Asymmetric Anchoring: Opening the MLLM Black Box for Forgery Detection

Zhiqiang Yang[1], Renshuai Tao[1,2⋆], Chunjie Zhang[1], Zhaoxiang Liu[3,4], Xiaolong Zheng[5], and Yao Zhao[1]

[1] Institute of Information Science, Beijing Jiaotong University
[2] Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University
[3] Data Science & Artificial Intelligence Research Institute, China Unicom
[4] Unicom Data Intelligence, China Unicom
[5] Institute of Automation, Chinese Academy of Sciences

Abstract. Multimodal large language models (MLLMs) are promising for forgery detection, but most methods treat them as fixed black boxes. Preliminary studies have shown that this passive use overlooks instabilities in the internal data flow, leading to representation drift. In this work, we introduce the Asymmetric Anchoring Paradigm (AAP), an open-box approach that reshapes the MLLM's internal data flow (rather than appending external components) by re-purposing its pre-trained visual encoder as an active truth anchor. AAP has two key steps: (1) for real images, we impose an anchor-alignment constraint that pulls representations toward the truth anchor, yielding a highly stable, low-variance manifold; (2) for tampered images, we measure their deviation from the anchor and use the resulting error map, which spatially quantifies departures from real-world priors, as a precise cue for the localization decoder. Comprehensive evaluations on SID-Set and OpenSDID demonstrate that AAP substantially improves detection accuracy and the localization IoU within manipulated regions. By opening the black box and anchoring to the encoder's real-world priors, AAP turns MLLMs from passive components into actively regularized detectors for practical forgery detection, offering a new insight for the field. The code is open-sourced and publicly available at `https://github.com/rstao-bjtu/AAP`.

1 Introduction

The proliferation of generative models, from the Sora series to various AIGC tools, has made the creation of forged content effortless, posing a severe threat to the credibility of digital information [32,43,50,58–61,72]. Traditional binary classification detectors [31,38,40,46,67] are struggling to meet the complex demands of modern forgery detection. In pursuit of more transparent and reliable evaluations, the research community has increasingly turned to Multimodal Large Language Models (MLLMs). As a result, a new research direction has emerged,

⋆ Corresponding author. Email: `rstao@bjtu.edu.cn`

harnessing the unique capabilities of MLLMs for comprehensive evaluations extending beyond simple classification to address complex tasks such as forgery detection, localization, explanation, and reasoning [17, 18, 34, 44, 65, 66, 69].

However, despite the progress made, existing methods face a fundamental limitation: almost universally, they treat these powerful MLLMs as immutable black boxes. Research has predominantly focused on designing clever visual/textual cues or appending supplementary external components.

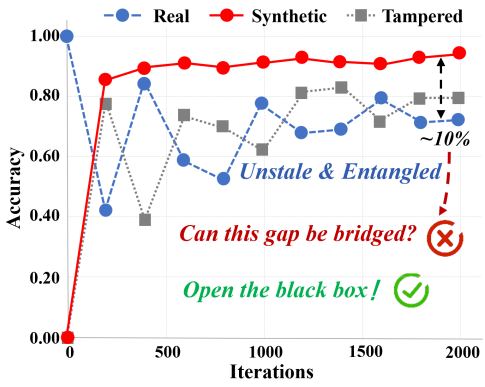

Fig. 1: Representation drift in black-box MLLM optimization. While the model quickly learns synthetic forgeries, optimization for real and tampered images is unstable, causing real image recognition to lag, revealing a performance bottleneck.

This black-box approach overlooks the inherent instability and untapped potential within the internal data flow, leaving a critical flaw unaddressed. To verify the existence of this defect, we analyze the training dynamics of models based on LLaVA [33] (Figure 1). Our results reveal an unexpected phenomenon: while the model quickly achieves high accuracy on fully synthetic categories, its performance on real and tampered images shows significant instability. During fine-tuning, the representations of these two categories become severely entangled. Instead of forming distinct clusters, their latent features collapse into an overlapping space (Figure 2).

While it is intuitively known that tampered images inherently share visual elements with authentic sources at the data level [9, 51], our diagnosis reveals a deeper issue occurring during the fine-tuning process. Specifically, we uncover an optimization conflict within the MLLM's internal representation space: as the model aggressively updates its weights to capture tampered forgery artifacts, it catastrophically corrupts the existing real-world knowledge, leading to a dynamic Representation Drift. We call this a specific type of asymmetry [68]. The model struggles to identify forgery traces in tampered images without misclassifying authentic real images. As a result, its accuracy on real images consistently lags behind, causing it to hit a performance bottleneck prematurely.

Based on this diagnosis, the observed drift is not an insurmountable flaw, but rather a missed opportunity. The solution lies not in adding complex external modules, but in harnessing an overlooked, pre-existing stabilizing force: the MLLM's foundational visual encoder, which serves as the missing "Truth Anchor". Extensively pre-trained on authentic images, this encoder is inherently rich in real-world priors. This prompts a crucial shift: **rather than simply appending components to a black-box paradigm, why not open and reshape the black box itself?** By intervening directly in the MLLM's internal

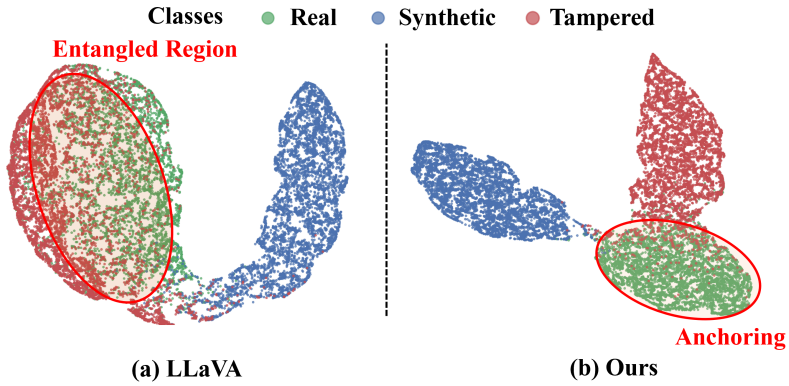

Fig. 2: T-SNE visualization of black-box LLaVA and AAP. (a) The black-box LLaVA clearly separates Full Synthetic images but shows highly diffuse and entangled representations for Real and Tampered images, limiting real image recognition. (b) In contrast, the AAP method effectively disentangles these representations.

data flow, intermediate visual representations can be realigned and anchored to the priors embedded within the visual encoder.

