TruthLens: Training-Free Data Verification for Deepfake Images via VQA-style Probing

Ritabrata Chakraborty [1] **Rajatsubhra Chakraborty** [2] **Ali Khaleghi Rahimian** [2]

Abstract

AI-generated imagery is on the rise to be outplacing our human ability to spot manipulations. Prevailing deepfake detectors cast as opaque binary classifiers offer little to no insights into their decisions. We introduce **TruthLens**, a training-free framework that reframes deepfake detection as a VQA task. We leverage large vision-language models (LVLMs) to reveal artifacts and GPT-4 to reason over the evidence to reach a coherent verdict by fusing visual and semantic cues. The framework explains which artifacts triggered its judgement, providing a deeper and newer mode of transparency. Evaluations demonstrate that TruthLens outperforms conventional methods while maintaining a strong emphasis on explainability and delivering instance-level data verification for large-scale generative models.

1. Introduction

> ***Every photograph is a fiction with pretensions to truth.***
>
> **Joan Fontcuberta**

The proliferation of manipulated and synthetic images, driven by advancements in generative models such as modern Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) such as StyleGAN (Karras et al., 2019) and diffusion models (Ho et al., 2020), has created significant challenges in distinguishing real from fake images. This has enabled the creation of highly photorealistic images, which are increasingly used in contexts ranging from entertainment to malicious disinformation campaigns. This surge has raised critical concerns in domains such as media integrity, cybersecurity, and ethical AI.

Traditional methods for detecting fake images rely heavily on binary classifiers, such as CNNDetection (Frankle et al., 2020), which uses pre-trained Convolutional Neural Networks (CNNs) like ResNet-50 to identify specific artifacts in GAN-generated images. While effective for early GAN models, these methods struggle with newer architectures that exhibit fewer detectable flaws. More recently, approaches like Diffusion Reconstruction Error (DIRE) (Zhou et al., 2023) have shown promise by leveraging the reconstruction inconsistencies of diffusion models to detect synthetic content. However, these methods focus solely on classification, offering little interpretability or insight into why an image is labeled as fake. Recent studies suggest that incorporating reasoning into classification tasks can enhance transparency and user trust (Ribeiro et al., 2016; Selvaraju et al., 2017). For instance, the LIME framework(Ribeiro et al., 2016) provides local explanations for machine learning predictions. Similarly, methods like Grad-CAM (Selvaraju et al., 2017) visualize important regions in images for CNN-based classifiers, but they are not inherently designed to address the complexities of synthetic image detection.

Inspired by advances in Large Vision-Language Models (LVLMs), such as LLaVA (Liu et al., 2023) and BLIP-2 (Jiang et al., 2023), we rethink the task of fake image detection as a multimodal Visual Question Answering (VQA) problem. We introduce TruthLens, an LVLM + LLM framework that bridges the gap between detection accuracy and interpretability for this task. By combining the detection capabilities of traditional models with the reasoning power of multimodal systems, TruthLens not only classifies images as real or fake but also provides detailed justifications for its decisions by leveraging both visual and textual features. The framework incorporates a structured pipeline that integrates multimodal querying, textual aggregation, and reasoning to deliver transparent and robust results[1].

Model-centric safety audits often treat training data as an af-

[1]Department of Computer Science and Engineering, Manipal University Jaipur, Jaipur, India [2]Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC, USA. Correspondence to: Ritabrata Chakraborty <ritabrata.229301716@muj.manipal.edu>.

Published at Data in Generative Models Workshop: The Bad, the Ugly, and the Greats (DIG-BUGS) at ICML 2025, Vancouver, Canada.

[1]For improved readability, we refer extensively to the following terms by their abbreviations: Visual Question Answering (**VQA**), Large Vision-Language Model (**LVLM**), and Large Language Model (**LLM**).

terthought (Papernot et al., 2018); we posit that trustworthy AI begins with data-centric safety and verification. By reframing deep-fake detection as a post-hoc data-verification task, TruthLens shows how LVLMs can audit each synthetic output without retraining.

The contributions of TruthLens are twofold:

We reframe fake image detection as a VQA task using LVLMs to find key visual artifacts and generate natural language explanations for classification decisions.

