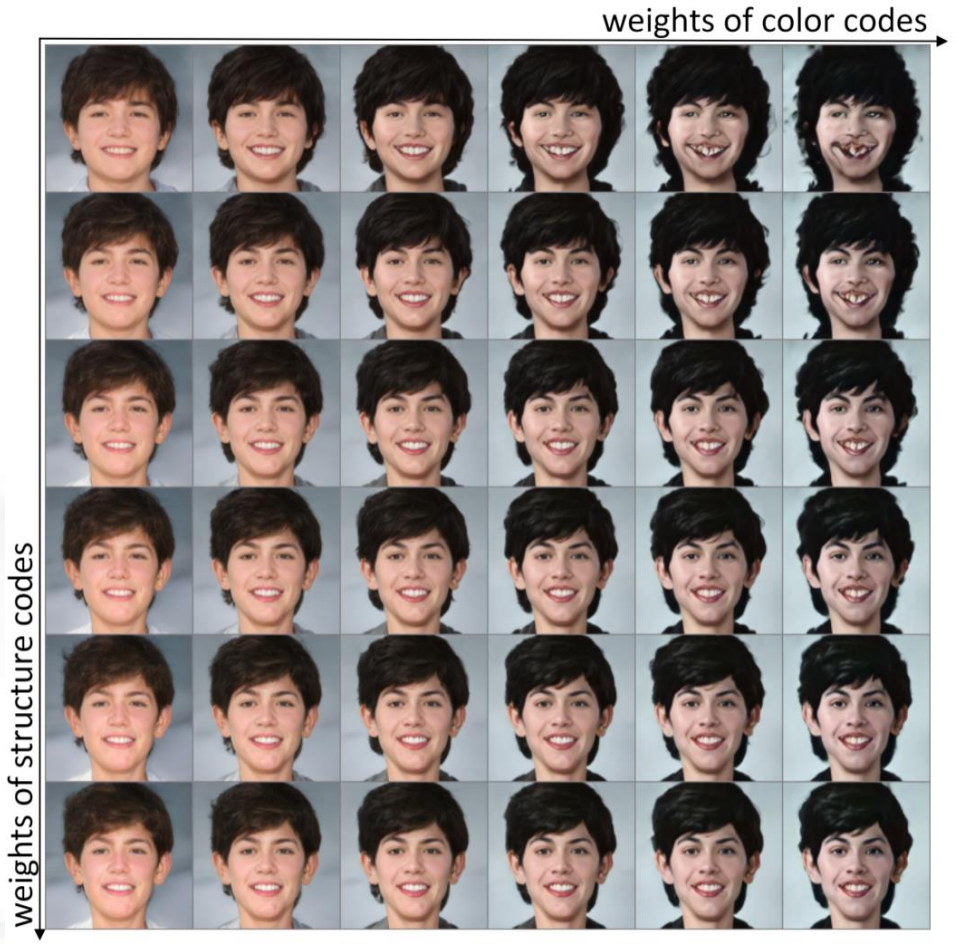


' A weird photo that may cause fear and discomfort '



