MTF-Bench: A Multi-Turn Forensic Benchmark for Explainable AI-Generated Image Detection

Anonymous Author(s)
Submission Id: 45

ACM Reference Format:
Anonymous Author(s). 2026. MTF-Bench: A Multi-Turn Forensic Benchmark for Explainable AI-Generated Image Detection. In . ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Dataset Composition

As shown in Figure 1, our dataset is systematically organized in terms of both content and generator types. The left pie chart illustrates the distribution of content across six major categories, each further divided into subcategories, capturing more fine-grained variations within broad thematic domains. This hierarchical content structure facilitates detailed statistical analyses, visualization, and experimental design. Meanwhile, the right pie chart presents the distribution of generator types used to create the dataset, providing insight into the diversity of generation methods. By adopting this dual categorization, we can better analyze the characteristics of the data and support diversified algorithm evaluation and comparative studies.

2 Details of Baselines

We compare MTF tuning with state-of-the-art AI-generated image detection methods, including CNNSpot [4], UnivFD [2], NPR [3], HyperDet [1], AIDE [5], VIB-Net [6], and the single-turn MLLM-based forensic method AIGI-Holmes [7]. For a fair and controlled comparison, all baseline methods are retrained using their official implementations under identical experimental settings and datasets. Below we briefly introduce each baseline method and its representative forensic principle.

CNNSpot [4]. A CNN-based detector that learns generation-specific visual patterns via supervised training and targeted data augmentations; it primarily exploits low- to mid-level pixel and texture artifacts characteristic of early generative models.

UnivFD [2]. A CLIP-driven approach that leverages pre-trained multimodal representations to improve zero-shot and cross-domain detection; by aligning visual features with semantic concepts, it attains strong generalization to unseen generators and domains.

NPR [3]. A method that focuses on local pixel correlations and neighborhood priors to identify subtle synthesis artifacts; NPR emphasizes robust, low-level cues that are less sensitive to semantic variations.

HyperDet [1]. A hybrid detector that fuses high-frequency (e.g., spectral or high-pass) signals with semantic representations, aiming to combine texture-level forensic evidence with higher-level context for improved robustness.

AIDE [5]. A frequency-aware synthetic image detector that combines patchwise DCT-based selection with residual feature extraction and semantic embeddings. AIDE explicitly separates high- and low-frequency patches and aggregates their statistics via an SRM-ResNet backbone, while incorporating CLIP-based semantic features as auxiliary cues.

VIB-Net [6]. A detector that incorporates an information bottleneck or representation-compression principle to discard generator-specific noise and learn compact, generalizable features, thereby enhancing cross-domain robustness.

AIGI-Holmes [7]. A single-turn MLLM-based forensic method that produces an authenticity judgment accompanied by a natural-language explanation; it represents the conventional one-shot vision–language fine-tuning paradigm against which we compare multi-turn forensic tuning.

3 Experimental Setup

All experiments are conducted on the MTF-Bench dataset. We select 58k images as the training set, comprising 29k synthesized images and 29k real images. The models are trained for 10 epochs with a batch size of 1, using full fine-tuning on 8 A800 80G GPUs. For the binary classification model, training is performed solely on image data, without any text-based fine-tuning.