TruthLens introduces a novel pipeline that combines multimodal prompting, response aggregation, and verdict synthesis through an external LLM. Unlike prior detection systems, our approach requires no additional fine-tuning and delivers fully explainable outputs via modular, interpretable stages.

2. TruthLens

TruthLens leverages multimodal reasoning to classify images and provide detailed justifications for its decisions. The pipeline consists of four main steps: (1) Question Generation, (2) Multimodal Reasoning, (3) Textual Aggregation, and (4) Final Decision Making.

2.1. Step 1: Question Generation

The first step in the pipeline involves generating a set of predefined prompts or questions that address specific visual and textual cues in an image. These prompts are carefully designed to probe various aspects of image authenticity, such as artifacts, inconsistencies, or visual features commonly associated with synthetic images. The set of prompts, denoted as $P = \{p_1, p_2, \ldots, p_N\}$, consists of N individual prompts, each corresponding to a specific artifact or visual clue. Prompts are crafted based on known patterns in synthetic images, such as lighting inconsistencies, unnatural textures, or boundary artifacts. By systematically querying the input image I with these prompts, the framework aims to extract detailed responses that highlight evidence supporting the classification task. The detailed prompt categories are highlighted in Appendix B.

These prompts are designed to capture flaws and artifacts typical to most deepfakes as best as possible. Unlike GAN-generated images, most fake images nowadays do not share the underlying artifacts that make them easy to spot (Frankle et al., 2020), and so we must prompt the model to consider more global visual features. Modern day deepfakes are much more sophisticated than they once were, but still have several common visual abnormalities that give them away: errors in texture, lighting, and anatomy are still commonplace in many current day deepfakes (Kamali et al., 2024). Prompting the model to focus on these visual abnormalities will give it the best chance of detecting fake images. Each prompt p_i acts as an instance-level verification probe that tests the input against known natural-image priors rather than against a closed classifier.

2.2. Step 2: Multimodal Reasoning

In the second step, a multimodal model, denoted as f_{MM}, processes the input image I alongside each prompt $p_i \in P$ to generate answers. The model combines visual and textual modalities to produce meaningful explanations. Formally, the multimodal model can be represented as $f_{\text{MM}}(I, p) : \mathcal{I} \times \mathcal{P} \to \mathcal{A}$, where $\mathcal{I}$ is the space of input images, $\mathcal{P}$ is the space of textual prompts, and $\mathcal{A}$ is the space of textual answers. For each prompt p_i, the model generates an answer $a_i = f_{\text{MM}}(I, p_i)$, resulting in a complete set of answers $A = \{a_1, a_2, \ldots, a_N\}$.

The multimodal model extracts visual features such as textures, edges, and artifacts using vision encoders and contextualizes these features with respect to the prompt using text encoders. It then generates natural language answers combining visual and textual evidence. By leveraging state-of-the-art models like LLaVA and BLIP-2, this step ensures that the extracted answers are both precise and interpretable.

2.3. Step 3: Textual Aggregation

Once the answers are generated, the next step involves aggregating these responses into a structured summary S that encapsulates the key observations from all prompts. This structured summary organizes the raw outputs into a coherent explanation. The aggregation process is represented as $g : \mathcal{A}^N \to \mathcal{S}$, where $\mathcal{S}$ is the space of structured summaries. The summary S is computed as $S = g(A) = g(f_{\text{MM}}(I, p_1), f_{\text{MM}}(I, p_2), \ldots, f_{\text{MM}}(I, p_N))$.

The goal of this step is to consolidate the multimodal reasoning into a concise, human-readable format that provides the foundation for the final classification and reasoning.

2.4. Step 4: Final Decision Making

In the final step, the structured summary S is passed to a language model f_{LM} for classification and reasoning. The language model determines whether the image is real or fake and generates a natural language explanation for its decision. This process can be formulated as $f_{\text{LM}}(S) : \mathcal{S} \to \mathcal{Y} \times \mathcal{R}$, where $\mathcal{Y} = \{\text{Real}, \text{Fake}\}$ represents the space of classification labels, and $\mathcal{R}$ represents the space of textual justifications. The final output is $(y, r) = f_{\text{LM}}(S)$, where y is the classification result (Real or Fake) and r is the explanation for the decision.