In this work, we propose the **A**symmetric **A**nchoring **P**aradigm (AAP), which applies distinct constraints for different image types: (1) For Real images, the primary classification objective and our internal alignment constraint work synergistically: the classification pulls features toward the Real class, while the alignment pulls them toward the Truth Anchor. This dual influence results in a stable, low-variance representation for real data, elegantly resolving the performance bottleneck (Figure 2). (2) For Tampered images, we decouple the objectives. The classification loss drives features away from the Real class, while we measure the alignment error against the Truth Anchor instead of applying an alignment loss. This measured deviation effectively signals inauthenticity. Crucially, this process generates a powerful byproduct: the *alignment error map*, which spatially quantifies local departures from real-world priors. We harness this map as a dense spatial cue for the localization decoder, enabling it to pinpoint tampered regions by contrasting their high alignment error against the low error of authentic regions. The contributions are summarized as follows:

- We are the first to demonstrate that the prevailing MLLM black-box paradigm neglects internal optimization instability, leading to representation drift that specifically bottlenecks real image recognition.
- We propose AAP, an open-the-black-box framework that redefines the visual encoder as a truth anchor to regularize representation optimization.
- We effectively utilize the alignment error map, a byproduct of the alignment process, as a spatial cue to enhance tampered image localization accuracy.
- Our method achieves state-of-the-art performance on both the SID-Set and OpenSDID benchmarks, thoroughly validating the superiority of our proposed open-the-black-box and representation anchoring paradigm.

2 Related Work

2.1 Traditional Forgery Detection and Localization

In the early stages of AIGC content forensics, forgery detection primarily focused on traditional image forensics techniques. Early works relied on analyzing specific statistical and frequency artifacts, such as compression artifacts, noise patterns, or camera fingerprints [13,23]. With the rise of deep learning, methods based on Convolutional Neural Networks became mainstream. These approaches achieved significant progress in binary classification tasks [6,31,38,46,70]. However, these models typically provide only opaque judgment, lacking interpretability and the ability to pinpoint specific tampered regions. These limitations make them insufficient for the multifaceted demands of modern forgery analysis.

To effectively address this limitation, a separate line of research focused on the task of forgery localization. These methods usually adopt semantic segmentation-like architectures, training an encoder-decoder model to output pixel-level forgery masks [15, 37, 41]. Although these specialized models excel at localization, they are often decoupled from the classification task, require extensive separate training, and lack the capability to provide natural language explanations. The fragmented nature of these traditional methods, coupled with the further demands placed on forensics by the high fidelity of generative techniques, has naturally led to the exploration of MLLMs for forgery detection applications.

2.2 Application of MLLM in Forgery Detection

In pursuit of a more transparent, interpretable, and comprehensive versatile forgery assessment, the research community has increasingly turned to MLLMs. The prevailing architectural paradigm of advanced contemporary MLLMs, such as LLaVA [33] and InstructBLIP [7], typically bridges a powerful, pre-trained visual encoder (e.g., CLIP-ViT [47]) with a large language model (LLM, e.g., LLaMA [62]) via a carefully designed lightweight projection module. This design endows the models with strong joint visual-language reasoning capabilities.

A primary branch of existing research focuses on leveraging the powerful reasoning and instruction-following capabilities of MLLMs. For instance, ForgeryGPT [34] and FakeShield [66] utilize meticulously designed prompting to guide the model in generating detailed textual analyses and forgery explanations. This MLLM-as-an-evaluator paradigm was subsequently extended: MLLM-as-a-Judge [65] validated the efficacy of MLLMs as reliable zero-shot image safety arbitrators, while HEIE [69] specifically employs their semantic understanding capabilities for an interpretable assessment of image impossibilities. Concurrently, other recent works have focused on augmenting MLLMs with new functionalities or optimizing their inputs. SIDA [18] and FakeShield [66] address the fundamental challenge of spatial localization by introducing a special `<SEG>` token, enabling the LLM to command an external SAM decoder [5]. In contrast, MM-FFD [17] concentrates on input-side engineering, designing a Visual-Knowledge Adapter

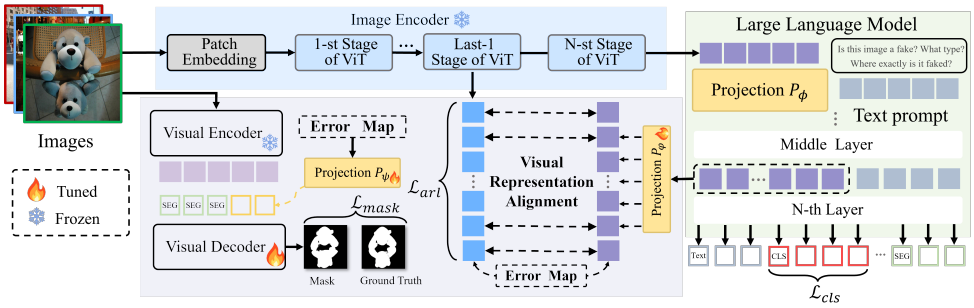

Fig. 3: The proposed open-box architecture for forgery detection and localization. The framework features two core modules: Truth Anchor-based internal representation alignment, and alignment error-driven spatial cue injection.

and a Text-Prompt Aligner to better align visual artifacts with text prompts before they enter the LLM, thereby enhancing interpretability.

However, as we noted in our introduction, these advancements almost universally follow a black-box paradigm. Despite being built upon MLLMs architectures, their innovations are primarily concentrated on external module design: for example, designing clever visual prompts, constructing multi-task instructions, or appending a decoder to the MLLM's final output layer [18, 65, 66, 69]. The MLLM itself is treated as an opaque, monolithic reasoning engine or an external component. Scant research has delved into its internal dataflow to investigate its stability and potential issues during fine-tuning on forensics tasks.

3 Asymmetric Anchoring Paradigm

To verify the effectiveness and universality of our proposed paradigm, we instantiate our framework using explainable forgery detection and localization task as an example. This complex and multifaceted task serves as an ideal testbed.

3.1 Preliminaries

The existing black box MLLM paradigm for e-IFDL [18] treats the system $\mathcal{M}$ as a sequential data processing pipeline. Formally, the input image $x_i \in \mathbb{R}^{H\times W\times 3}$ is first encoded by a visual encoder V (either a pretrained model or a custom-designed architecture) into a sequence of N image patch features $z_v = V(x_i)$, where $z_v \in \mathbb{R}^{N\times D_v}$. These visual embeddings are then mapped by a projector P into the lexical embedding space of a large language model $\mathcal{L}$.

The language model $\mathcal{L}$ then processes these projected features along with textual cues, deriving a high-level semantic representation from its final hidden state. In this paradigm, optimization is entirely driven by the endpoint objective acting on this final state: the classification loss $\mathcal{L}_{cls}$ determines the image category, while the segmentation loss $\mathcal{L}_{seg}$ guides the semantically labeled localization decoder $\mathcal{D}_{seg}$. Existing methods, often inspired by LISA [29], typically exploit this by extracting representations for `<CLS>` and `<SEG>` tokens from the

final layer. This dependence on the endpoint signal, lacking internal control, leads to the Representation Drift identified in Sec. 1.

3.2 Anchoring Representations

We propose an open box reshaping paradigm (Fig. 3) that intervenes in the MLLM's internal dataflow to resolve the optimization instability identified.

