Deep Residual Injection for Full-Spectrum Forensic Signal Perception in Multimodal Large Language Models

Kaiqing Lin, Zhiyuan Yan, Ruoxin Chen, Ke-Yue Zhang, Yue Zhou, Caiyong Piao, Bin Li†, Taiping Yao, Bo Wang, Youchang Xiao, Shouhong Ding

Background

What is MLLM-based AIGI detection?

- Detect real vs. AI-generated images with MLLM-based semantic reasoning and forensic explanations.

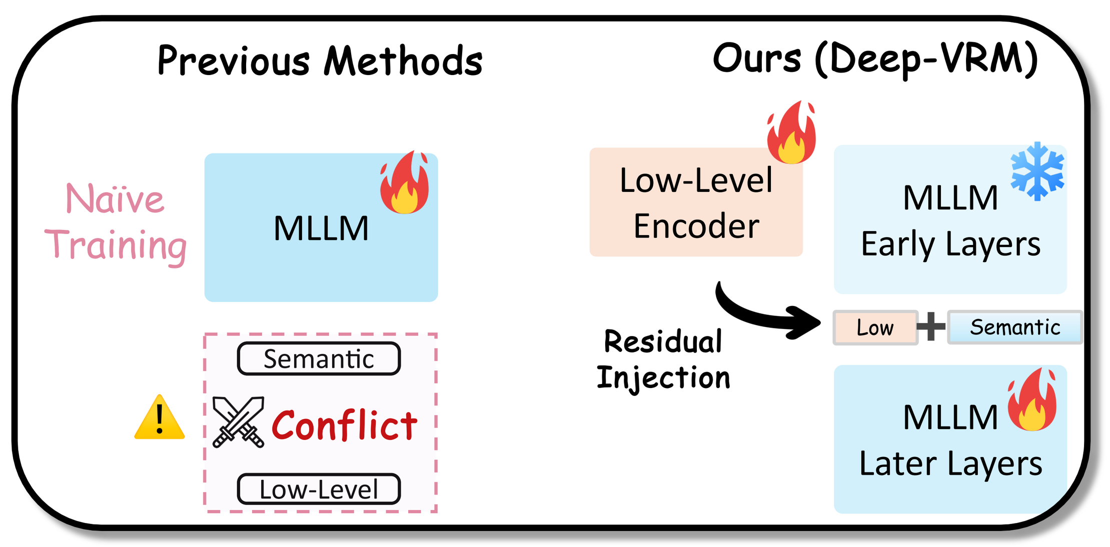
What is full-spectrum forensic perception?

- Capture both high-level semantic anomalies and low-level generator traces.

OUR MOTIVATION

MLLM-based detectors face two limitations:

- Learning low-level artifacts conflicts with preserving pre-trained semantic knowledge.
- Existing methods mix semantic and low-level cues in one path, ignoring layer-wise roles.