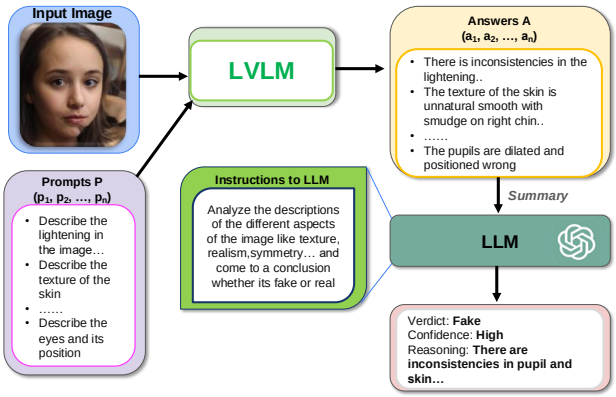

Figure 1. Overview of the detection pipeline used in the TRUTHLENS framework.

3. Experiments and Discussion

We evaluate the performance of TruthLens on various datasets and compare it against existing state-of-the-art methods, including CNNDetection and DIRE. Our experiments are designed to measure detection accuracy, reasoning quality, and robustness across different types of fake images.

3.1. Datasets

LDM Consists of 1000 fake images generated by Latent Diffusion Models (LDM) alongside corresponding 1000 real images from FFHQ (Karras et al., 2018b). This dataset challenges models with high-quality synthetic images that closely mimic real-world distributions.

ProGAN Includes 1000 fake images generated by ProGAN derived from ForgeryNet dataset (He et al., 2021) . ProGAN's images exhibit traditional GAN artifacts, making this dataset suitable for evaluating the performance of models on GAN-based generation techniques.

Figure 2. Overview of the evaluation dataset on the left hand side we have Real images from FFHQ dataset (Karras et al., 2018b) and on the right we have ProGAN generated images from ForgeryNet dataset (He et al., 2021) and Latent Diffusion Model(LDM) (Rombach et al., 2021) generated images.

3.2. Metrics

We report results using several key metrics. Accuracy measures the percentage of correct classifications, whether the data is real or fake. AUC (Area Under the ROC Curve) indicates the model's ability to distinguish between classes. Additionally, we evaluate Precision, Recall, and F1-Score, which offer a comprehensive understanding of the balance between true positives and false negatives. Qualitative Analysis includes visualizations and reasoning quality to assess the interpretability of the results.

3.3. Verification Results

Comparison of AUC Scores. The AUC scores of various methods across LDM and ProGAN datasets are shown in Table 1. Our framework achieves superior performance compared to CNNDetection and DIRE, demonstrating its robustness across diverse generation techniques.

Table 1. Comparison of AUC scores across datasets generated by LDM and ProGAN.

Method	LDM (%)	ProGAN (%)
DIRE	46.47	58.12
CNNDetection	86.50	40.44
TruthLens (Ours)	**95**	**97.5**

Impact of Prompt + LLM in classification. Table 2 compares the classification accuracy of different models on real and fake images (LDM and ProGAN), with and without the use of prompts and the language model (LLM). The inclusion of prompts and LLM significantly enhances detection performance, especially for challenging LDM datasets.

Table 2. Classification results for real and fake data (LDM and ProGAN) using different models, with and without Prompt + LLM.

Method	Model	Real (%)	Fake (%)	
			LDM	ProGAN
Yes or No Question	BLIP2	52	20	50
	CogVLM	62	25	52
	LLAVA 1.5	50	22	48
	ChatUniVi	54	24	52
Prompts + LLM	BLIP2	74	68	72
	CogVLM	85	82	90
	LLAVA 1.5	76	70	74
	ChatUniVi	**98**	**92**	**97**

Performance Breakdown by Metric. Table 3 provides a detailed breakdown of model performance, including precision, recall, and F1-score for LDM and ProGAN datasets. These metrics highlight the strengths of our framework in accurately identifying fake images across different generation techniques. We explore ablation insights and feature-

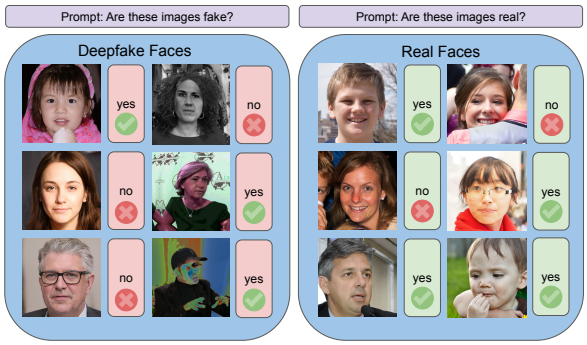

Figure 3. Visualization of the yes/no prompts supplied to LVLMs and their responses.

specific performance in Appendix C.

