Guard Me If You Know Me: Protecting Specific Face-Identity from Deepfakes

Kaiqing Lin, Zhiyuan Yan*, Ke-Yue Zhang, Li Hao, Yue Zhou, Yuzhen Lin, Weixiang Li, Taiping Yao†, Shouhong Ding, Bin Li†*

NEURAL INFORMATION PROCESSING SYSTEMS

PERSONALIZED DEEPFAKE DETECTION

What is the deepfake detection?

- Deepfake detection identifies whether a facial image or video has been manipulated or generated.

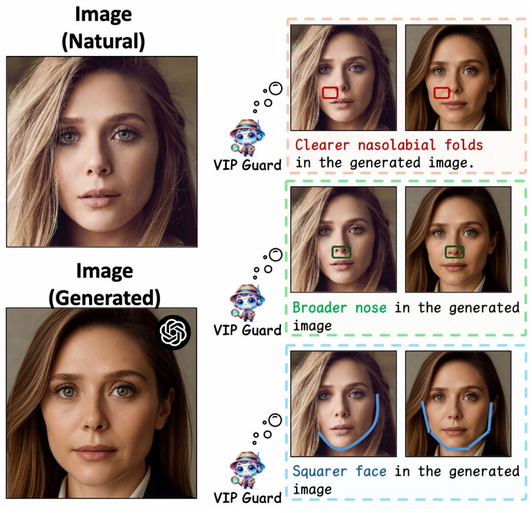
What is the personalized deepfake detection?

- Personalized deepfake detection protects specific individuals by using their known facial priors to detect identity-targeted manipulations.
- In this setting, only a few authentic images of a specific individual are available.

OUR MOTIVATION

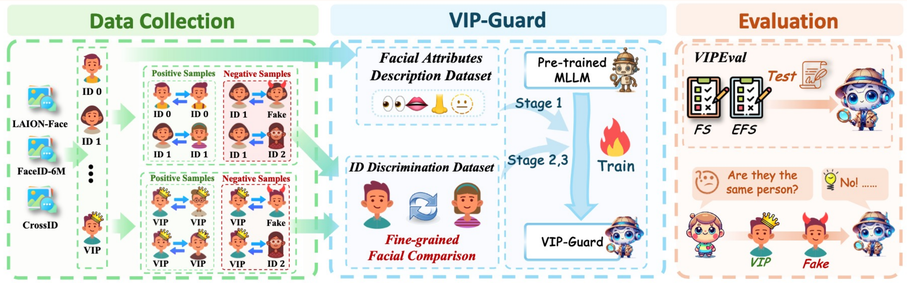
Current deepfake detectors struggle to protect specific individuals due to two critical oversights:

- **Ignorance of Identity Priors:** General detectors treat faces generically, overlooking the **valuable authentic reference data available for target individuals** (VIPs).
- **Neglect of Fine-Grained Details:** Existing methods rely on coarse global features, **failing to spot the subtle local inconsistencies** (e.g., eye bags) present in high-fidelity forgeries.