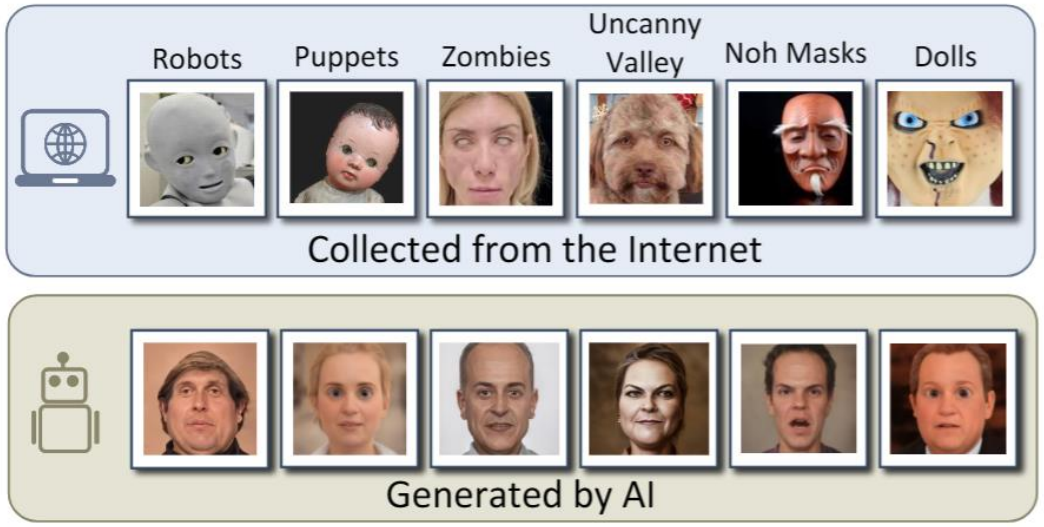
Facial Discomfort Dataset

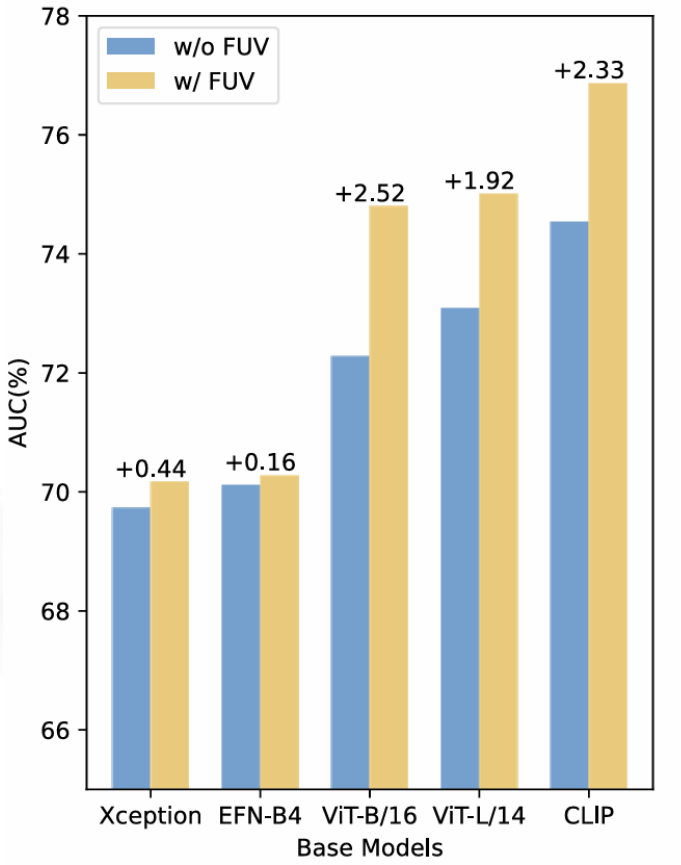
Collecting data from the Internet



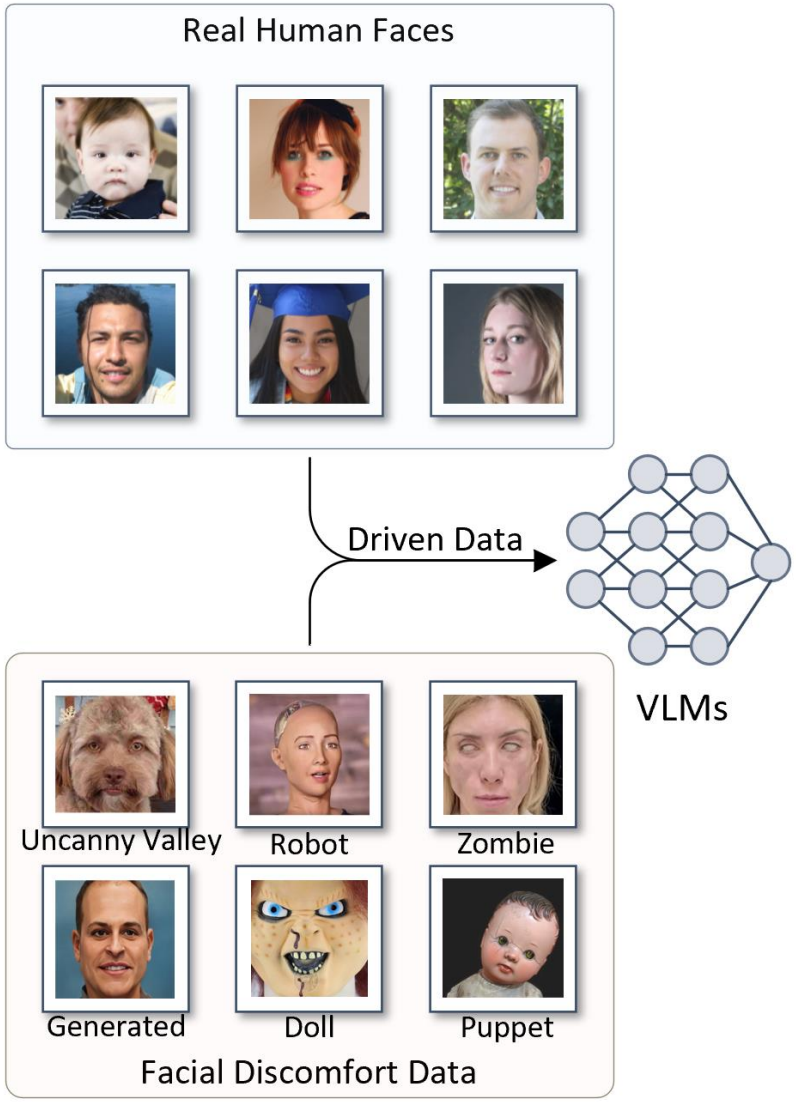
Generating data utilizing AIGC Models

