TranX-Adapter: Bridging Artifacts and Semantics within MLLMs for Robust AI-generated Image Detection

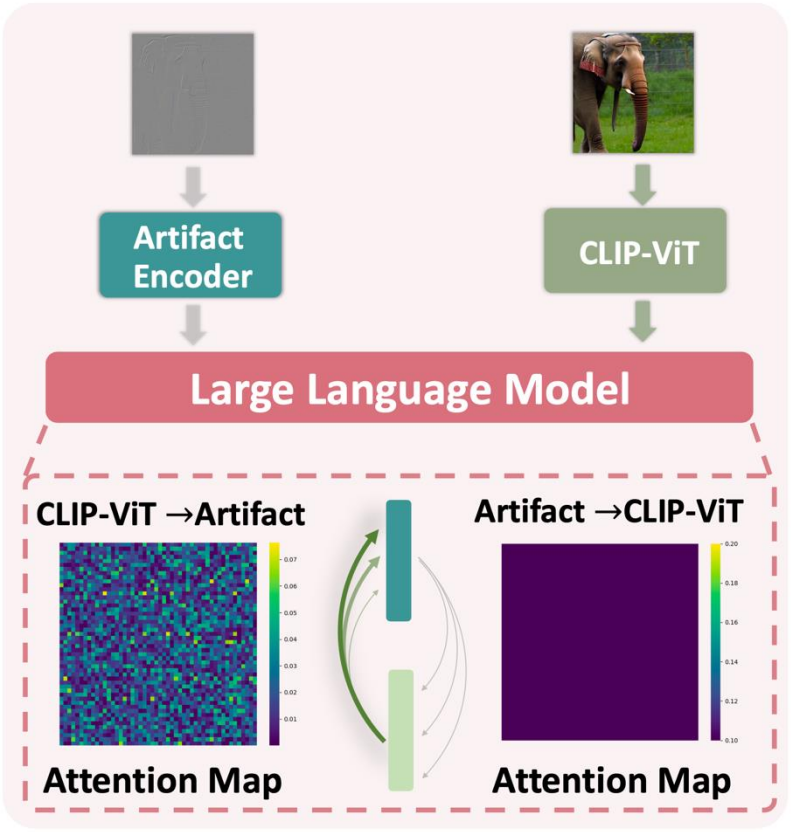
Wenbin Wang[1], Yuge Huang[2], Jianqing Xu[2], Yue Yu[2], Jiangtao Yan[2],
Shouhong Ding[2], Pan Zhou[3], Yong Luo[1]