Table 3. Performance breakdown of different methods on LDM and ProGAN datasets. TruthLens demonstrates superior precision, recall, and F1-score across both datasets.

Method	Precision (%)		Recall (%)		F1-Score (%)	
	LDM	ProGAN	LDM	ProGAN	LDM	ProGAN
DIRE	49.97	49.67	**98.90**	**97.70**	66.40	65.86
CNNDetection	**97.59**	60.00	8.10	0.30	14.96	0.60
TruthLens (Ours)	90.99	**90.16**	90.57	96.28	**95.55**	**95.91**

Qualitative Analysis. In addition to quantitative results, we conducted qualitative analyses to assess the interpretability of our model's decisions. Figure 3 illustrates verdicts generated by TruthLens, highlighting interpretibility that contributed to the classification. A detailed outlook into model outputs and justifications into deepfake detection verdicts is shown in 4 of Appendix. These demonstrate the framework's ability to identify subtle artifacts in fake images.

3.4. Discussion

The success of ChatUniVi and Large Language Models (LLMs) within the TruthLens framework lies in their ability to integrate and reason across visual and textual modalities effectively. ChatUniVi excels by unifying image and video tokens into a shared representation, enabling a holistic understanding of visual artifacts and their contextual significance. This unified processing enhances the detection of subtle patterns, such as lighting inconsistencies and unnatural textures, which are often overlooked by traditional models. When paired with the reasoning capabilities of LLMs, ChatUniVi elevates detection accuracy and provides interpretable justifications for its decisions. The natural language explanations foster transparency and user trust, addressing critical challenges in high-stakes domains like media integrity and ethical AI.

Advantages. By eliminating the dependency on task-specific training, this approach ensures adaptability to emerging generative techniques without requiring large annotated datasets. This paradigm accelerates deployment and reduces computational overhead, making it highly scalable for various applications. Moreover, the reliance on pre-trained state-of-the-art models, like ChatUniVi and LLaVA, leverages their extensive training on diverse data, enabling the framework to detect artifacts in novel scenarios with minimal resource usage. Crucially, this paradigm shifts the focus to interpretable outputs by reframing detection as a Visual Question Answering (VQA) task, offering detailed natural language justifications that enhance trust and accountability. These attributes position TruthLens as a transformative solution for combating synthetic media, combining cutting-edge detection performance with transparency and user-centric design.

Limitations & Outlook. Our current study benchmarks TruthLens on two face-centric datasets (ProGAN and LDM), leaving other high-value domains like scene-level images and video deepfakes unexplored. Extending the probe set and evaluation suite to these modalities is a natural next step and will let us quantify domain transfer more rigorously. In addition, the framework issues several LVLM queries and one LLM aggregation per image. While still training-free, this design introduces non-trivial inference latency and API cost. Ongoing work on prompt batching, answer caching, and lightweight distillation aims to cut runtime and make real-time deployment feasible.

4. Conclusion

TruthLens delivers state-of-the-art fake-image detection and transparent rationales, directly addressing a central mandate of technical AI governance: balancing accuracy with accountability. By pairing vision-language models with interpretable reasoning, the framework equips regulators, platforms, and civic auditors to verify claims, audit decision paths, and meet disclosure obligations envisioned in policies such as the EU AI Act and C2PA provenance guidelines. Its modular, training-free design eases deployment across domains and datasets, while human-readable explanations foster public trust. Future work includes extending the framework to handle emerging synthetic media types and exploring more advanced interpretability techniques to enhance user understanding.