Defining the Truth Anchor. We first define the Truth Anchor z_a as the patch representations extracted from the penultimate layer of the visual encoder, V_p. Our choice of the penultimate layer is deliberate. As demonstrated in prior work [33], the penultimate layer retains a richer, unprojected representation of these visual priors—such as texture, edges, and local artifacts. Therefore, z_a represents the ground truth perceptual representation of the input image x_i, as interpreted by V's extensive real world pretraining.

$$z_a^i = V_p(x_i) \in \mathbb{R}^{N \times D_a} \tag{1}$$

Since V remains frozen, the Truth Anchor z_a serves as a stable, corruption-free perceptual reference. The performance improvement derives from targeted regularization rather than parameter expansion. Unlike randomly initialized modules, this visual encoder has been pre-trained on massive datasets of authentic images, embedding an overwhelmingly strong normative real-world prior into its latent space. Recent studies [42, 52, 55] have shown that such pre-trained features can implicitly capture forgery artifacts. Even for high-fidelity forgeries that appear realistic to the human eye, subtle inconsistencies in low-level statistics and micro-textures invariably disrupt the manifold of authentic representations [30, 56]. We argue that this capability stems precisely from the fact that forgeries deviate from the encoder's ingrained priors. Therefore, the Truth Anchor z_a serves not as a completely forgery-blind baseline, but rather as an authentic gravitational center. For authentic images, z_a provides a highly stable reference to prevent representation drift during MLLM fine-tuning. For synthetic or tampered images, their inherent unnaturalness causes a measurable perturbation against this strong authentic prior. This perceptual deviation from the Truth Anchor subsequently serves as the crucial source signal for our Error Map.

Anchoring Representations. Our framework's optimization is defined by two competing objectives. The first is the standard classification objective $\mathcal{L}_{cls}$. Applied to the model's final output, it acts as a semantic force encouraging the separation of Real and Fake representations into distinct clusters. The second, which forms the core of our intervention, is our novel **Anchoring Representation Loss (ARL)**, $\mathcal{L}_{arl}$. We intervene by extracting the intermediate patch representations $z_l = \{z_l^p\}_{p=k}^N$ from a middle layer k of the language model $\mathcal{L}$. This choice is motivated by recent studies [21, 73] on MLLM internal information flow. These works reveal a processing hierarchy: early layers aggregate global visual context, intermediate layers capture fine grained and spatially localized features, and later layers integrate information for response generation.

As the intermediate layers are critical for visual understanding and pivotal for visual grounding [19,22,71], we target them for our internal feature alignment

process. To effectively compare these **dimensionally** disparate representations, we introduce a lightweight alignment projector P_φ (a simple MLP, detailed in Supplementary Materials) to map $\mathbf{z}_l$ to the anchor's dimension. $\mathcal{L}_{arl}$ is then defined as the mean cosine distance between the projected representations and the Truth Anchor, summed over all N spatial patches:

$$\mathcal{L}_{arl} = \frac{1}{N}\sum_{p=1}^{N}\left(1 - \frac{P_\varphi(z_l^p)\cdot z_a^p}{\|P_\varphi(z_l^p)\|\|z_a^p\| + \epsilon}\right) \tag{2}$$

The $\mathcal{L}_{arl}$ term serves as an auxiliary perceptual constraint, applied asymmetrically. As defined in our overall objective (Eq. 8), the $\mathcal{L}_{arl}$ term is computed exclusively for real images ($y_i = real$). For these images, the semantic and perceptual goals are perfectly aligned. This synergy forces the model to learn an ultra-stable, low-variance representation for genuine data, effectively resolving the performance bottleneck. Conversely, for tampered or synthetic images, the classification and alignment objectives conflict. The classification loss ($\mathcal{L}_{cls}$) drives the MLLM's intermediate representation z_l to capture forgery features, inherently pushing it further away from the Truth Anchor's fixed perceptual baseline z_a. To avoid this optimization conflict, our asymmetric design excludes $\mathcal{L}_{arl}$ from the loss function for non-real images. Instead, we compute the patch-wise alignment deviation d_p (Eq. 3) purely as a quantitative metric of this perceptual divergence. Detached from backpropagation, this deviation forms the Error Map (M_e), providing a powerful spatial cue for localization (as detailed in Sec. 3.3).

Theoretical Analysis from a Gradient Perspective. Our asymmetric contrastive regularization provides a sound optimization mechanism. For real images, the optimization is governed by $\nabla_z\mathcal{L} = \nabla_z\mathcal{L}_{cls} + \lambda\nabla_z\mathcal{L}_{arl}$. The $\mathcal{L}_{arl}$ term mathematically imposes a **linear restoring force** (conceptually a "centripetal force") that constrains the feature trajectory within the pre-trained low-variance manifold, effectively neutralizing representation drift. Conversely, for tampered images, this constraint is released, allowing $\nabla_z\mathcal{L}_{cls}$ to dominate as a "centrifugal force" that actively pushes these tampered features further away from the real manifold. Consequently, this asymmetric dynamic naturally guarantees a mathematical margin $\Delta = \|z_{tamp} - z_a\| - \|z_{real} - z_a\| > 0$, resolving gradient conflicts without the need for explicit negative pairs.

3.3 Error Map as a Spatial Prompt

Defining the Error Map. M_e is a 2D map, corresponding to the patch resolution. We first define the alignment failure d_p for each patch p as the cosine distance (the term inside the summation of Eq. 2):

$$d_p = 1 - \frac{P_\varphi(z_l^p)\cdot z_a^p}{\|P_\varphi(z_l^p)\|\|z_a^p\| + \epsilon} \tag{3}$$

The Error Map M_e is then the spatial collection of these individual patch errors, reshaped into the $H_p \times W_p$ patch grid (where $N = H_p \times W_p$):

$$M_e = \text{Reshape}_{H_p \times W_p}\left([d_1, d_2, \ldots, d_N]\right) \tag{4}$$

For authentic images, this map is uniformly low. For tampered images, M_e presents high-error hotspots over manipulated regions, precisely where the internal representations z_l deviate from the Truth Anchor z_a.

Injecting the Spatial Cue. We uniquely leverage this byproduct as an explicit spatial prior for localization, directly indicating regions of perceptual deviation to the decoder without relying on external dense annotations. Specifically, the patch-level error map M_e is bilinearly upsampled to match the spatial resolution $(H_e \times W_e)$ of the localization encoder's feature map E_{enc}, yielding M_e^r. Treating this map as a dense visual prompt, we employ a lightweight convolutional projector P_ψ to map M_e^r into a high-dimensional discrepancy embedding E_e, aligning its channel dimension C_e with that of the encoder features:

$$M_e^r = \text{Upsample}_{H_e \times W_e}(M_e) \tag{5}$$

$$E_e = P_\psi(M_e^r) \in \mathbb{R}^{C_e \times H_e \times W_e} \tag{6}$$

Our method injects this signal by fusing two distinct prompt types. The discrepancy embedding (E_e), which carries low-level perceptual evidence of conflict, is combined via element-wise addition with the primary semantic prompt (E_p). This semantic prompt is derived by the prompt encoder from the LLM's `<SEG>` token embeddings and represents the high-level intent to segment. The resulting feature E_p^A becomes the final dense prompt for the decoder:

$$E_p^A = E_p + E_e \tag{7}$$

This fused prompt, E_p^A, is then passed to the mask decoder, while the original visual features E_{enc} are passed unmodified. This step (Eq. 7) prompts the decoder, providing it with not just the location of the conflict (via M_e^r) but also a learned, high-dimensional interpretation of that conflict.