[1]Wuhan University, [2] Tencent YouTu Lab, [3]Singapre Management University

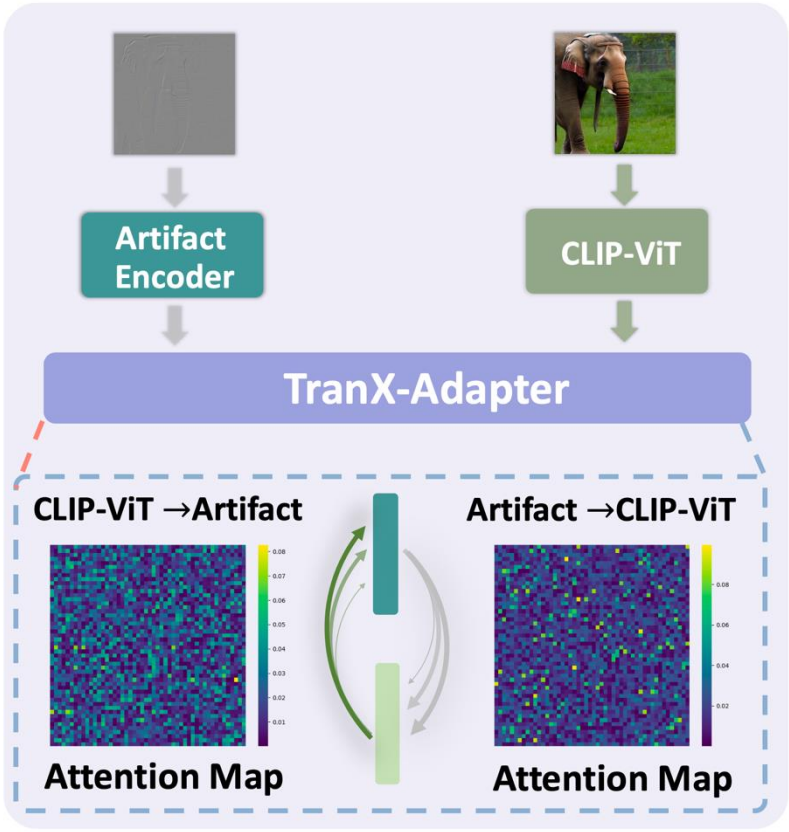
Motivation

Artifact image

RGB image



(a) Previous Method



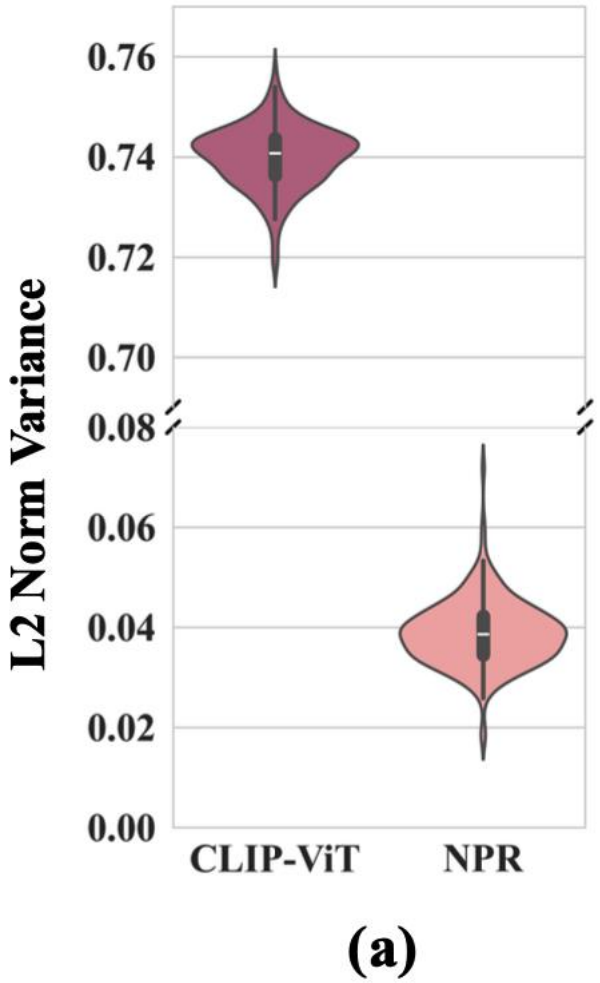
(b) Our TranX-Adapter

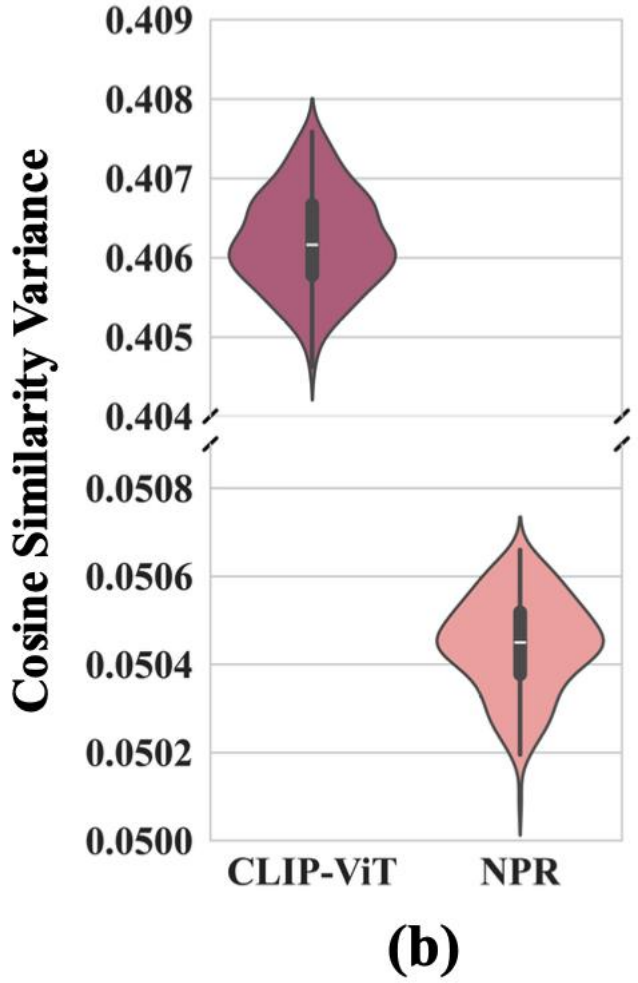
Summary of contributions

- Pilot Study
  - Identifying that fusing artifact and semantic features within MLLMs is hindered by the high intra-feature similarity of artifact representations.