Impact Statement

TruthLens advances technical AI governance by turning deep-fake detection from an opaque, accuracy-only exercise into an auditable, reasoning-centred process. Since the framework is training-free and built on publicly available LVLMs/LLMs, hence regulators, newsrooms, and civil-society auditors can reproduce our results, inspect the natural-language rationales, and contest misclassifications—directly supporting transparency mandates in emerging policies such as the EU AI Act (European Union, 2024) and C2PA provenance standards [2] . TruthLens stores no biometric embeddings beyond transient inference, preserving user privacy, and its explanations are designed to inform, not reveal, sensitive identity attributes. We recommend that platforms adopting our system pair it with provenance watermarks and public reporting dashboards so that technical safeguards are reinforced by clear governance mechanisms and independent oversight.

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Chen, Z., Xie, L., Chen, X., Zhang, L., and Tian, Q. Towards interpretable face manipulation detection. In *Proceedings of the European Conference on Computer Vision*, 2022.

Ekko, Z. et al. Fakebench: Benchmarking multimodal models for fake image detection and reasoning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 1689, 12 July 2024, 2024. Available at http://data.europa.eu/eli/reg/2024/1689/oj.

Frankle, J., Schwab, B., et al. Cnn-generated images are surprisingly easy to spot...for now. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 470–477, 2020.

[2]Coalition for Content Provenance and Authenticity (C2PA), *C2PA Technical Specification*, Version 1.3, April 2024. Available at https://c2pa.org/specifications/specifications/1.3.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

He, Y., Gan, B., Chen, S., Zhou, Y., Yin, G., Song, L., Sheng, L., Shao, J., and Liu, Z. Forgerynet: A versatile benchmark for comprehensive forgery analysis. *arXiv preprint arXiv:2103.05630*, 2021.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, pp. 6840–6851, 2020.

Jiang, Z. et al. Blip-2: Bootstrapped language-image pre-training. *arXiv preprint arXiv:2301.12597*, 2023.

Jin, P., Takanobu, R., Zhang, W., Cao, X., and Yuan, L. Chat-univi: Unified visual representation empowers large language models with image and video understanding, 2024. URL https://arxiv.org/abs/2311.08046.

Kamali, N., Nakamura, K., Chatzimparmpas, A., Hullman, J., and Groh, M. How to distinguish ai-generated images from authentic photographs, 2024. URL https://arxiv.org/abs/2406.08651.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2018a.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948, 2018b. URL http://arxiv.org/abs/1812.04948.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Li, Y. and Lyu, S. Interpretable deepfake detection: A survey. *ACM Computing Surveys*, 2023.

Liu, H., Lin, J., et al. Visual instruction tuning: Teaching large language models to see and describe. *arXiv preprint arXiv:2304.14485*, 2023.

Papernot, N., McDaniel, P., Sinha, A., and Wellman, M. P. Sok: Security and privacy in machine learning. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 399–414, 2018. doi: 10.1109/EuroSP.2018.00035.

Ribeiro, M. T., Singh, S., and Guestrin, C. "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2021.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Smith, J. et al. Cifake: A dataset of real and ai-generated synthetic images. *Papers with Code*, 2022. https://paperswithcode.com/dataset/cifake.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., Xu, J., Xu, B., Li, J., Dong, Y., Ding, M., and Tang, J. Cogvlm: Visual expert for pretrained language models, 2024. URL https://arxiv.org/abs/2311.03079.

Wang, Z., Zhao, H., Wang, Y., Wang, S., Li, H., and Shi, Y. Fakebench: A comprehensive benchmark for deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Zellers, R., Holtzman, A., West, P., and Choi, Y. Can language models detect deepfakes? *arXiv preprint arXiv:2305.13244*, 2023.

Zhou, T. et al. Diffusion reconstruction error: A metric for fake image detection. *Proceedings of the NeurIPS Conference on Artificial Intelligence*, pp. 1234–1245, 2023.

A. Related Work

The task of detecting synthetic images has evolved significantly with advancements in generative models. Early methods focused on binary classification using handcrafted features, while modern approaches leverage deep learning models, which excel at uncovering subtle artifacts in fake images. Additionally, advancements in multimodal systems, large language models, and vision-language models have opened new avenues for interpretable fake image detection.