3.4 Overall Objective and Implementation

Our model is trained end-to-end by optimizing a composite, multi-task objective $\mathcal{L}$. The final loss target is as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{seg} + \lambda_3 \mathcal{L}_{arl} \tag{8}$$

where $\mathcal{L}_{seg}$ is the standard combination of Dice and BCE loss for the mask and λ are balancing hyperparameters.

4 Experiments

4.1 Settings

Implementation Details. We adopt LLaVA-v1.5-7B [33] and LISA-7B-v1 [29] as our default large vision-language foundation models. Both models couple the

Vicuna-1.5 language model [4] with the CLIP visual encoder [48]. Furthermore, we also employ SAM [5] as our localization encoder/decoder to facilitate precise mask generation and spatial understanding. We fine-tune the models using the LoRA technique, setting the hyperparameter α to 16 and the dropout rate to 0.05. The input images are resized to 1024×1024. All models are trained for 20 epochs using the Adam optimizer [24] with a fixed learning rate of 1e-4 and a batch size of 16. Based on our empirical analysis, the loss hyperparameters λ_{cls}, λ_{seg}, and λ_{arl} (from Eq. 8) are set to 1.0, 1.0, and 3.0, respectively. All experiments are conducted on two NVIDIA H100 (80GB) GPUs.

Evaluation Protocols and Dataset. For evaluation, we comprehensively assess our method across three forensic dimensions: detection, localization, and explanation. For detection, we report image-level Accuracy and F1 scores. For pixel-level localization, we adopt AUC, F1, and Intersection over Union (IoU). **For interpretability**, the implementation details and qualitative evaluations are provided in the **Supplementary Materials** due to space constraints. Crucially, our experiments aim beyond standard generalization to validate our core hypothesis: **anchoring stabilizes optimization and resolves representation drift**. We deliberately select SID-Set [18] and OpenSDID [64] as our primary testbeds. Their high-fidelity, three-category paradigm (*real*, *fully synthetic*, and *tampered*) explicitly triggers the severe representation entanglement and optimization bottlenecks observed in baseline MLLMs (Sec. 1), providing the ideal environment to verify our Asymmetric Anchoring Paradigm. SID-Set comprises 300,000 images, focusing on high-fidelity social media content. Its dataset is evenly divided into three parts: 100,000 real images, 100,000 high-fidelity synthetic images (generated by Flux [28]), and 100,000 intricately manipulated tampered images (via GPT-4o and Latent Diffusion). We utilize the standard training and test splits of SID-Set. This dataset directly tests our method's ability to maintain the integrity of the real-world prior against state-of-the-art forgery artifacts. OpenSDID is curated for spotting AI-generated content in the open world, integrating a suite of SOTA T2I diffusion models (SD1.5 [49], SD2.1 [49], SDXL [45], SD3 [10], and Flux.1 [28]). We follow its open-world protocol: training exclusively on SD1.5 and testing across all generators. Here, our goal is to evaluate **prior robustness**: if our Truth Anchor successfully aligns the MLLM with fundamental authentic priors rather than superficial artifacts, the model should naturally exhibit superior cross-generator stability, resolving the performance degradation seen in artifact-reliant detectors.

4.2 Detection Evaluation

We comprehensively evaluate the detection performance of our method on two benchmarks: SID-Set and OpenSDID. Partial results are cited from [29, 64].

Results on SID-Set. As shown in Table 1, our method achieves leading performance across all three categories and demonstrates strong performance with 97.1% Acc and 97.0% F1 on average metrics. Most notably, our method directly solves the core bottleneck identified in the introduction: the recognition of real images. This validates its strength on real images.

Methods	Real		Synthetic		Tampered		Overall	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Antifake [2]	88.9	89.1	97.5	97.9	90.9	80.4	92.4	89.1
CnnSpott [12]	89.0	90.8	90.7	88.1	68.1	64.0	82.6	80.9
FreDect [11]	46.0	47.6	60.9	66.0	37.1	53.0	47.6	55.4
Fusing [20]	89.2	92.7	88.1	89.1	27.0	31.4	68.1	71.0
Gram-Net [36]	89.2	91.7	97.9	98.6	89.9	86.9	92.1	92.4
UnivFD [42]	68.3	68.5	86.4	98.0	92.5	90.0	82.4	85.4
LGrad [57]	62.0	76.1	58.0	67.3	**99.1**	**98.8**	73.0	80.6
LNP [1]	14.4	23.0	36.2	35.6	<u>93.3</u>	94.6	48.0	51.1
LLaVA-7B [33]	83.1	81.7	98.0	98.9	82.2	82.8	88.3	87.8
SIDA [18]	89.1	91.0	98.7	98.6	92.7	91.0	93.5	93.5
AAP-llava	**98.2**	**95.9**	<u>99.8</u>	<u>99.7</u>	92.0	94.9	<u>96.6</u>	<u>96.6</u>
AAP-lisa	<u>97.8</u>	<u>95.2</u>	**99.9**	**99.8**	<u>93.3</u>	<u>95.2</u>	**97.1**	**97.0**

Table 1: Performance comparison of our method and other deepfake detectors on the SID-Set dataset. All methods were fine-tuned on the SID-Set training data. Bold text represents the best results, and underlined text represents the second-best results.

The baseline LLaVA achieves only 83.1% Acc on the Real category, while the previous SOTA method, SIDA, reaches only 89.1%. In contrast, our method (AAP-lisa) achieves an impressive 97.8% accuracy. This is an improvement of over 14 percentage points compared to the baseline. This result powerfully validates our hypothesis. The exclusive application of $\mathcal{L}_{arl}$ to the Real category, which is the core of our asymmetric anchoring paradigm, forces the model to dedicate its optimization power to perfecting a stable, low-variance representation for authentic data. This directly resolves the representation drift bottleneck that plagues baseline models, leading to the near-perfect separation of real images visually reflected by the T-SNE visualization in Figure 2.

Results on OpenSDID. We hypothesized that by anchoring to fundamental realness priors instead of superficial forgery artifacts, our method would yield superior generalization. The evaluation on OpenSDID (Table 2) confirms this inference. As the results show, artifact-reliant detectors (e.g., RINE) suffer from performance degradation when moving from in-domain (SD1.5) to unseen, advanced generators (e.g., SDXL, Flux.1). In contrast, our method maintains the highest or second-highest performance across all unseen domains, achieving the best overall average. This indicates that our Truth Anchor acts as a powerful regularizer. Because the anchor itself is pre-trained on diverse, real-world data and is agnostic to any specific generator's artifacts, our $\mathcal{L}_{arl}$ constraint prevents the model from overfitting to these superficial tells. Instead, the model is compelled to learn a more fundamental and robust decision boundary based on this concept of realness, ensuring high performance even on unseen generators. The detection results for specific categories are shown in Figure 3, further confirming that our detection performance for the real category remains highly effective.

4.3 Localization Evaluation

We further evaluate the performance of our method in pixel-level localization.