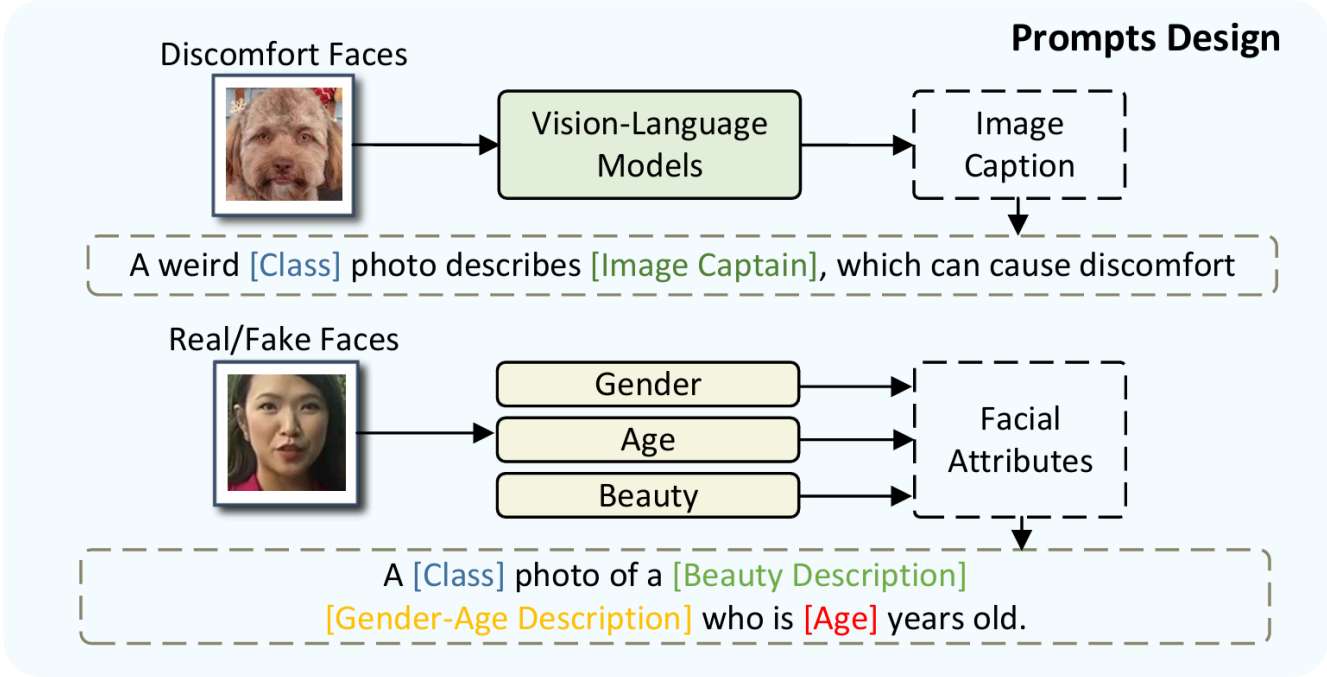
Facial Discomfort Dataset



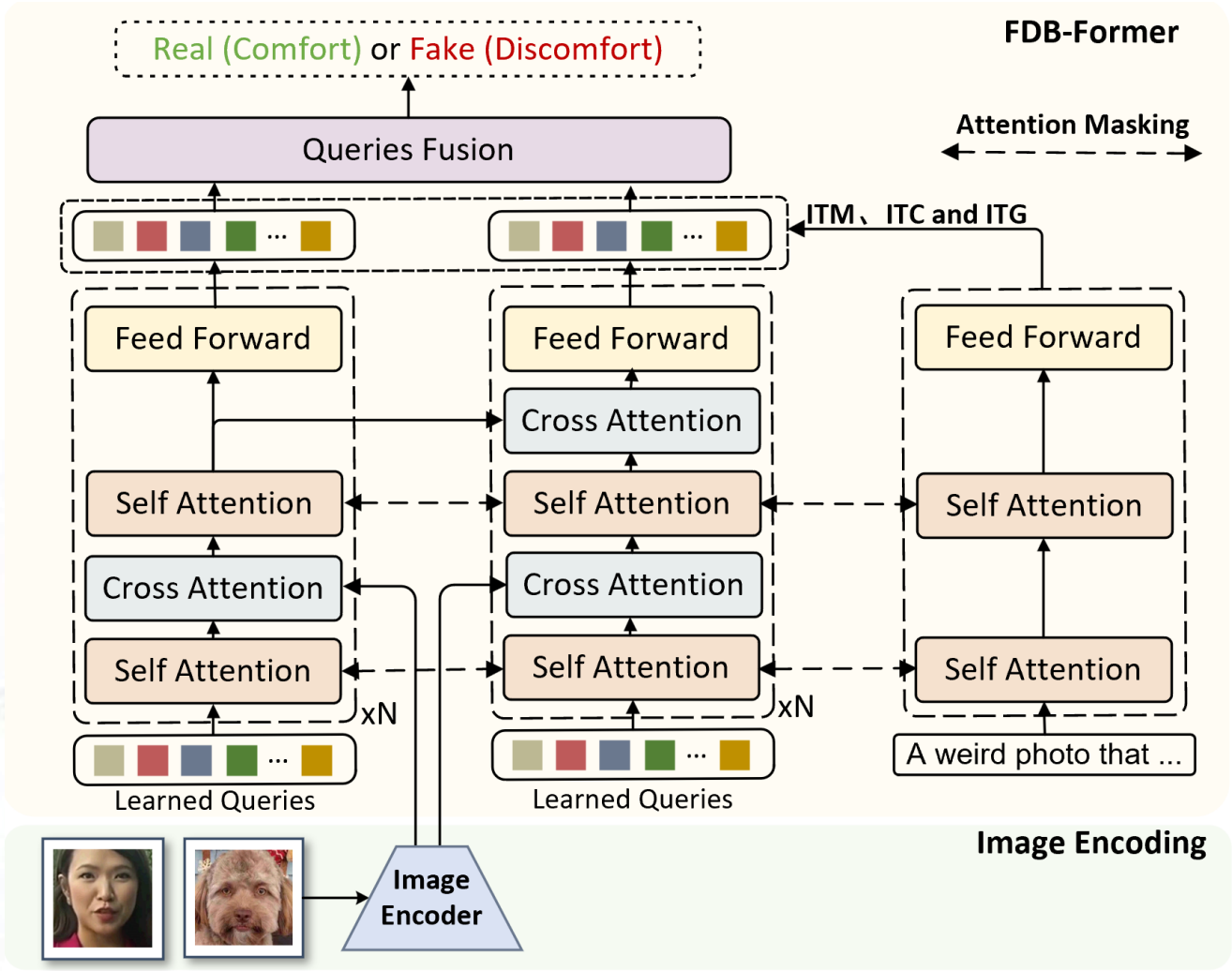


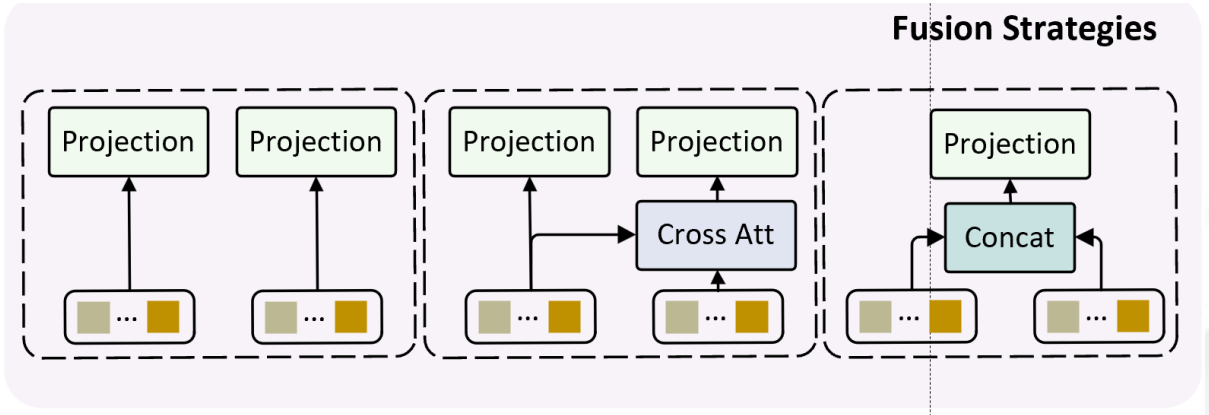
Prompts Design

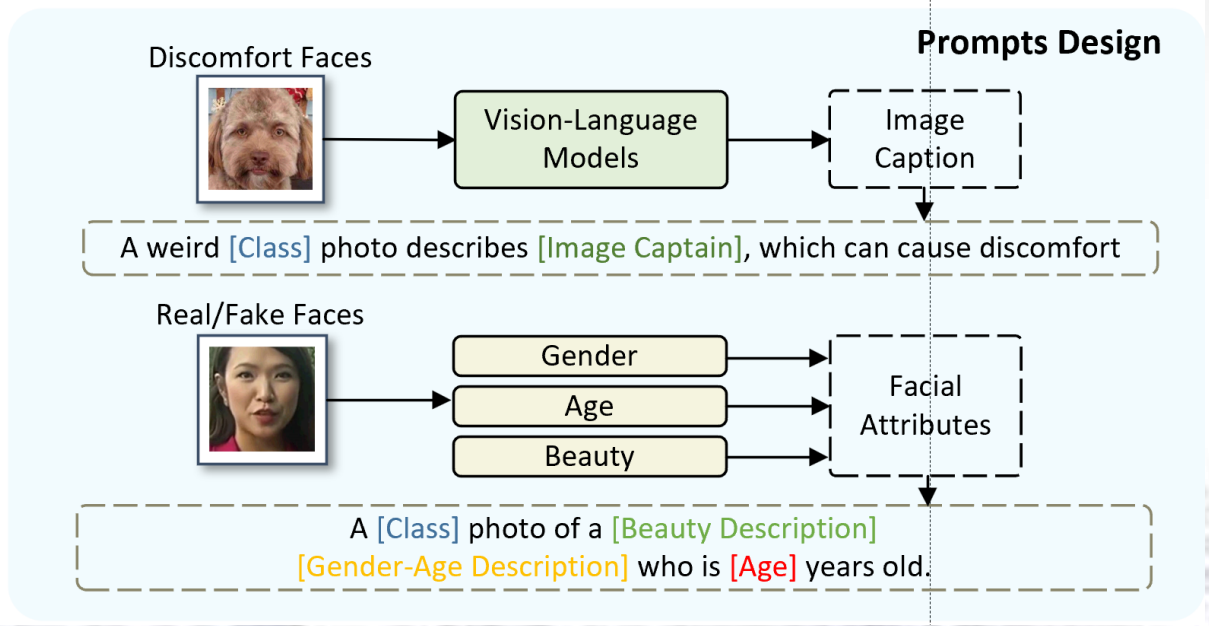