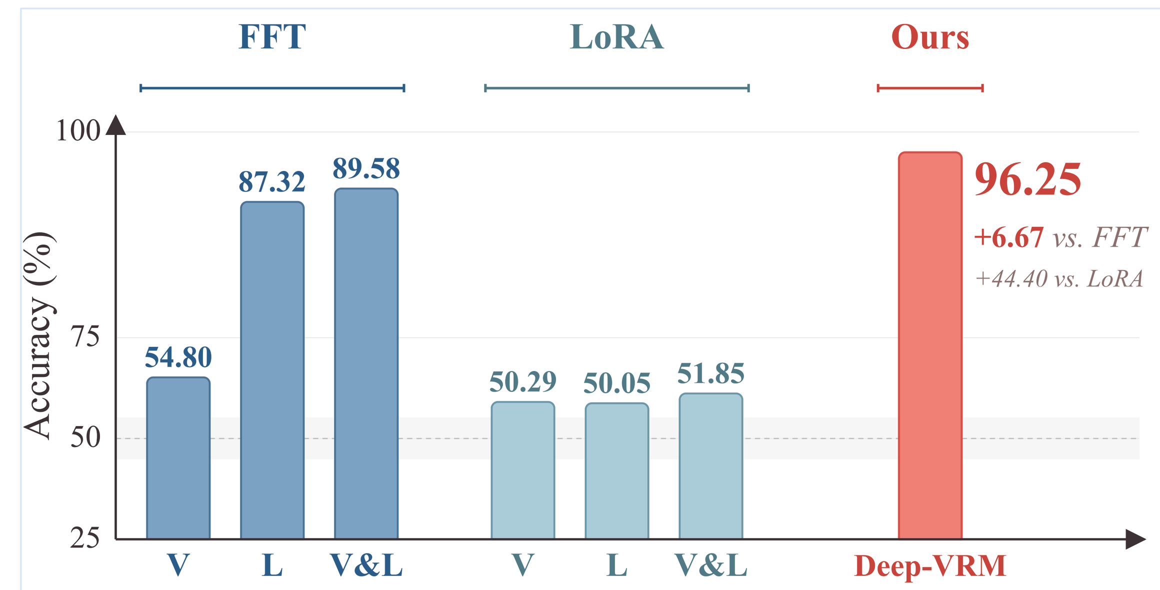
OUR CONTRIBUTION

- ✓ **New Analysis:** We **reveal a representation conflict** between semantic reasoning and low-level artifact perception in MLLMs.
- ✓ **New Architecture:** We propose Deep-VRM, which **preserves early semantics and injects artifact cues** into intermediate LLM layers via deep residual injection.

OUR FINDING

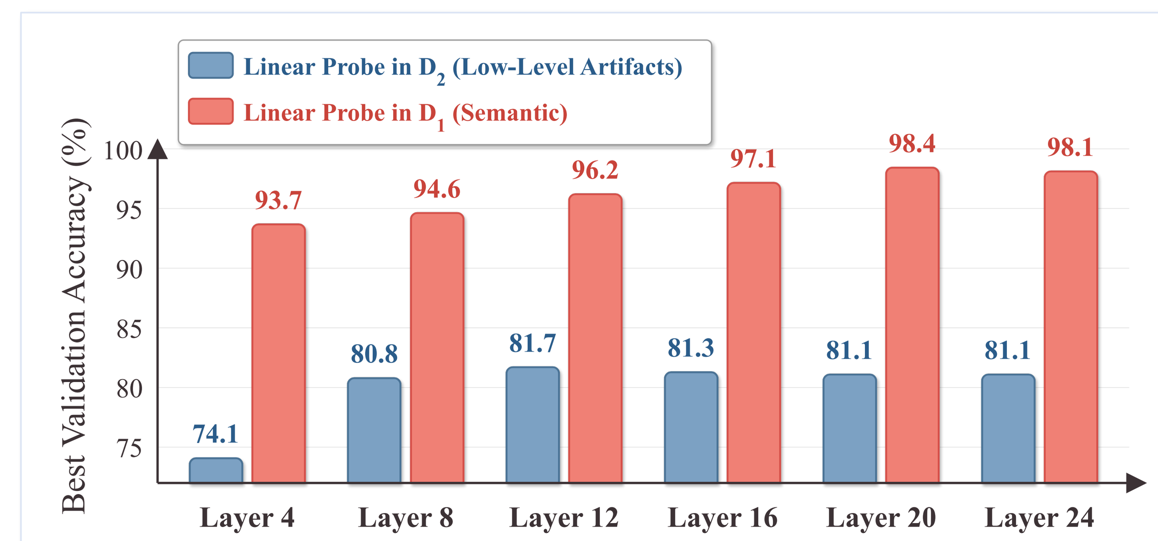
Takeaway 1: Feature-Space Conflict

- **Native MLLMs lack low-level sensitivity.** Generalizable artifact learning requires **large feature-space shifts**.
- **Large shifts cause semantic forgetting.** Full tuning drops common benchmarks: BLINK -93%▼, Real World VQA -83%▼