Method	SD1.5		SD2.1		SDXL		SD3		Flux.1		AVG	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
FreqNet [56]	75.9	77.7	61.0	68.4	53.2	64.0	53.5	64.4	38.5	57.1	56.4	66.3
NPR [54]	79.4	79.3	81.7	81.8	72.1	74.3	73.4	75.5	67.6	71.4	74.9	76.5
UniFD [42]	77.5	77.6	80.6	81.9	70.7	74.8	71.1	75.2	61.1	69.1	72.2	75.7
RINE [25]	91.1	91.0	87.5	88.1	73.4	78.8	72.1	76.8	55.9	67.0	76.0	80.3
MVSS-Net [3]	93.5	93.7	79.3	82.3	59.9	70.4	62.8	72.1	27.6	56.8	64.6	75.1
PSCC-Net [35]	**96.1**	**96.1**	76.9	80.9	55.7	68.8	59.8	70.9	51.8	67.0	68.0	76.8
ObjFormer [63]	71.7	75.2	66.8	72.6	49.2	62.9	48.3	62.5	37.9	58.1	54.8	66.3
DeCLIP [53]	80.7	78.3	84.0	82.8	70.7	70.6	69.9	68.4	51.8	65.6	71.4	73.1
IML-ViT [39]	94.5	75.7	69.7	61.2	41.0	50.0	44.7	51.3	18.2	43.6	53.6	56.4
SIDA-7B [39]	91.3	91.2	87.3	88.7	74.6	77.7	76.6	80.2	66.7	73.3	79.3	82.2
OpenSDI [64]	92.6	92.7	88.7	89.5	78.0	81.2	73.1	78.0	56.5	68.5	77.8	82.0
AAP-llava (**Ours**)	92.2	91.6	90.2	**90.5**	81.7	83.5	**84.9**	**85.9**	**73.1**	**76.7**	**84.4**	**85.6**
AAP-lisa (**Ours**)	91.2	90.3	**90.6**	90.1	**83.9**	**85.0**	82.8	84.2	73.0	**76.7**	84.3	85.3

Table 2: Performance comparison of our method and other deepfake detectors on the OpenSDID dataset. All methods were fine-tuned on the OpenSDID training data. Bold text represents the best results, and underlined text represents the second-best results.

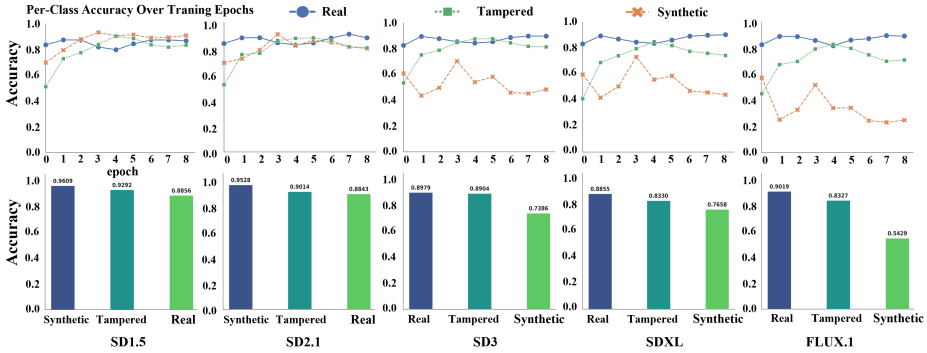

Table 3: The training and optimization process on the OpenSDID dataset, and the accuracy of the three categories in the best results.

Results on SID-Set. As shown in Table 5, our method achieves SOTA performance on the SID-Set localization task. This leap in performance is directly attributable to the injection of M_e into the decoder, which propels the IoU score to 56.8%. This dramatically surpasses the previous SOTA (SIDA-7B at 43.8%) by 13.0 full percentage points. This significant gain is mirrored on the AUC metric, where our method improves from 87.3% to 93.9%. This performance gain validates the efficacy of our alignment byproduct as a spatial prior. Crucially, it confirms that the measured deviation from the Truth Anchor is far from a mere abstract scalar; it inherently captures the spatial footprint of tampering, thereby providing explicit and precise guidance for the localization decoder.

Results on OpenSDID. Table 4 further shows that this spatial cue is also highly generalizable. In the cross-generator localization test, our method again surpasses the SOTA method OpenSDI, which was designed specifically for pixel-

Method	SD1.5		SD2.1		SDXL		SD3		Flux.1		AVG	
	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1
MVSS-Net [8]	57.9	65.3	44.9	51.8	14.7	18.5	26.9	32.7	4.8	6.4	29.8	34.9
CAT-Net [27]	66.4	<u>74.8</u>	54.6	62.3	25.5	30.7	35.6	42.1	5.0	6.6	37.4	43.3
PSCC-Net [35]	54.7	64.2	36.7	44.8	19.7	26.1	29.3	37.3	8.2	11.6	29.7	36.8
ObjFormer [63]	51.2	65.7	47.4	41.4	7.4	9.8	9.4	12.6	5.3	7.3	24.1	27.4
TruFor [14]	63.4	71.0	<u>54.7</u>	61.9	26.6	31.9	32.3	38.5	7.6	9.7	36.9	42.6
DeCLIP [53]	37.2	43.4	35.7	41.9	14.6	18.2	27.3	33.4	11.2	14.3	25.2	30.3
IML-ViT [39]	<u>66.5</u>	73.6	44.8	50.6	21.5	26.0	23.6	28.4	6.1	7.9	32.5	37.3
OpenSDI [64]	**67.1**	**75.6**	**55.5**	62.9	31.0	37.0	<u>43.8</u>	51.2	16.2	20.3	<u>42.7</u>	49.4
LISA-7B-v1 [29]	49.4	61.8	49.8	<u>63.3</u>	<u>32.2</u>	<u>45.5</u>	41.7	<u>54.2</u>	<u>23.1</u>	<u>36.3</u>	39.2	<u>52.2</u>
AAP-lisa (**Ours**)	53.0	65.2	54.1	**67.8**	**35.2**	**49.7**	**44.4**	**58.2**	**31.6**	**44.8**	**43.7**	**57.1**

Table 4: Pixel-level localization performance on the OpenSDID dataset. All models were trained on the sd1.5 training set. Bold text represents the best results, and underlined text represents the second-best results.

level localization, on average (AVG) IoU and F1 scores. This advantage is most pronounced on the most challenging cross-architecture generator, Flux.1, where our IoU is nearly double that of the SOTA method. This indicates that the deviation-based anchor-deviation map is a more fundamental localization signal than the specific generator artifacts other models may rely on.

Symmetric vs. Asymmetric Anchoring. A critical question is whether $\mathcal{L}_{arl}$ should also be applied to forged images. To investigate this, we investigate applying $\mathcal{L}_{arl}$ symmetrically to both Real and Tampered images (Table 6). While detection accuracy only drops by ∼3%, as the endpoint $\mathcal{L}_{cls}$ still dominates the final prediction, the localization IoU plummets to 40.2%. This perfectly validates our hypothesis: forcing forged images to align with the Truth Anchor artificially suppresses their inherent artifacts and erases the perceptual deviation signal (d_p). Consequently, the Error Map (M_e) becomes homogeneous, leaving the localization decoder without spatial guidance. The asymmetric design is thus essential for preserving the spatial conflict required for precise localization.