A.1. Large Language Models

Large Language Models (LLMs) are characterized by their massive number of parameters and large-scale training corpora. Models such as GPT (Brown et al., 2020) and LLaMA (Touvron et al., 2023) exemplify this class of models and are renowned for their versatility and power. Using transformer architectures, LLMs process language in a manner that surpasses traditional methods, achieving human-like or near-human performance in natural language tasks.

Many LLMs, like GPT (Brown et al., 2020), are so large, that they can almost be considered a distillation of nearly all human knowledge. This size needs the support of a large training corpus, which is typically sourced from all across the internet. With such large models and training corpus, LLMs are considered to operate at a near-human level in ideal conditions. Additionally, the number of tasks they can be generalized to is quite high. Nowadays, many vision models make use of pre-trained LLMs in order to expand their own capabilities, as there is a great amount of overlap between many vision and language tasks.

A.2. Large Vision-Language Models

LLMs are not limited to language tasks; they have also been adapted to handle computer vision tasks. Vision-Language Models (VLMs) combine visual and textual encoders within a unified architecture, allowing them to perform both categories of tasks simultaneously. Notable examples include OpenAI's GPT-for-vision (GPT4V), which processes both images and text for multi-modal reasoning, and LLaVA (Liu et al., 2023), which integrates several powerful vision and language models to enhance understanding of complex instruction-following tasks. These models are well-known for their generalization capabilities and broad applicability, proving particularly effective for reasoning over multi-modal inputs.

Some notable models include Chat-UniVi (Jin et al., 2024), which merges image and video tokens into a single unified representation. This merging provides Chat-UniVi with the ability to understand both forms of media, and enhances its LLM capabilities (Jin et al., 2024). Other models, like BLIP-2 (Jiang et al., 2023) take a different approach to improving model capabilities, and focus on different pre-training strategies in order to teach the model better multi-modal representations of data. BLIP-2 specifically learns in two stages: a representation learning stage, using a frozen image encoder, and a generative learning stage, using a frozen LLM (Jiang et al., 2023). This "bootstrapping" of pre-trained models allows BLIP-2 to take advantage of powerful vision and language models, and combine their capabilities into one. Finally, CogVLM (Wang et al., 2024) modifies the typical frozen language and vision encoders, by adding an extra "expert layer" (Wang et al., 2024) inside the transformers. This extra layer is meant to connect the two normally separate encoders, integrating both visual and linguistic features into one.

A.3. Deepfakes and Deepfake Detection

The artificial generation and manipulation of human faces—commonly referred to as "deepfakes"—gained prominence with the release of StyleGAN (Karras et al., 2019). These deepfakes have grown increasingly realistic over time, making their detection a critical research area.

Early detection methods, such as CNNDetection (Frankle et al., 2020), employed pre-trained CNNs like ResNet-50 to classify real and fake images based on pixel-level inconsistencies. While effective for earlier GANs, the flaws exploited by these methods are less prevalent in modern GANs and diffusion-based image generation.

Diffusion Reconstruction Error (DIRE) (Zhou et al., 2023) was introduced as an alternative to binary classifiers. This method reconstructs input images using pre-trained diffusion models and calculates the difference between the input and reconstructed image. DIRE assumes that diffusion-generated images share a similar probability distribution, allowing it to effectively detect diffusion-based fakes. However, it struggles with images generated by models outside this distribution.

A.4. Datasets for Deepfake Detection

Several benchmark datasets have been developed to support the training and evaluation of deepfake detection models. *CIFAKE* (Smith et al., 2022) combines real and synthetic images derived from the CIFAR-10 dataset, providing a compact yet challenging benchmark. *CelebA-HQ Resized* (Karras et al., 2018a) features high-resolution images of celebrity faces, making it particularly useful for evaluating generative models. Furthermore, benchmarks like *FakeBench* (Ekko et al., 2023) include datasets such as FakeClass and FakeClue, designed to assess both detection accuracy and reasoning capabilities. Another notable dataset, *FakeQA*, offers over 40,000 question-answer pairs, enabling open-ended evaluation of multi-modal models in reasoning and detection tasks. Lastly, *ForgeryNet* (He et al., 2021) provides a large-scale dataset with over 2.9 million images and videos, encompassing diverse manipulation techniques to facilitate robust and comprehensive evaluations.