- **A Lightweight Adapter**
  - TranX-Adapter: TOP-Fusion and X-Fusion
- **Analysis**
  - Results on cross-generators benchmark and in the wild AIGI detection benchmarks
  - Ablation Study
  - …

Pilot Study

- Uniform Artifact and Variance Semantic

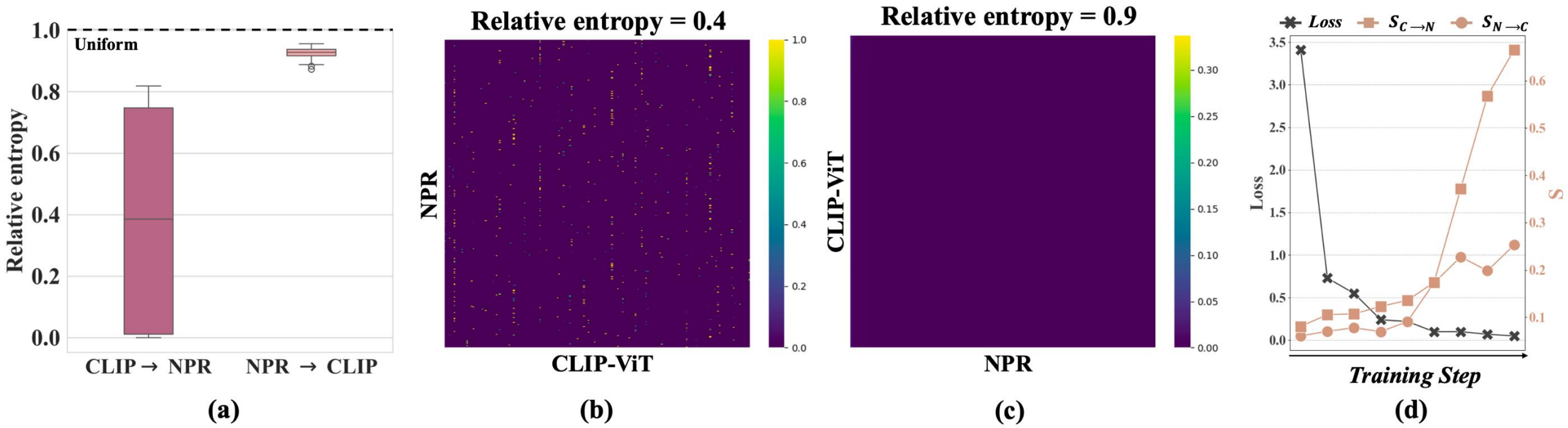




Artifact representations are more homogeneous and less discriminative in feature space.

