REVEAL – Reasoning and Evaluation of Visual Evidence through Aligned Language

The proliferation of powerful diffusion-based generators has made image forgeries increasingly indistinguishable from authentic content. Synthetic images threaten news, creative industries, and cybersecurity by amplifying misinformation and eroding trust in digital media. Traditional forgery detection pipelines often rely on supervised training for a specific manipulation, which limits their adaptability when new types of generative models emerge. While multimodal large language models offer scene-level reasoning and natural language explanations, existing fine-tuned solutions such as FakeShield [1] are locked to specific datasets. We reframe forgery detection as a prompt-driven reasoning task for vision–language models (VLMs), emphasizing generalisation, interpretability, and zero-shot performance.

We design REVEAL (Reasoning and Evaluation of Visual Evidence through Aligned Language) to probe forgery cues through structured prompts systematically. In our experiments, we benchmark VLMs on Photoshop tampering (CASIA1+, Columbia, IMD2020), DeepFake (FFHQ, FaceApp, Seq-DeepFake), and AI-generated content (AIGC-Editing) datasets. We assess open VLMs including LLaVA, GPT-4.1, GPT-4o, and Gemini 2.5-pro. We establish a binary baseline as well as propose 2 alternative approaches from first principles. The binary baseline uses a zero-shot prompt ("Is this image real or fake?") for classification. *Approach I* introduces a holistic prompt strategy where each image is evaluated across eight forensic dimensions: lighting/shadow, reflections, perspective/geometry, repetition, edge consistency, semantic coherence, anomaly/artifacts, and realism of human/object rendering. Each factor is rated on a Likert scale (1=authentic to 5=forged), followed by concise justifications that accumulate into a global tampering score. *Approach II* leverages region-wise prompting by overlaying a 3×3 labeled grid over images. Models must provide local anomaly reasoning per cell before synthesizing a global judgment. This design draws on insights from Set-of-Mark prompting [2], which demonstrates that explicit overlays outperform imagined spatial partitioning for visual grounding tasks.

We observe that not only structured prompting significantly outperforms baseline binary classification but also performs similar to finetuned baselines such as Fakeshield. For example, GPT-4.1 F1-score on CASIA1+ improves from 0.80 in the baseline to 0.92 in both approaches and also comes close to 0.95 in Fakeshield. Gemini's F1-score on the Columbia dataset rises from 0.39 (baseline) to 0.85 under holistic prompting. Performance gains are most pronounced on DeepFake (matching Fakeshield F1 of 0.93) and AIGC datasets, where manipulations are subtle or spatially localized, demonstrating the advantage of explicit prompt scaffolding. In Approach II, Region-wise prompts excel at detecting local splicing or synthetic insertions, while holistic evaluation in Approach I captures scene-wide inconsistencies such as mismatched shadows or semantic contradictions. We plot the ROC curves of the global tampering score obtained in Approach I for threshold-independent comparison between different VLMs.

REVEAL highlights how structured prompting enables scalable, interpretable, and domain-agnostic forgery detection. By reasoning from first principles, VLMs adapt across manipulation types without retraining. Beyond metrics, the method provides human-interpretable rationales, offering forensic transparency and explanability. Future work will integrate REVEAL with segmentation and activation models like Grad-CAM, SAM etc. to further improve localization, enabling fine-grained attribution of tampered regions.

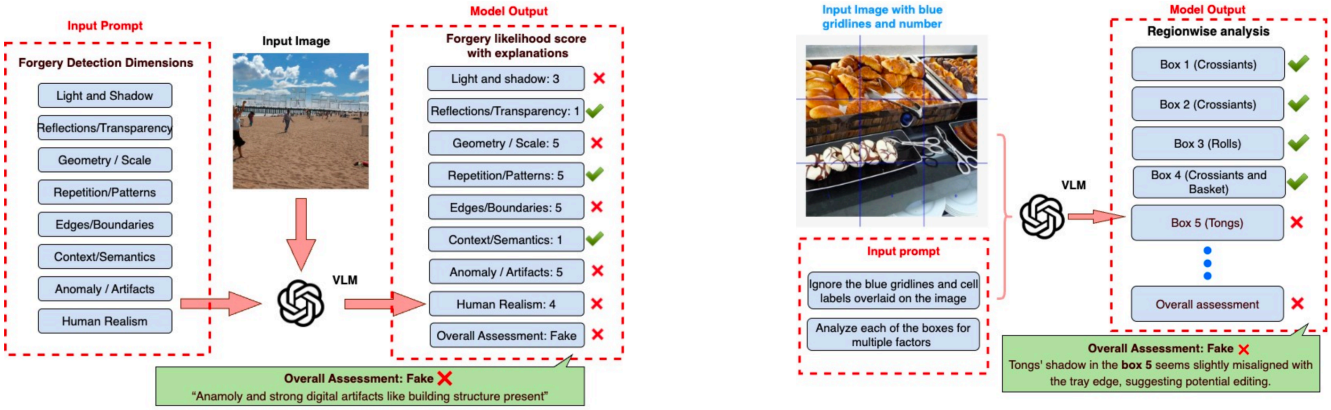

Figure: (L) Holistic scene-level evaluation across 8 forensic dimensions with Likert scoring. (R) Region-wise anomaly detection using a 3×3 labelled grid for localised inconsistencies

[1] Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. Fakeshield: Explainable image forgery detection and localisation via multi-modal large language models, 2025

[2] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4v, 2023