CONTRIBUTION OF OUR PAPER

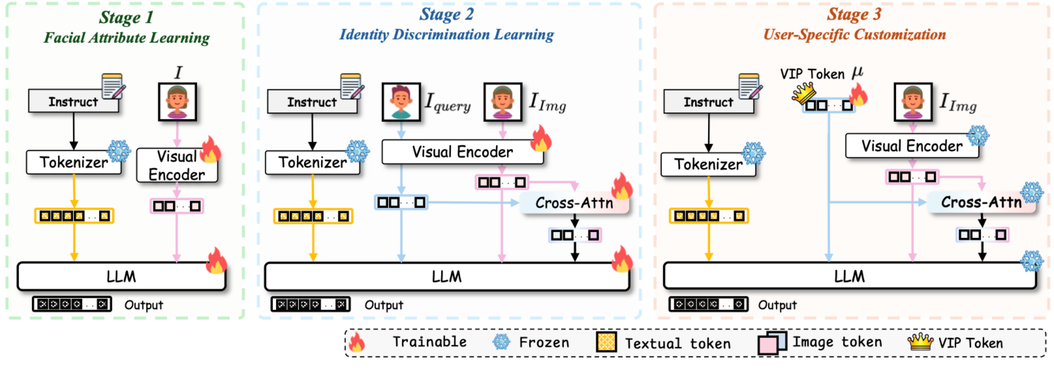
- ✓ **New Formulation:** We reformulate personalized deepfake detection as **a fine-grained face recognition task** leveraging both global identity cues and detailed facial attributes.
- ✓ **New Framework:** We present **VIP-Guard**, an identity-aware detection framework that combines pretrained facial priors with MLLMs to provide personalized deepfake detection and explanation through identity features.
- ✓ **New Benchmark:** We introduce **VIPBench**, the first comprehensive benchmark for identity-aware deepfake detection, featuring 22 real identities and 80,080 images from 14 manipulation methods for fine-grained, realistic evaluation.



METHODOLOGY OF VIP-GUARD

VIPGuard is a three-stage framework trained using contrastive learning on facial image pairs.

- ☐ **Stage 1:** Understand Facial Details. Fine-tunes the MLLM on a facial attribute dataset to **master fine-grained facial semantics** (e.g., eye shape, skin texture).
- ☐ **Stage 2:** Identity Discrimination Learning (Compare Arbitrary Faces). Trains the model to reason about and distinguish **discrepancies between any pair of faces**.
- ☐ **Stage 3:** User-Specific Customization (Personalized Protection). Freezes the main model and learns a lightweight **"VIP token" for a specific target identity** to enable personalized forgery detection.



EVALUATION FOR DETECTION

VIP-Guard achieves SOTA performance across all settings.

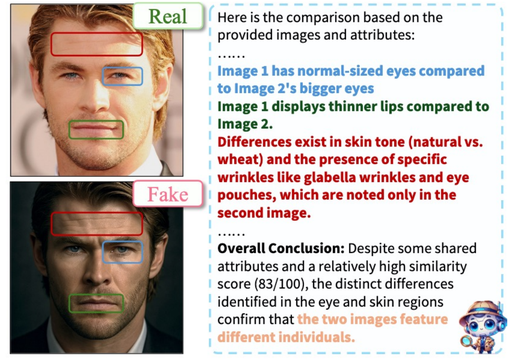
Performance Evaluation of Face-Swapping Detection

Methods	BlendFace [52]	Ghost3 [17]	HifiFace [60]	InSwap [27]	MobileSwap [64]	SimSwap [5]	UniFace [62]
Xception [8]	53.89 / 46.59	61.08 / 42.27	71.70 / 34.09	64.79 / 38.64	98.12 / 7.50	64.91 / 38.41	59.91 / 42.50
EfficientNet [53]	33.94 / 60.91	43.52 / 54.32	61.22 / 42.95	37.34 / 58.64	89.67 / 64.45	56.55 / 45.00	46.53 / 50.23
UCF [69]	64.31 / 38.18	65.92 / 35.68	65.62 / 36.82	66.02 / 34.32	87.33 / 20.00	63.11 / 36.82	67.25 / 34.32
ProDet [7]	59.84 / 42.95	53.18 / 47.73	56.48 / 46.36	35.91 / 59.32	73.01 / 34.32	56.27 / 45.68	49.34 / 50.00
RECCE [3]	56.11 / 45.00	61.53 / 39.54	56.93 / 45.91	58.85 / 43.41	88.40 / 19.32	60.05 / 42.95	63.01 / 39.10
CDFA [39]	59.86 / 45.23	70.50 / 36.82	85.43 / 24.09	70.92 / 36.59	98.75 / 5.00	71.22 / 35.91	73.30 / 34.77
RepDFD [38]	70.34 / 35.00	78.76 / 28.18	80.09 / 27.50	71.66 / 34.32	95.76 / 10.45	79.40 / 27.95	82.36 / 25.45
Effort [66]	91.87 / 17.61	95.28 / 13.07	96.60 / 10.23	85.53 / 23.86	96.23 / 13.07	97.16 / 8.52	92.48 / 13.64
DiffID [73]	83.33 / 24.47	66.02 / 38.38	87.23 / 20.33	66.22 / 38.06	75.71 / 30.89	72.03 / 33.91	83.30 / 24.40
ICT-Ref [15]	88.67 / 13.37	86.52 / 15.51	85.90 / 14.05	84.34 / 17.87	87.45 / 13.52	86.57 / 14.65	82.73 / 19.04
VIPGuard	**99.48 / 0.81**	**97.97 / 4.39**	**99.63 / 0.61**	**96.40 / 8.24**	**99.55 / 1.01**	**99.43 / 1.02**	**99.69 / 0.26**