Pilot Study

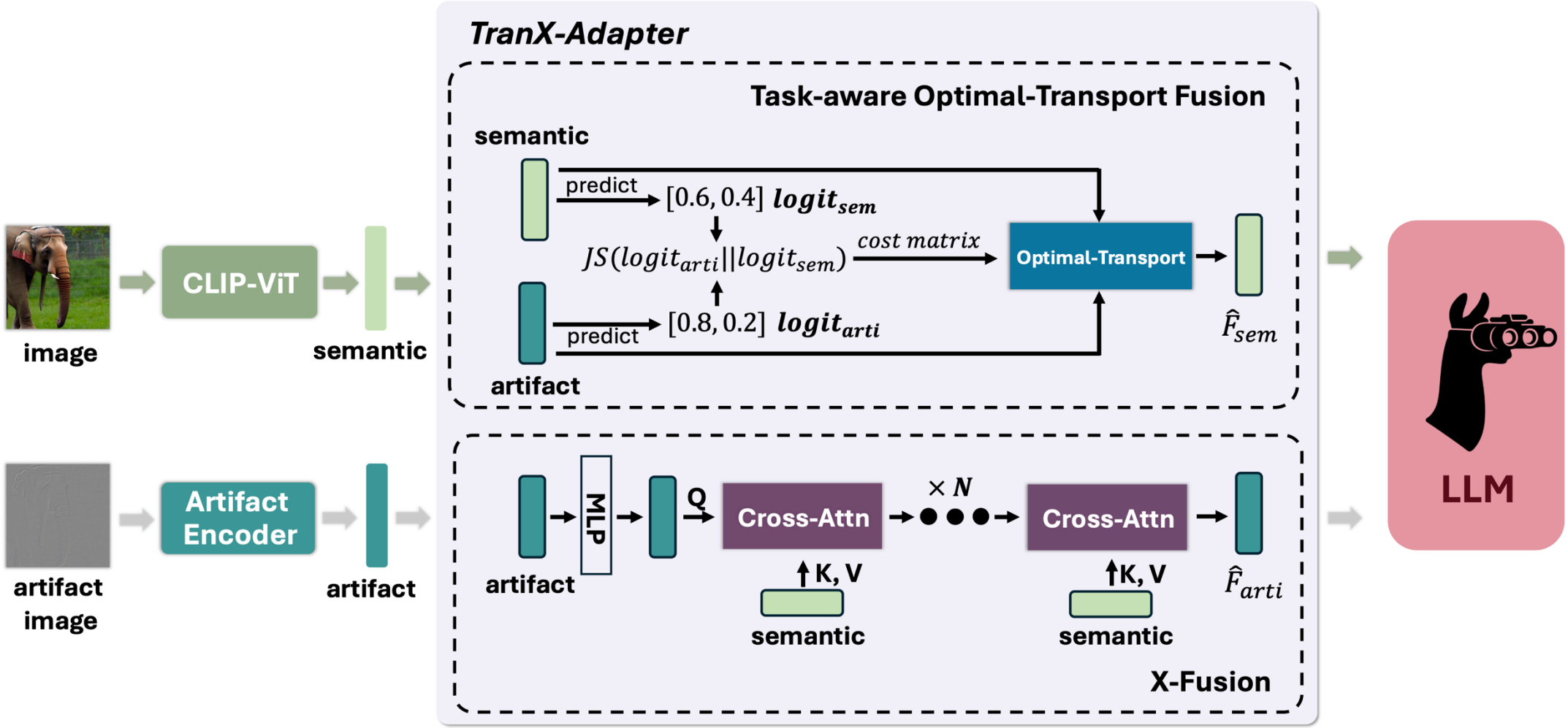
- Uniform Artifact and Variance Semantic



The artifact information encoded by NPR is difficult to transfer into the semantic feature space.

Our method: TranX-Adapter

➢ Overview



Analysis

➢ Results on GenImage Benchmark

Table 1. **Cross-model accuracy performance on the GenImage Dataset.** Accuracy (%) of different detectors (rows) in distinguishing real images from those produced by various generative models (columns). The best result are marked in **bold**. "†" indicates direct concatenation of artifact and semantic features following Zhou et al. (2025).

Method	*Midjourney*	*SD v1.4*	*SD v1.5*	*ADM*	*GLIDE*	*Wukong*	*VQDM*	*BigGAN*	*Mean*
CNNSpot (Wang et al., 2020)	52.8	96.3	95.9	50.1	39.8	78.6	53.4	46.8	64.2
F3Net (Qian et al., 2020)	50.1	**99.9**	**99.9**	49.9	50.0	**99.9**	49.9	49.9	68.7
DIRE (Wang et al., 2023b)	60.2	99.9	99.8	50.9	55.0	99.2	50.1	50.2	70.7
GenDet (Zhu et al., 2023a)	89.6	96.1	96.1	58.0	78.4	92.8	66.5	75.0	81.6
PatchCraft (Zhong et al., 2023b)	79.0	89.5	89.3	77.3	78.4	89.3	83.7	72.4	82.3
UnivFD (Ojha et al., 2023)	73.2	84.2	84.0	55.2	76.9	75.6	56.9	80.3	73.3
AIDE (Yan et al., 2025)	79.4	99.7	99.8	78.5	**91.8**	98.7	80.3	66.9	86.9
NPR (Tan et al., 2024)	89.8	90.7	90.7	84.6	90.3	90.7	87.0	81.8	88.3
LLaVA-1.6-mistral 7B† (Liu et al., 2024)	88.6	94.0	94.0	80.3	86.8	93.4	82.6	76.2	87.3
AIGI-Holmes (Zhou et al., 2025)	81.6	91.3	91.4	**88.4**	91.5	89.5	**90.9**	**94.5**	89.8
Qwen3-VL 2B† (Bai et al., 2025)	88.1	87.3	84.1	78.6	81.7	81.7	84.1	72.2	82.2
Qwen3-VL 4B† (Bai et al., 2025)	87.3	98.4	96.0	65.1	91.3	96.0	73.8	78.6	85.8
w/ our TranX-Adapter									
Qwen3-VL 2B	90.5	97.6	96.0	83.3	89.7	92.9	82.5	71.4	88.0
Qwen3-VL 4B	92.1	97.6	97.6	81.0	83.3	94.4	84.9	87.3	89.8
LLaVA-1.6-mistral 7B	**94.9**	96.4	96.4	87.0	88.0	94.9	90.1	85.9	**91.9**