A.5. Interpretability in Machine Learning

Interpretability has become a crucial aspect in fostering trust in AI systems, particularly in the domain of deepfake detection. While general-purpose frameworks like LIME (Ribeiro et al., 2016) and Grad-CAM (Selvaraju et al., 2017) have proven effective in explaining model predictions across various tasks, there is a growing need for specialized interpretability methods tailored to the nuances of deepfake analysis.

Recent research has explored the potential of Large Language Models (LLMs) in detecting deepfakes. For instance, studies have investigated whether LLMs can distinguish between real and AI-generated content, leveraging their broad knowledge and contextual understanding (Zellers et al., 2023). These approaches aim to complement traditional image-based detection methods by analyzing textual and semantic inconsistencies that may be present in deepfake content.

In parallel, initiatives like FAKEBENCH have emerged to provide comprehensive evaluation frameworks for deepfake detection algorithms (Wang et al., 2023). FAKEBENCH offers a standardized platform for assessing the performance of various detection methods, including those that incorporate interpretability components. This benchmark not only evaluates detection accuracy but also considers the explainability of the models, addressing the critical need for transparent and trustworthy AI systems in combating digital misinformation.

The development of interpretable deepfake detection models presents unique challenges due to the sophisticated nature of modern forgery techniques (Li & Lyu, 2023). Researchers are working on adapting existing explainability methods and developing new ones that can effectively highlight the subtle artifacts and inconsistencies that characterize deepfakes (Chen et al., 2022). These specialized approaches aim to provide more precise and relevant explanations for deepfake detection decisions, potentially improving both the accuracy and trustworthiness of detection systems.

A.6. Large Vision-Language Models for Deepfake Detection

Recent advancements in vision-language models have demonstrated their potential for combining detection and reasoning. Models like BLIP-2 (Jiang et al., 2023) and LLaVA (Liu et al., 2023) excel at joint visual and textual understanding. FakeBench (Ekko et al., 2023) explores their applicability to deepfake detection, evaluating not only detection accuracy but also reasoning capabilities. However, these works primarily focus on benchmarking rather than developing end-to-end systems for detection and reasoning.

Building on these advancements, our proposed framework, TruthLens, integrates detection and interpretability by re-framing deepfake detection as a Visual Question Answering (VQA) task. By leveraging state-of-the-art vision-language models, TruthLens provides both accurate classification and detailed reasoning for image authenticity, addressing limitations in existing methods.

B. A detailed look into prompts for TruthLens

The prompts used are listed as follows:

- **Lighting and Shadows:** *"Describe the lighting in the image. Does it appear natural or does it show any inconsistencies, such as unrealistic shadows or lighting direction?"*
- **Texture and Skin Details:** *"Analyze the texture of the skin in this image. Does the skin appear to have natural imperfections like pores, wrinkles, or blemishes, or is it unnaturally smooth?"*

- **Symmetry and Proportions:** *"Describe the facial symmetry in the image. Are there any noticeable asymmetries in the eyes, nose, mouth, or face shape?"*
- **Reflections and Highlights:** *"Examine the reflections in the eyes or any shiny areas on the skin. Do they appear to be consistent with the environment, or do they seem artificial or inconsistent?"*
- **Facial Features and Expression:** *"Describe the facial expression in the image. Does it appear natural, or are there any signs of a forced or unnatural expression?"*
- **Facial Hair (if applicable):** *"If there is facial hair in the image, describe its appearance. Does it seem realistic in terms of texture, growth pattern, and interaction with the lighting?"*
- **Eyes and Pupils:** *"Describe the appearance of the eyes in the image. Do the pupils appear natural in size, shape, and positioning, or are there any abnormalities?"*
- **Background and Depth Perception:** *"Describe the background of the image. Does it seem well-integrated with the face in terms of depth, focus, and lighting, or does it appear artificially blurred or detached?"*
- **Overall Realism of the Face:** *"Taking into account the lighting, texture, symmetry, and other features, describe the overall realism of the face. Does it show any signs of being digitally manipulated or generated?"*

C. Ablation Studies

In this section, we analyze the impact of specific image features on the detection accuracy of synthetic images through targeted ablation studies. Table 4 presents the accuracy achieved when prompts are designed to focus on distinct categories of visual cues. Each category probes a specific aspect of the image, helping to identify patterns or inconsistencies that contribute to the model's overall performance.