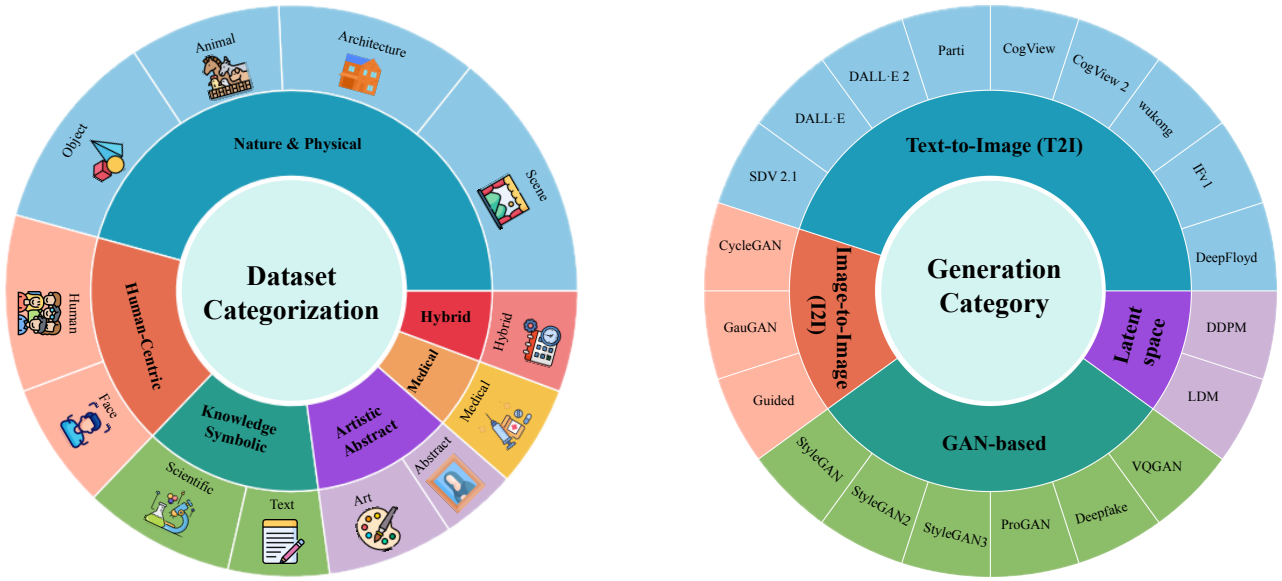
For optimization, we adopt the AdamW optimizer with a base learning rate of 1e-5 and a warmup ratio of 0.05 to mitigate instability during the initial training iterations. Throughout training, gradient accumulation and gradient clipping are applied to enhance the stability of large-scale model updates, and a linear learning rate decay schedule is employed for the remainder of the training process.

4 Evaluation Metrics

Following existing research in AI-generated image detection, we adopt classification accuracy (ACC) as the primary evaluation metric. Accuracy is defined as the ratio of correctly classified samples to the total number of samples, and it reflects the overall prediction correctness of a detection model. In our setting, the multimodal large language model (MLLM) produces detection results in the form of interpretable textual outputs (i.e., *Real* or *Fake*). We therefore map these textual predictions to binary labels before computing accuracy. For all baseline methods that output confidence scores or logits, we strictly follow their official implementations and apply the default decision thresholds provided by the authors

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn



(a) Content Distribution **(b) Generator Type Distribution**

Figure 1: Overview of the dataset composition. Left shows the content distribution across categories; right shows the distribution of generator types.

Table 1: Compared with single-turn forensic tuning, multi-turn forensic tuning yields a 5.38% absolute improvement in cross-domain accuracy, indicating stronger generalization capability.

Method	MTF-Bench	Midjourney	SD v1.4	SD v1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	*Avg.Acc*
Single-Turn Forensic										
Phi-3.5	78.25	65.30	63.50	60.65	57.50	51.90	64.40	52.35	79.75	63.73
Qwen2.5-VL-3B	84.16	86.30	85.50	85.40	77.55	68.55	84.65	78.65	83.70	81.61
Qwen2.5-VL-7B	83.35	86.25	90.65	88.80	82.95	76.00	89.35	85.25	83.90	85.17
LLaVA-v1.5-7B	89.71	81.10	90.85	90.20	90.45	84.30	89.55	91.15	90.85	88.68
LLaVA-v1.5-13B	89.76	81.50	90.20	89.50	90.35	87.80	89.15	90.15	90.40	88.76
Multi-Turn Forensic										
Phi-3.5	82.28	73.30	74.40	72.95	55.95	61.40	76.40	58.75	90.00	71.71
Qwen2.5-VL-3B	93.99	95.15	95.95	**96.30**	80.80	74.35	94.95	85.30	94.20	90.11
Qwen2.5-VL-7B	95.33	**98.35**	**96.70**	96.10	87.75	87.70	**95.35**	90.65	91.35	93.25
LLaVA-v1.5-7B	95.27	77.10	93.25	92.00	94.05	80.70	92.15	95.20	**97.30**	90.78
LLaVA-v1.5-13B	**96.41**	84.90	96.15	95.45	**96.00**	**91.30**	94.60	**96.40**	96.05	**94.14**

to obtain binary predictions. It is worth noting that, unlike conventional classifiers, the MLLM does not output continuous confidence scores or logit values. As a result, we do not report evaluation metrics that rely on ranked scores or logits, such as Average Precision (AP) or AUROC. Instead, ACC is adopted as a unified and directly comparable metric across all methods in our experiments.