Analysis

➢ Results on Cross-dataset Benchmark

Table 2. **Cross-dataset accuracy performance on the Chameleon testset.** "†" denotes the setting following Zhou et al. (2025), where the artifact features and semantic features are directly concatenated.

Method	*Training Set*	
	SDv1.4	*All GenImage*
UnivFD (Ojha et al., 2023)	55.6	60.4
DIRE (Wang et al., 2023b)	59.7	57.8
PatchCraft (Zhong et al., 2023b)	56.3	55.7
NPR (Tan et al., 2024)	58.1	57.8
LLaVA-1.6-mistral 7B† (Liu et al., 2024)	69.4	81.9
AIDE (Yan et al., 2025)	62.6	65.8
PatchAll/CLIP (Yang et al., 2025)	63.9	69.3
PatchAll/DINOv2 (Yang et al., 2025)	66.6	72.1
Qwen3-VL 2B† (Bai et al., 2025)	68.5	78.8
Qwen3-VL 4B† (Bai et al., 2025)	69.4	78.3
w/ our TranX-Adapter		
Qwen3-VL 2B	71.8	82.3
Qwen3-VL 4B	72.6	83.6
LLaVA-1.6-mistral 7B	**75.8**	**85.1**

Chameleon

Table 3. **Cross-dataset accuracy performance on the RRDataset.** The "***Ori.***" represents the Original, "***Trans.***" represents the Transmission, and "***Re.***" represents the Re-digitization. "†" indicates the direct concatenation of artifact and semantic features (Zhou et al., 2025).

Method	*RRDataset*			
	Ori.	*Trans.*	*Re.*	*Avg.*
MLLMs (Zero-shot)				
GPT-4o (Achiam et al., 2023)	94.5	84.7	73.1	84.1
Claude-3.7-sonnet (Ant)	89.9	83.8	73.9	82.5
Gemini-2-flash (Team et al., 2023)	85.3	74.8	71.8	77.3
Qwen2VL-72B (Wang et al., 2024)	59.9	56.4	59.8	58.7
Detectors (Fine-tuned)				
DIRE (Wang et al., 2023b)	94.0	94.1	50.2	79.4
AIDE (Yan et al., 2025)	79.0	76.8	79.6	78.4
GramNet (Liu et al., 2020)	78.0	77.6	70.8	75.4
CNNSpot (Wang et al., 2020)	80.8	77.3	64.9	74.3
NPR (Tan et al., 2024)	72.7	62.6	65.6	67.0
C2P-CLIP (Tan et al., 2025)	57.4	64.2	54.2	58.6
LLaVA-1.6-mistral 7B† (Liu et al., 2024)	94.8	69.3	85.5	83.2
Qwen3-VL 2B† (Bai et al., 2025)	88.9	89.9	68.8	82.5
Qwen3-VL 4B† (Bai et al., 2025)	96.0	89.5	71.3	85.6
w/ our TranX-Adapter				
Qwen3-VL 2B	97.5	95.3	78.3	88.9
Qwen3-VL 4B	**98.1**	**95.5**	**79.0**	**90.9**
LLaVA-1.6-mistral 7B	96.6	93.0	77.1	88.9

RRDataset

Conclusions

(1) High intra-feature similarity of artifact features leads to attention dilution under self-attention, while discrepancy-aware fusion is more effective.

(2) The LLM increasingly relies on visual information during training.

(3) Artifact-semantic fusion predominantly occurs in shallow layers.

Thanks !