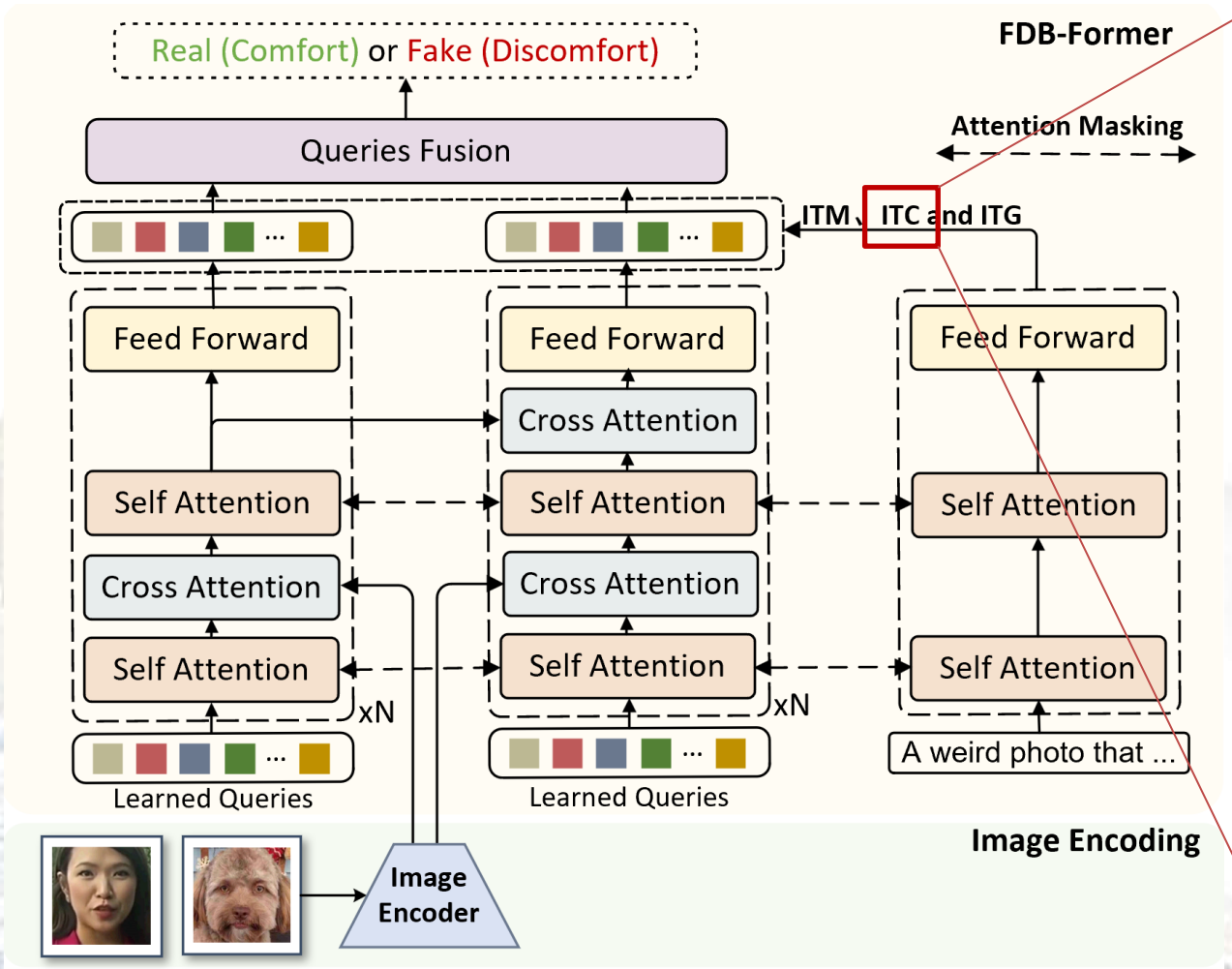
Forgery-Discomfort Joint Learning

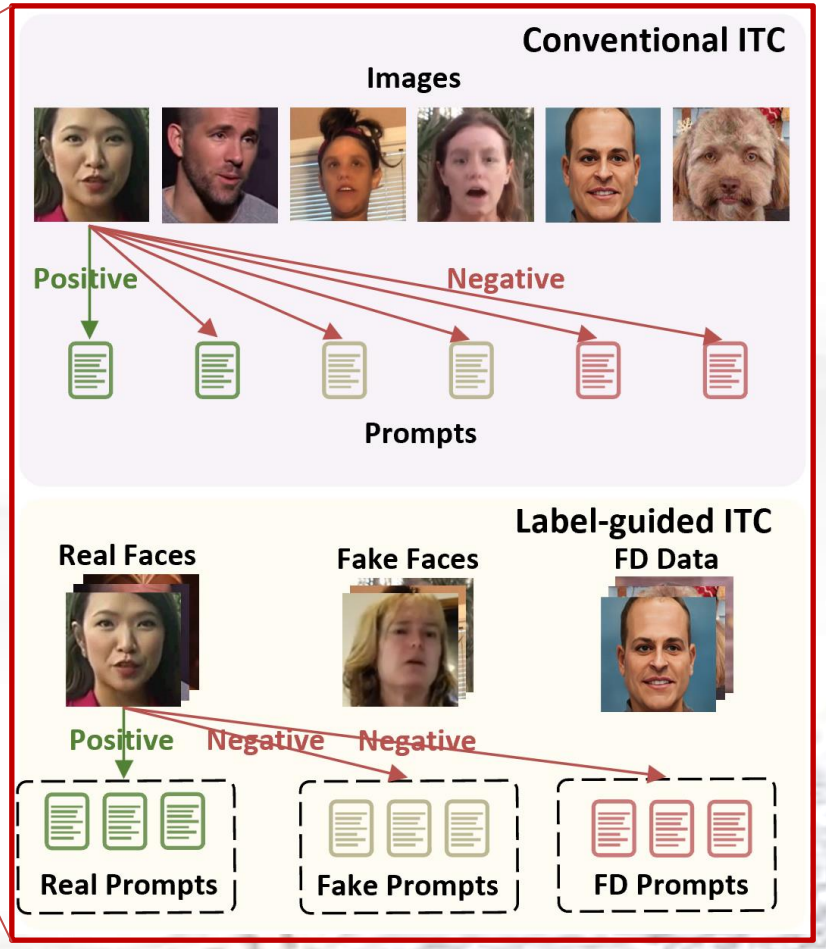




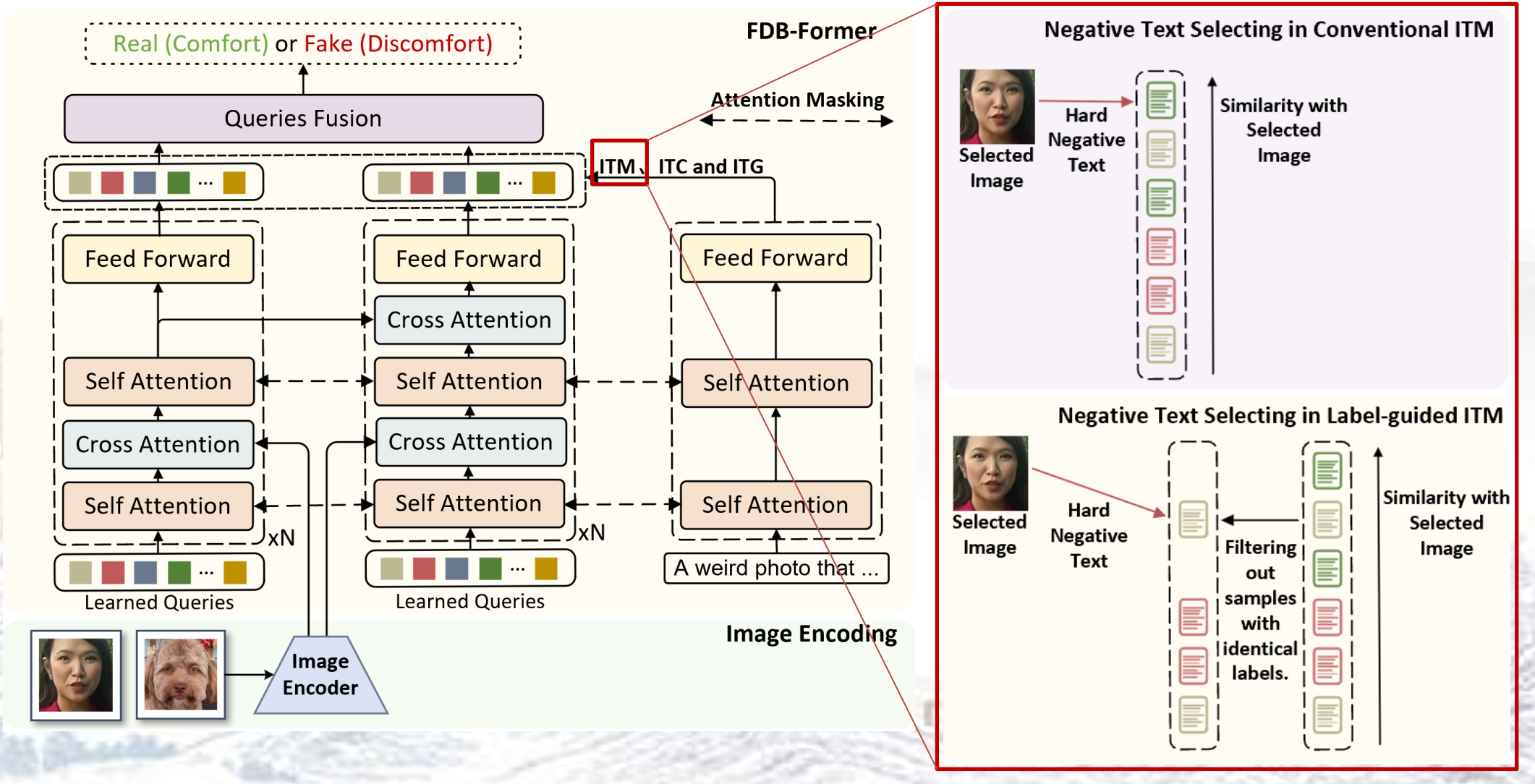


Forgery-Discomfort Joint Learning





Forgery-Discomfort Joint Learning



FDJL: Results

Table 1. Comparison experiments under the standard cross-dataset testing protocol. Methods marked with the symbol $*$ are reproduced using the official codes. The best performance is highlighted in **bold**, and the second best is underlined. The testing results show that the proposed FDJL framework achieves state-of-the-art results among all the compared methods, including methods that take frames and videos as input.