Takeaway 2: Layer-wise Misalignment

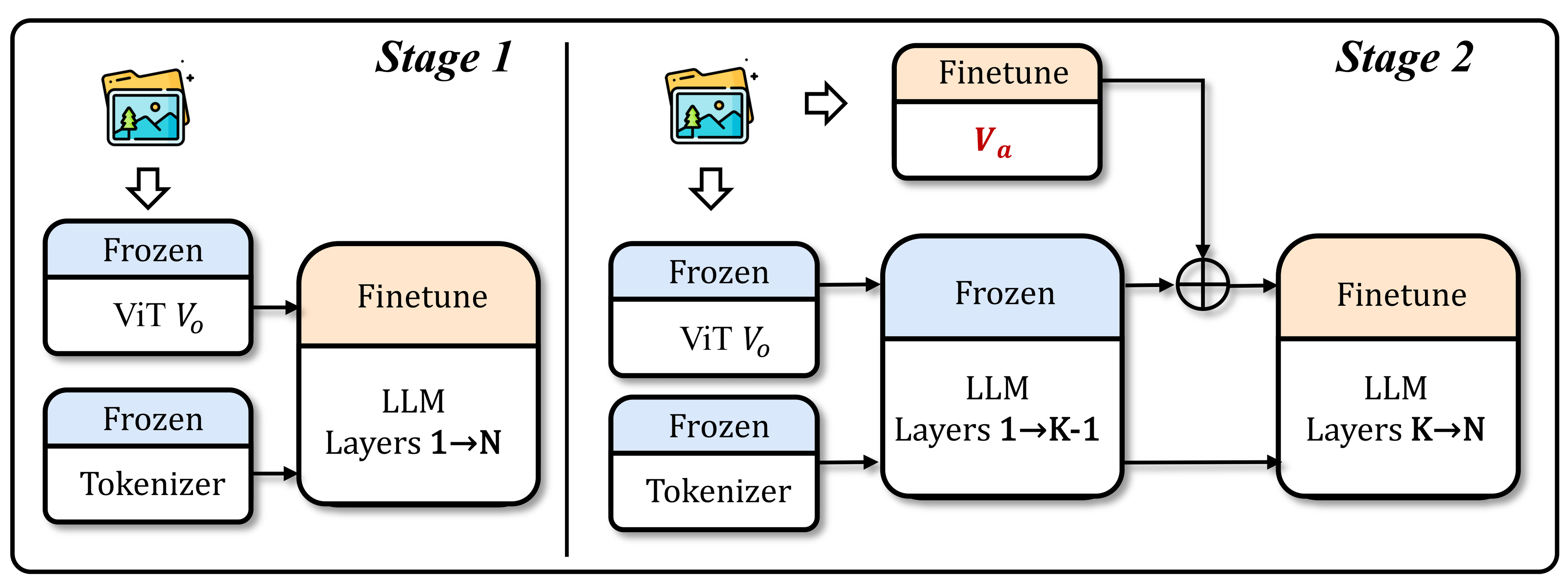
- **Early-to-middle layers encode semantic-features.** Semantic data **peaks at Layers 16–20**.
- **Low-level artifacts are weakly separable across layers.** Low-Level data stays around 81%▼ across depths.



OUR METHOD

Stage 1: Fine-tune the LLM on Semantic data while **freezing other modules to activate high-level semantic priors** for AIGI detection.

Stage 2: Freeze early LLM layers, **inject Low-Level artifact residuals before layer K**, and train later layers.



EVALUATION

Deep-VRM achieves SOTA performance

Model	MidJ	SDv1.4	SDv1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	**AVG**
Xception	57.97	98.06	97.98	51.16	57.51	97.79	50.34	48.74	69.94
CNNSpot	61.25	98.13	97.54	51.50	55.13	93.51	51.83	51.06	69.99
F3Net	52.26	99.30	99.21	49.64	50.46	98.70	45.56	49.59	68.09
GramNet	63.00	94.19	94.22	48.69	46.19	93.79	49.20	44.71	66.75
UnivFD	77.29	97.01	96.67	50.94	78.47	91.52	65.72	55.91	76.69
NPR	62.00	**99.75**	99.64	56.79	82.69	97.89	54.43	52.26	75.68
AIDE	79.38	99.74	**99.76**	78.54	91.82	98.65	80.26	66.89	86.88
DIRE	51.11	55.07	55.31	49.93	50.02	53.71	49.87	49.85	51.86
DRCT/Conv-B	94.43	99.37	99.19	66.42	73.31	99.25	76.85	59.41	83.53
DRCT/UniFD	85.82	92.33	91.87	75.18	87.44	92.23	89.12	87.38	87.67
OMAT	90.36	97.52	97.46	83.82	97.41	97.62	95.53	97.34	94.63
Deep-VRM	**96.16**	99.65	99.42	**89.82**	**99.08**	**99.44**	**97.29**	**98.50**	**97.42**

Method	*WildRF*				*AIGI Bench*		
	FaceBook	Reddit	Twitter	**AVG**	SocialRF	CommunityAI	**AVG**
FatFormer	64.38	76.65	40.00	60.34	57.98	50.62	54.30
CO-SPY	50.00	56.79	73.30	60.03	55.54	53.02	54.28
C2P-CLIP	54.38	68.40	47.27	56.68	53.13	50.98	52.06
SAFE	62.50	61.70	40.33	54.84	58.00	54.25	56.13
AIDE	75.00	55.48	48.00	59.49	57.80	54.15	55.98
Deep-VRM	**85.94**	**91.07**	**91.53**	**89.51**	**83.68**	**97.28**	**90.48**

VISUALIZATION EXAMPLE

Can you describe this photo?

Certainly. The image shows a zebra ...... The overall composition appears plausible at first glance.

Determine whether the image is real or fake.

This image is **fake**. It has **artifacts not related to the content itself but rather to the underlying digital texture and patterns**, which deviate from what would be expected in a real photograph.

Classify the image as fake or real. Answer using only one word.

fake

What evidence supports your decision?

The **car's front license plate is distorted and illegible**, ... Furthermore, **the vehicle's left side mirror is missing**, which contradicts standard automotive design and safety regulations.

Homepage

Paper

GitHub