Methods	Backbone	Localization		
		AUC	F1	IOU
MVSS-Net [8]	ResNet-50	48.9	38.3	23.7
HIFI-Net [16]	CNN	64.0	34.8	21.1
PSCC-Net [35]	HRNetV2p	82.1	52.6	35.7
LISA-7B-v1 [29]	LLaVA-7B	78.4	49.1	32.5
SIDA-7B [18]	LLaVA-7B	87.3	60.9	43.8
SAFIRE [26]	SAM	-	62.8	45.8
AAP-lisa (**Ours**)	LLaVA-7B	**93.9**	**72.4**	**56.8**

Table 5: Results of pixel-level localization performance evaluation on the SID-Set dataset. Partial results are cited from [29].

Anchoring Strategy	Detection Acc (%)	Localization IoU (%)
Symmetric (Real + Tampered)	94.8	40.2
Asymmetric (**Ours**)	**97.1**	**56.8**

Table 6: Symmetric vs. Asymmetric Anchoring. Performance comparison on the SID-Set when applying $\mathcal{L}_{arl}$ symmetrically to all images versus asymmetrically.

4.4 Ablation Study and Analysis

We conducted a series of ablation studies to validate the effectiveness of the key components in our framework.

Language Model	Vision Encoder	$\mathcal{L}_{\text{ARL}}$	SD1.5	SD2.1	SDXL	SD3	Flux.1	SID-Set
Vicuna-1.5-7B [33]	CLIP	✗	88.31%	88.25%	78.90%	81.60%	71.25%	91.78%
		✓	**91.56%**(+3.25)	**90.45%**(+2.20)	**83.50%**(+4.60)	**85.90%**(+4.30)	**76.70%**(+5.45)	**96.60%**(+4.82)
	SigLIPv2	✗	87.32%	85.67%	75.19%	78.45%	70.83%	93.21%
		✓	**88.79%**(+1.47)	**88.92%**(+3.25)	**80.53%**(+5.34)	**83.12%**(+4.67)	**75.91%**(+5.08)	**97.04%**(+3.83)
LISA-7B-v1 [29]	CLIP	✗	91.17%	88.70%	77.65%	80.20%	73.25%	93.50%
		✓	**90.25%**(-0.92)	**90.08%**(+1.38)	**85.00%**(+7.35)	**84.20%**(+4.00)	**76.72%**(+3.47)	**97.10%**(+3.60)
Vicuna-1.5-13B [29]	CLIP	✗	92.17%	89.21%	78.26%	81.14%	74.26%	94.13%
		✓	**92.56%**(+0.39)	**90.58%**(+1.37)	**84.84%**(+6.58)	**84.90%**(+3.76)	**77.52%**(+3.26)	**97.24%**(+3.11)

Table 7: Visual representation anchoring impact. We evaluate models trained with and without $\mathcal{L}_{arl}$ across various visual encoders and LLM backbones.

Impact of Anchoring Representation Loss ($\mathcal{L}_{arl}$). To isolate the effectiveness of our core contribution, $\mathcal{L}_{arl}$, we compare MLLMs with (✓) and without (✗) the $\mathcal{L}_{arl}$ constraint in Table 7. The results clearly show that adding $\mathcal{L}_{arl}$ can bring consistent and significant performance improvements for LLM base models or visual encoders (CLIP, SigLIPv2) with different weights. Notably, it boosts LLaVA-v1.5-7B accuracy from 83.35% to 87.45% (+4.1%). These sustained improvements across diverse architectures and scales (7B to 13B) validate $\mathcal{L}_{arl}$ as a robust, model-agnostic solution for mitigating representation drift.

error map	OpenSDID	SID-Set	Avg.
×	39.23	43.80	41.52
✓	43.65(+4.42)	56.85(+13.0)	50.25(+8.73)

Table 8: The impact of error map on localization.

Impact of Error Map Injection. Next, we validate the contribution of injecting the alignment error map M_e as a spatial cue into the localization decoder. As shown in Table 8, we compare the performance on the localization task with (✓) and without (×) using the error map. The results show that on both the SID-Set and OpenSDID datasets, the model using error map injection outperforms the baseline that does not use it (Avg. 50.25 vs 41.52). Specifically, our method achieves an average IoU of 50.25, surpassing the baseline's 41.52 by a commanding margin of 8.73% points. This confirms that the decoder is not merely treating this map as noise, but is effectively learning to leverage it as a strong spatial prior for fine-tuning the mask boundaries. Concurrently, this directly validates our hypothesis that the alignment error is not an abstract value but a potent spatial cue, providing critical guidance for the localization decoder.

Quantitative and Qualitative Analysis of Representation Space. To qualitatively validate our diagnosis, we present a T-SNE feature visualization in Figure 2. As diagnosed, the Real and Tampered representations of the baseline model are severely entangled. In contrast, our anchoring method effectively pulls the Real representation into a highly compact, low-variance cluster, distinctly

Model	Intra-class Var (↓)	Inter-class Dist (↑)
Baseline	3.5225	5.4550
AAP (**Ours**)	**2.9590**	**13.2344**

Table 9: Quantitative metrics for representation margin.

separated from the forgery category. To provide rigorous quantitative support for these visual observations, we further calculate the intra-class variance and inter-class distance for the Real and Tampered categories, as reported in Table 9. The quantitative metrics confirm that our Asymmetric Anchoring Paradigm effectively minimizes intra-class variance (from 3.52 to 2.96) while significantly enlarging the inter-class distance (from 5.46 to 13.23), explicitly expanding the decision boundary and resolving the representation drift.

4.5 Qualitative Results

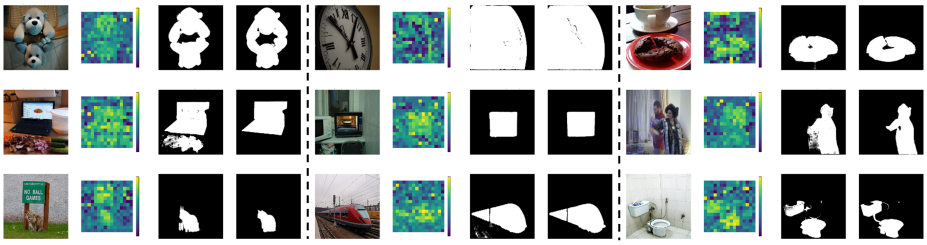

Fig. 4: Qualitative results on the SID dataset. From left to right: original tampered image, error heatmap, predicted mask, and ground truth mask.

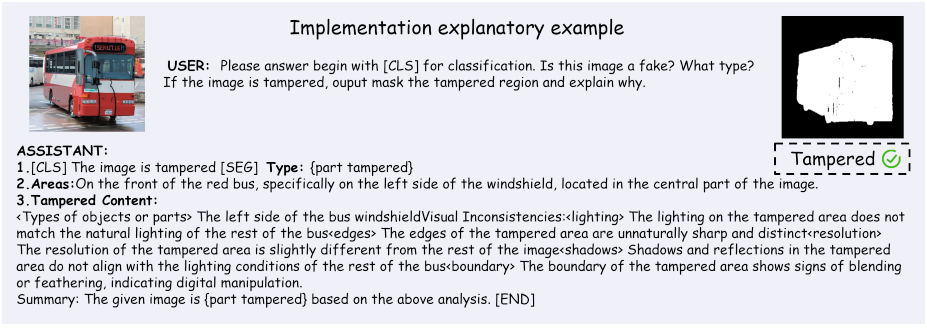

Fig. 5: Qualitative results of forgery detection and explanation on the SID-Set.