Method	Input Type	Training Set		Test Set AUC (%)			
		Real	Fake	Celeb-DF	DFDC	DFDCP	FFIW
Xception* [3]	Frame	✓	✓	65.27	69.73	73.74	59.45
Face X-ray+BI [17]	Frame	✓	✓	79.50	65.50	80.92	-
F^3-Net* [23]	Frame	✓	✓	77.92	67.35	73.54	70.11
LTW* [28]	Frame	✓	✓	77.14	69.00	74.58	76.63
PCL+I2G [6]	Frame	✓		90.03	67.52	74.37	-
CORE [22]	Frame	✓		79.45	72.83	75.74	75.19
DCL [29]	Frame	✓	✓	82.30	76.71	-	-
SBI [27]	Frame	✓		93.18	72.42	86.15	<u>84.83</u>
F^2Trans [19]	Frame	✓	✓	89.87	76.15	-	-
SeeABLE [15]	Frame	✓		87.30	75.90	86.30	-
AUNet [1]	Frame	✓		92.77	73.82	86.16	81.45
MoE-FFD [14]	Frame	✓	✓	91.28	-	84.97	-
UCF [37]	Frame	✓	✓	82.40	80.50	-	-
LAANet [21]	Frame	✓		95.40	-	86.94	-
RAE [33]	Frame	✓		<u>95.50</u>	<u>80.20</u>	<u>89.50</u>	-
FTCN* [40]	Video	✓	✓	86.90	71.00	74.00	74.47
RealForensics [8]	Video	✓	✓	86.90	75.90	-	-
TALL [35]	Video	✓	✓	90.79	76.78	-	-
TALL++ [36]	Video	✓	✓	91.96	78.51	-	-
NACO [38]	Video	✓	✓	89.50	76.70	-	-
Ours(FDJL+FD dataset)	Frame	✓		**95.62**	**81.91**	**89.58**	**86.74**

FDJL: Results

Table 2. Comparative experiments when facing unknown forgery methods. All methods are reproduced using the official codes. The best performance is highlighted in **bold**, and the second best is underlined. Our FJDL achieves state-of-the-art results among all the compared methods, indicating that the proposed method performs better generalization for unknown forgery types.

Method	AUC (%)					
	2D Face Morphing				3D Face Morphing	Face Generation
	OpenCV	Face Morpher	Web Morph	AMSL	Shape Transfer	StyleGAN2
MOCO [9]	57.80	54.44	50.51	52.72	49.91	43.25
CLIP [24]	67.56	72.64	68.04	73.33	63.25	70.71
OC-FakeDect [11]	69.74	70.57	63.81	66.35	60.37	56.61
Xception [3]	63.03	57.39	55.76	51.13	51.84	53.23
EfficientNetB4 [2]	59.23	59.34	52.69	54.91	52.69	55.08
LTW [28]	74.67	63.00	62.81	61.09	57.86	61.10
F^3-Net [23]	71.99	67.19	65.77	64.02	55.30	63.48
CORE [22]	74.89	69.11	70.47	68.87	64.57	67.14
STDD [39]	73.15	66.95	67.56	65.63	61.59	65.35
FaRL [41]	75.76	74.20	69.84	67.43	67.69	68.88
FakeOut [13]	77.31	71.52	66.04	68.93	64.46	68.13
DCL [29]	79.65	74.91	71.39	70.26	71.61	70.84
SBI [27]	81.72	78.65	72.50	71.69	70.12	66.97
Ours(FDJL+FD dataset)	**87.25**	**81.86**	**75.56**	**74.92**	**74.73**	**79.83**

FDJL: Results

Train	Method	DF	F2F	FS	NT
DF	EN-b4	*99.97*	76.32	46.24	72.72
	DCL	*99.98*	77.13	61.01	75.01
F2F	EN-b4	84.52	*99.20*	58.14	63.71
	DCL	91.91	*99.21*	59.58	66.67
FS	EN-b4	69.25	67.69	*99.89*	48.61
	DCL	74.80	69.75	*99.90*	52.60
NT	EN-b4	85.99	48.86	73.05	*98.25*
	DCL	91.23	52.13	79.31	*98.97*
BI	Face X-ray	99.17	98.57	98.21	98.13
SBIS	SBI	**99.99**	99.88	99.91	98.79
SBIs+FD	Ours	99.98	**99.97**	**99.93**	**99.04**

FDJL: Ablation Study

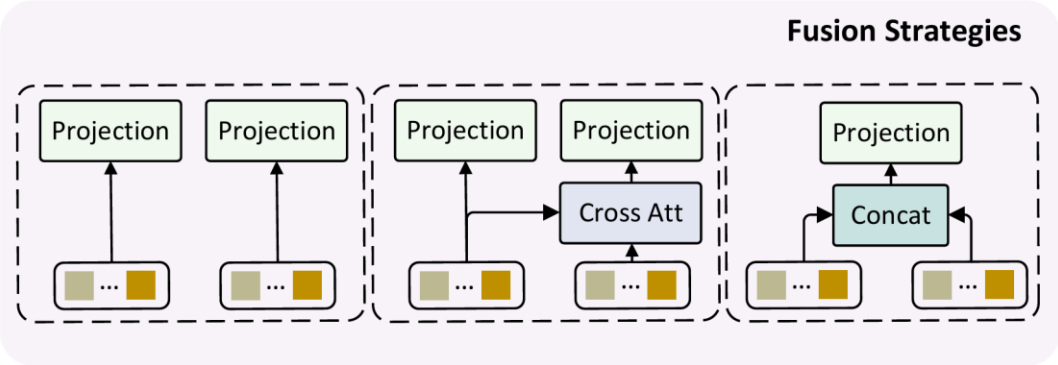


Table 4. Ablation study of the Facial Discomfort dataset.