The results indicate that certain features, such as "Eyes and Pupils" (82.5% accuracy) and "Facial Hair" (74.5% accuracy), provide the most reliable cues for distinguishing real and fake images. These features are likely less prone to generative model artifacts, making them critical for accurate classification. On the other hand, categories like "Texture and Skin Details" (54.5% accuracy) and "Overall Realism of the Face" (55.6% accuracy) exhibit lower accuracy, suggesting these aspects are either less distinctive or more challenging for models to evaluate effectively.

The study highlights the importance of leveraging feature-specific prompts to enhance the detection process. By identifying high-impact categories, future improvements can prioritize these areas, leading to more focused and efficient detection strategies. This analysis also underscores the need for diverse and comprehensive prompts to cover a wide range of potential artifacts in synthetic images.

Table 4. Accuracy of Image Features across Categories.

Prompt Category	**Accuracy (%)**
Lighting and Shadows	62.0
Texture and Skin Details	54.5
Symmetry and Proportions	67.8
Reflections and Highlights	66.6
Facial Features and Expression	67.0
Facial Hair	74.5
Eyes and Pupils	82.5
Background and Depth Perception	60.0
Overall Realism of the Face	55.6

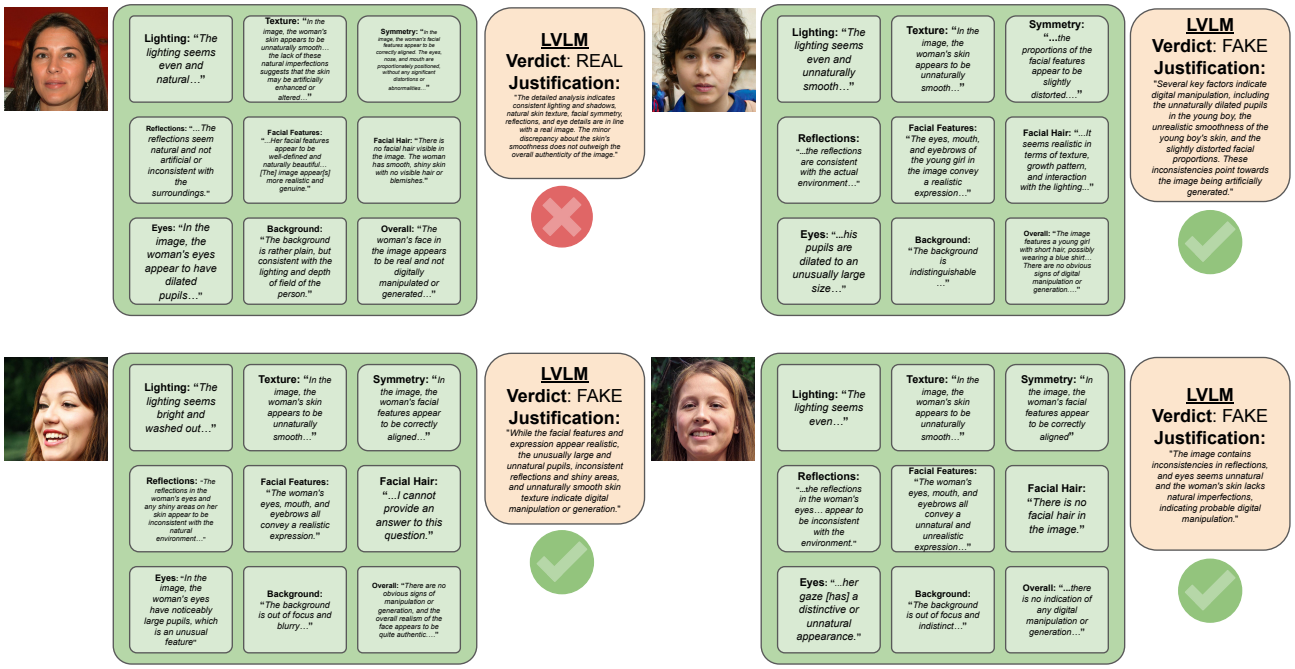

Figure 4. A visualization of the output of the model for each of the prompts, and the LVLM's final verdict on whether each sample is real or fake.