Performance Evaluation of Entire Face Synthesis Detection

Method	Open-Source				Commercial-API		
	ConsistentID [25]	Arc2Face [44]	PuLID [22]	GPT-4o [43]	Jimeng AI [29]	TongYi [55]	Kling AI [31]
Xception	42.02 / 54.77	51.87 / 48.86	59.23 / 44.09	58.13 / 46.36	57.77 / 44.55	34.36 / 62.05	44.34 / 54.32
EfficientNet	33.81 / 61.14	44.96 / 50.91	47.19 / 50.23	75.35 / 28.86	55.04 / 45.00	41.15 / 55.68	42.40 / 54.32
UCF [69]	62.16 / 40.91	56.62 / 46.36	54.16 / 46.36	59.06 / 42.73	71.08 / 32.78	82.38 / 22.73	63.31 / 40.23
ProDet [7]	63.68 / 37.95	59.62 / 40.91	67.02 / 36.82	59.23 / 40.91	59.71 / 42.73	89.80 / 18.41	72.53 / 32.73
RECCE [3]	68.56 / 36.82	57.79 / 47.72	63.85 / 40.00	83.63 / 22.73	69.00 / 35.23	97.00 / 7.50	70.94 / 37.72
CDFA [39]	77.62 / 30.00	67.09 / 39.09	67.93 / 37.50	73.46 / 32.73	71.98 / 34.09	90.32 / 17.95	77.47 / 30.00
RepDFD [38]	83.52 / 24.32	61.67 / 41.59	74.65 / 32.27	73.62 / 32.27	62.78 / 40.91	93.80 / 14.09	60.10 / 43.41
Effort [66]	58.68 / 47.35	57.03 / 44.89	56.31 / 44.89	49.63 / 48.30	63.60 / 40.34	82.93 / 23.86	56.84 / 44.89
DiffID [73]	75.85 / 32.01	78.47 / 28.86	70.36 / 35.72	45.26 / 52.97	64.29 / 39.35	84.66 / 23.58	69.51 / 35.56
ICT-Ref [15]	63.15 / 39.84	70.27 / 32.84	72.36 / 31.93	58.59 / 40.93	65.36 / 35.73	74.88 / 24.41	50.05 / 45.98
VIPGuard	**99.69 / 0.45**	**98.05 / 4.80**	**98.96 / 1.90**	**89.03 / 16.14**	**97.04 / 5.36**	**99.76 / 0.23**	**99.27 / 1.25**

VISUALIZATION EXAMPLE



ACKNOWLEDGEMENT

This work was supported by the following foundations:

- NSFC (Grant U23B2022, U22B2047, 62202310, 62572328)
- Shenzhen R&D Program (Grant JCYJ20250604181211O1, SYSPG20241211174032004)
- Guangdong Basic and Applied Basic Research Foundation (Grant 2025A1515010292)

PROJECT

Paper

Dataset

GitHub