Collected	Generated	AUC(%)			
		Celeb-DF	DFDC	DFDCP	FFIW
		91.93	74.28	79.72	77.46
✓		94.47	78.17	87.40	82.50
	✓	93.82	76.70	83.05	82.98
✓	✓	**95.62**	**81.91**	**89.58**	**86.74**

Table 6. Ablation study on fusion strategies.

Fusion Strategy	AUC(%)			
	Celeb-DF	DFDC	DFDCP	FFIW
Concat	87.36	74.69	76.58	69.89
Cross-Att	94.60	81.41	88.66	85.51
Split	**95.62**	**81.91**	**89.58**	**86.74**

Table 5. Ablation study on the different losses.

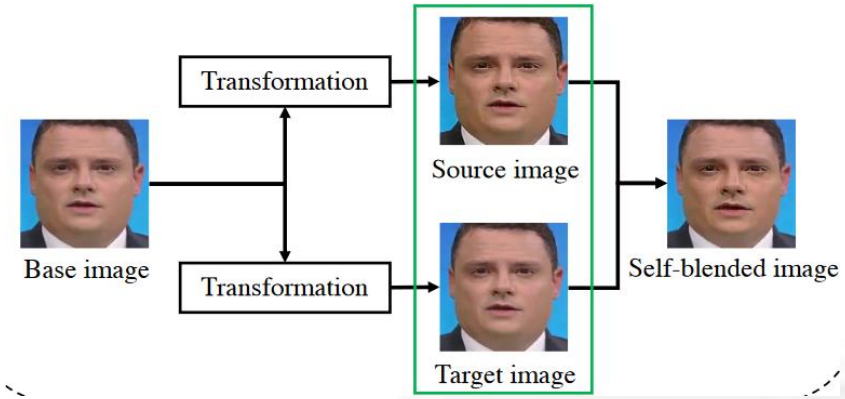
$\mathcal{L}_{CLS}$	$\mathcal{L}_{ITA}$	$\mathcal{L}_{FD}$	AUC(%)			
			Celeb-DF	DFDC	DFDCP	FFIW
✓			87.50	69.73	75.90	73.02
✓	✓		92.11	75.59	76.17	80.85
✓		✓	91.93	73.75	78.14	82.19
✓	✓	✓	**95.62**	**81.91**	**89.58**	**86.74**

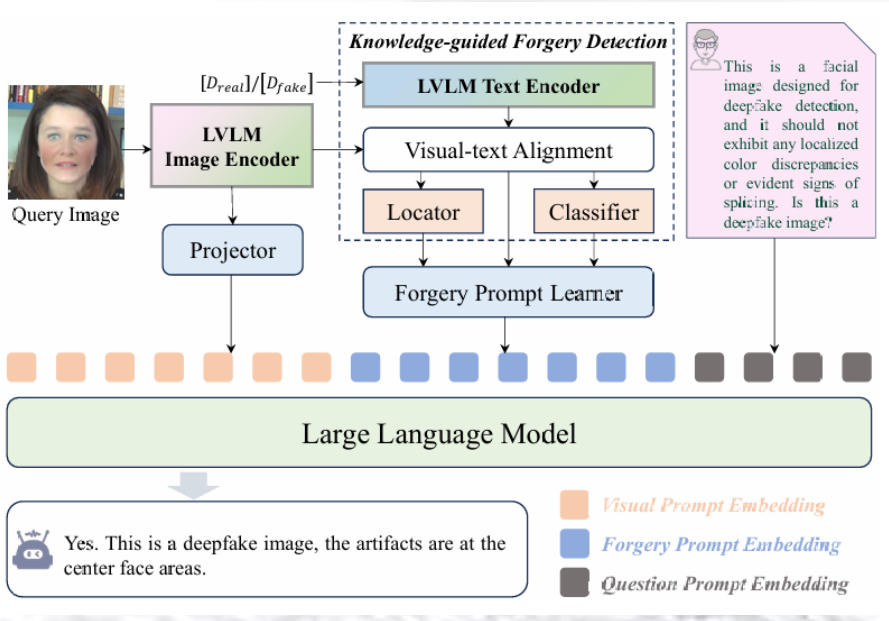
Take home message

- Robust and discriminant features for deepfake detection is necessary. Combining different types for features (e.g., semantic and detailed features) is a possible way .

- Fundamental (essential) features rather than database biased features is the key to realize high generalization performance.

- Multi-modal large model is a promising direction. Not only output the classification results, but also with explainable texts.





Thank You!