Finally, we present qualitative localization results on SID-Set in Figure 4. The error heatmap is particularly insightful. This map is a byproduct of the $\mathcal{L}_{arl}$ anchoring process, generated automatically without any pixel-level supervision. We can clearly see that the heatmap (bright areas) accurately highlights the tampered regions, while the authentic background remains dark. Guided by this precise cue, our final predicted mask is able to highly conform to the ground truth mask. Figure 5 demonstrates our model's explanatory capabilities. By stabilizing internal representations, Asymmetric Anchoring significantly reduces

visual-textual hallucinations, grounding MLLM reasoning in robust real-world priors. Implementation details are provided in the Supplementary Materials.

Error Map Dynamics and Failure Analysis. While the error map provides a crucial spatial prior, we also analyze its behavior in highly complex or failure cases, as illustrated in Figure 6. We observe that the raw alignment error map ($\mathcal{M}_e$) sometimes exceeds the exact tampered boundaries. This occurs because the alignment deviation constraint is computed at the feature level, which can be partially influenced by the MLLM's inherent object-level attention (e.g., highlighting an entire manipulated object rather than just the spliced edge). However, our AAP elegantly resolves this through the multi-modal fusion design: by synergizing the dense error map with the semantic `<SEG>` prompt (Eq. 7), the localization decoder effectively filters out these perceptual overflows, ensuring the final predicted mask remains highly precise even when the raw error map contains noise.

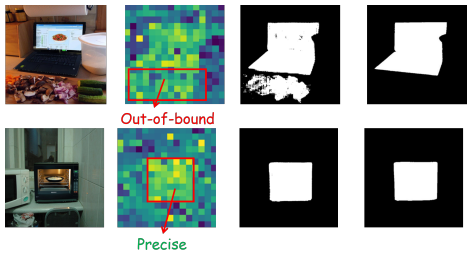

Fig. 6: Failure case and error map dynamics analysis. The raw error map may sometimes overflow due to object-level attention, but the synergy with the semantic prompt filters this noise to yield precise masks.

5 Conclusion

We identify Representation Drift as a critical bottleneck suppressing MLLM performance on authentic images. To resolve this, we propose the Asymmetric Anchoring Paradigm (AAP), an open-box framework repurposing the visual encoder as a Truth Anchor. Asymmetrically aligning authentic images to this anchor successfully stabilizes the real-world prior, while the byproduct alignment error map provides a precise spatial cue for localization. Comprehensive evaluations on SID-Set and OpenSDID confirm AAP's state-of-the-art performance in both forgery detection and localization.

Acknowledgements

This work was supported by Natural Science Foundation of China (No. 62506030, U24B20179, 62476021), Beijing Natural Science Foundation (No. L242021), and the Open Project of Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University (No. MMC202406).

References

1. Bi, X., Liu, B., Yang, F., Xiao, B., Li, W., Huang, G., Cosman, P.C.: Detecting generated images by real images only. Arxiv (2023)
2. Chang, Y., Yeh, C., Chiu, W., Yu, N.: Antifakeprompt: Prompt-tuned vision-language models are fake image detectors. Arxiv (2023)
3. Chen, X., Dong, C., Ji, J., Cao, J., Li, X.: Image manipulation detection by multi-view multi-scale supervision. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14185–14193 (2021)
4. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See https://vicuna. lmsys. org (accessed 14 April 2023) **2**(3), 6 (2023)
5. Choi, J., Kim, T., Jeong, Y., Baek, S., Choi, J.: Exploiting style latent flows for generalizing deepfake video detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
6. Cui, X., Li, Y., Luo, A., Zhou, J., Dong, J.: Forensics adapter: Adapting clip for generalizable face forgery detection. arXiv preprint arXiv:2411.19715 (2024)
7. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems **36**, 49250–49267 (2023)
8. Dong, C., Chen, X., Hu, R., Cao, J., Li, X.: Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. IEEE TPAMI (2023)
9. Dong, S., Wang, J., Liang, J., Fan, H., Ji, R.: Explaining deepfake detection by analysing image matching pp. 18–35 (2022)
10. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first International Conference on Machine Learning (2024)
11. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition. In: ICML (2020)
12. Frank, J., Holz, T.: Cnn-generated images are surprisingly easy to spot...for now. Arxiv (2021)
13. Goljan, M., Chen, M., Comesaña, P., Fridrich, J.: Effect of compression on sensor-fingerprint based camera identification. Electronic Imaging **28**, 1–10 (2016)
14. Guillaro, F., Cozzolino, D., Sud, A., Dufour, N., Verdoliva, L.: Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20606–20615 (2023)
15. Guo, X., Liu, X., Ren, Z., Grosz, S., Masi, I., Liu, X.: Hierarchical fine-grained image forgery detection and localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3155–3165 (2023)
16. Guo, X., Liu, X., Ren, Z., Grosz, S., Masi, I., Liu, X.: Hierarchical fine-grained image forgery detection and localization. In: CVPR (2023)
17. Guo, X., Song, X., Zhang, Y., Liu, X., Liu, X.: Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 105–116 (2025)
18. Huang, Z., Hu, J., Li, X., He, Y., Zhao, X., Peng, B., Wu, B., Huang, X., Cheng, G.: Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28831–28841 (2025)

19. Jiang, Z., Chen, J., Zhu, B., Luo, T., Shen, Y., Yang, X.: Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25004–25014 (2025)
20. Ju, Y., Jia, S., Ke, L., Xue, H., Nagano, K., Lyu, S.: Fusing global and local features for generalized ai-synthesized image detection. In: ICIP (2022)
21. Kaduri, O., Bagon, S., Dekel, T.: What's in the image? a deep-dive into the vision of vision language models, 2024. URL https://arxiv. org/abs/2411.17491
22. Kang, S., Kim, J., Kim, J., Hwang, S.J.: Your large vision-language model only needs a few attention heads for visual grounding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9339–9350 (2025)
23. Kee, E., O'brien, J.F., Farid, H.: Exposing photo manipulation from shading and shadows. ACM Transactions on Graphics **33**(5), 1–21 (2014)
24. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
25. Koutlis, C., Papadopoulos, S.: Leveraging representations from intermediate encoder-blocks for synthetic image detection. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 394–411. Springer Nature Switzerland, Cham (2025)
26. Kwon, M.J., Lee, W., Nam, S.H., Son, M., Kim, C.: Safire: Segment any forged image region. In: Proceedings of the AAAI conference on artificial intelligence. vol. 39, pp. 4437–4445 (2025)
27. Kwon, M.J., Nam, S.H., Yu, I.J., Lee, H.K., Kim, C.: Learning jpeg compression artifacts for image manipulation detection and localization. International Journal of Computer Vision **130**(8), 1875–1895 (2022)
28. Labs, B.F.: Flux.1: A 12 billion parameter rectified flow transformer for text-to-image generation. `https://github.com/black-forest-labs/flux` (2024), a 12 billion parameter rectified flow transformer capable of generating images from text descriptions
29. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9579–9589 (2024)
30. Li, J., Xie, H., Li, J., Wang, Z., Zhang, Y.: Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
31. Lin, L., He, X., Ju, Y., Wang, X., Ding, F., Hu, S.: Preserving fairness generalization in deepfake detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16815–16825 (2024)
32. Liu, F., Ye, M., Du, B.: Learning a generalizable re-identification model from unlabelled data with domain-agnostic expert. Visual Intelligence **2**(1), 28 (2024)
33. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
34. Liu, J., Zhang, F., Zhu, J., Sun, E., Zhang, Q., Zha, Z.J.: Forgerygpt: Multimodal large language model for explainable image forgery detection and localization. arXiv preprint arXiv:2410.10238 (2024)
35. Liu, X., Liu, Y., Chen, J., Liu, X.: Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. IEEE Transactions on Circuits and Systems for Video Technology **32**(11), 7505–7517 (2022)
36. Liu, Z., Qi, X., Torr, P.H.S.: Global texture enhancement for fake face detection in the wild. In: CVPR (2020)

37. Lou, Z., Cao, G., Guo, K., Yu, L., Weng, S.: Exploring multi-view pixel contrast for general and robust image forgery localization. IEEE Transactions on Information Forensics and Security (2025)
38. Luo, A., Kong, C., Huang, J., Hu, Y., Kang, X., Kot, A.C.: Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection. IEEE Transactions on Information Forensics and Security (2023)
39. Ma, X., Du, B., Liu, X., Hammadi, A.Y.A., Zhou, J.: Iml-vit: Image manipulation localization by vision transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence (2024)
40. Nguyen, D., Mejri, N., Singh, I.P., Kuleshova, P., Astrid, M., Kacem, A., Ghorbel, E., Aouada, D.: Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
41. Niloy, F.F., Bhaumik, K.K., Woo, S.S.: Cfl-net: Image forgery localization using contrastive learning. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 4642–4651 (2023)
42. Ojha, U., et al.: Towards universal fake image detectors that generalize across generative models. In: CVPR. pp. 24480–24489 (2023)
43. Pan, S., Zhang, Z., Wei, K., Yang, X., Deng, C.: Few-shot generative model adaptation via style-guided prompt. IEEE Transactions on Multimedia (2024)
44. Peng, S., Wang, Z., Gao, L., Zhu, X., Zhang, T., Liu, A., Zhang, H., Lei, Z.: Mllm-enhanced face forgery detection: A vision-language fusion solution. arXiv preprint arXiv:2505.02013 (2025)
45. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=di52zR8xgf
46. Qiao, T., Xie, S., Chen, Y., Retraint, F., Luo, X.: Fully unsupervised deepfake video detection via enhanced contrastive learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(7), 4654–4668 (2024)
47. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)
48. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
49. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
50. Shi, Y., Gao, Y., Lai, Y., Wang, H., Feng, J., He, L., Wan, J., Chen, C., Yu, Z., Cao, X.: Shield: An evaluation benchmark for face spoofing and forgery detection with multimodal large language models. Visual Intelligence **3**(1), 9 (2025)
51. Shiohara, K., Yamasaki, T.: Detecting deepfakes with self-blended images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18720–18729 (2022)
52. Smeu, S., Oneata, E., Oneata, D.: Declip: Decoding clip representations for deepfake localization pp. 149–159 (2025)
53. Smeu, S., Oneata, E., Oneata, D.: Declip: Decoding clip representations for deepfake localization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2025)

54. Tan, C., Liu, H., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection (2023)
55. Tan, C., Tao, R., Liu, H., Gu, G., Wu, B., Zhao, Y., Wei, Y.: C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection **39**(7), 7184–7192 (2025)
56. Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: Frequency-aware deepfake detection: Improving generalizability through frequency space learning (2024)
57. Tan, C., Zhao, Y., Wei, S., Gu, G., Wei, Y.: Learning on gradients: Generalized artifacts representation for gan-generated images detection. In: CVPR (2023)
58. Tao, R., Le, M., Tan, C., Liu, H., Qin, H., Zhao, Y.: Oddn: Addressing unpaired data challenges in open-world deepfake detection on online social networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 799–807 (2025)
59. Tao, R., Qin, Z., Ding, Y., Tan, C., Wang, J., Wang, W.: Unlocking the potential of lightweight quantized models for deepfake detection. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI). pp. 520–528 (2025)
60. Tao, R., Tan, C., Liu, H., Wang, J., Qin, H., Chang, Y., Wang, W., Ni, R., Zhao, Y.: Sagnet: Decoupling semantic-agnostic artifacts from limited training data for robust generalization in deepfake detection. IEEE Transactions on Information Forensics and Security (2025)
61. Tao, R., Tang, S., Qin, H., Wang, W., Wei, Y., Zhao, Y.: Lednet: a multimodal foundation model for robust deepfake detection. Science China Information Sciences **68**(6), 160106 (2025)
62. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
63. Wang, J., Wu, Z., Chen, J., Han, X., Shrivastava, A., Lim, S.N., Jiang, Y.G.: Objectformer for image manipulation detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2364–2373 (2022)
64. Wang, Y., Huang, Z., Hong, X.: Opensdi: Spotting diffusion-generated images in the open world. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4291–4301 (2025)
65. Wang, Z., Hu, S., Zhao, S., Lin, X., Juefei-Xu, F., Li, Z., Han, L., Subramanyam, H., Chen, L., Chen, J., et al.: Mllm-as-a-judge for image safety without human labeling. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14657–14666 (2025)
66. Xu, Z., Zhang, X., Li, R., Tang, Z., Huang, Q., Zhang, J.: Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. arXiv preprint arXiv:2410.02761 (2024)
67. Yan, Z., Luo, Y., Lyu, S., Liu, Q., Wu, B.: Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
68. Yan, Z., Wang, J., Jin, P., Zhang, K.Y., Liu, C., Chen, S., Yao, T., Ding, S., Wu, B., Yuan, L.: Orthogonal subspace decomposition for generalizable ai-generated image detection. arXiv preprint arXiv:2411.15633 (2024)

69. Yang, F., Zhen, R., Wang, J., Zhang, Y., Chen, H., Lu, H., Zhao, S., Ding, G.: Heie: Mllm-based hierarchical explainable aigc image implausibility evaluator. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3856–3866 (2025)
70. Yermakov, A., Cech, J., Matas, J.: Unlocking the hidden potential of clip in generalizable deepfake detection. arXiv preprint arXiv:2503.19683 (2025)
71. Yoon, H., Jung, J., Kim, J., Choi, H., Shin, H., Lim, S., An, H., Kim, C., Han, J., Kim, D., et al.: Visual representation alignment for multimodal large language models. arXiv preprint arXiv:2509.07979 (2025)
72. Zhang, S., Yang, Y., Zhou, Z., Sun, Z., Lin, Y.: Dibad: A disentangled information bottleneck adversarial defense method using hilbert-schmidt independence criterion for spectrum security. IEEE Transactions on Information Forensics and Security (2024)
73. Zhang, Z., Yadav, S., Han, F., Shutova, E.: Cross-modal information flow in multimodal large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19781–19791 (2025)