5 Comparison with Single-Turn Forensic Baselines

Table 1 presents a quantitative comparison between the proposed multi-turn forensic tuning (MTF tuning) and representative single-turn forensic baselines on MTF-Bench. For a fair comparison, all models are trained under the same setting using the MTF-Bench training split. We report both *in-domain* performance evaluated on

the MTF-Bench test set and *out-of-domain* generalization results evaluated on the GenImage dataset.

As shown in the table, MTF tuning consistently outperforms single-turn forensic approaches across all evaluation settings, demonstrating its superior capability in AI-generated image detection. While single-turn forensic baselines can achieve competitive accuracy on the in-domain MTF-Bench evaluation, their performance gains are generally limited and rely heavily on task-specific heuristics learned from a single interaction.

In contrast, multi-turn forensic tuning enables the model to iteratively refine its reasoning process by aggregating complementary forensic cues from multiple perspectives across successive turns. This structured interaction mechanism leads to more robust decision-making and yields substantially improved generalization when transferred to the out-of-domain GenImage benchmark.

Overall, the results in Table 1 indicate that MTF tuning provides clear and consistent performance gains over single-turn forensic baselines in both in-domain and cross-domain evaluations, validating the effectiveness of multi-turn interaction and structured forensic reasoning for enhancing detection accuracy and generalization.

6 Qualitative Comparisons

Figures 2 and 3 present representative results of multi-turn forensic models, single-turn forensic models, and closed-source general-purpose large models on real images, while Figures 4 and 5 further illustrate typical examples of these three approaches on AI-generated images.

As can be observed, the multi-turn forensic model is able to progressively conduct multi-perspective analysis by focusing on potential synthesis artifacts within an image, performing fine-grained and interpretable forensic evaluations across different regions and feature levels. This process enables the construction of a more comprehensive and coherent chain of forensic evidence.

In contrast, closed-source large models lack task-specific priors and dedicated forensic reasoning mechanisms for synthetic image detection. As they are not explicitly trained for this task, their final predictions often exhibit insufficient confidence and require further verification to reach a definitive conclusion, as illustrated in Fig.2 and Fig.5. Moreover, their inference behavior tends to emphasize high-level semantic and content understanding, making them less reliable in consistently identifying generation artifacts.

The single-turn forensic model, constrained by its one-shot reasoning paradigm, adopts a relatively limited forensic perspective and fails to systematically capture diverse forgery cues. Consequently, its overall detection capability and accuracy are significantly inferior to those of the multi-turn forensic model.

References

[1] Huangsen Cao, Yongwei Wang, Yinfeng Liu, Sixian Zheng, Kangtao Lv, Zhimeng Zhang, Bo Zhang, Xin Ding, and Fei Wu. 2024. HyperDet: Generalizable Detection of Synthesized Images by Generating and Merging A Mixture of Hyper LoRAs. *arXiv preprint arXiv:2410.06044* (2024).

[2] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. 2023. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24480–24489.

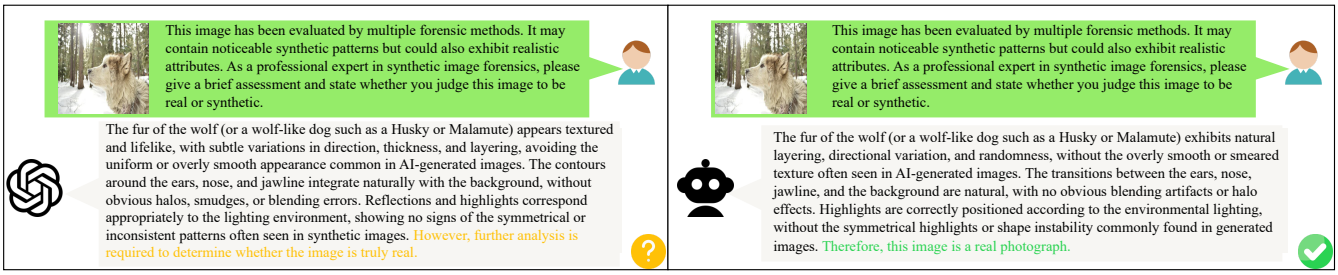
[3] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. 2024. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 28130–28139.

[4] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. 2020. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8695–8704.

[5] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. 2024. A sanity check for ai-generated image detection. *arXiv preprint arXiv:2406.19435* (2024).

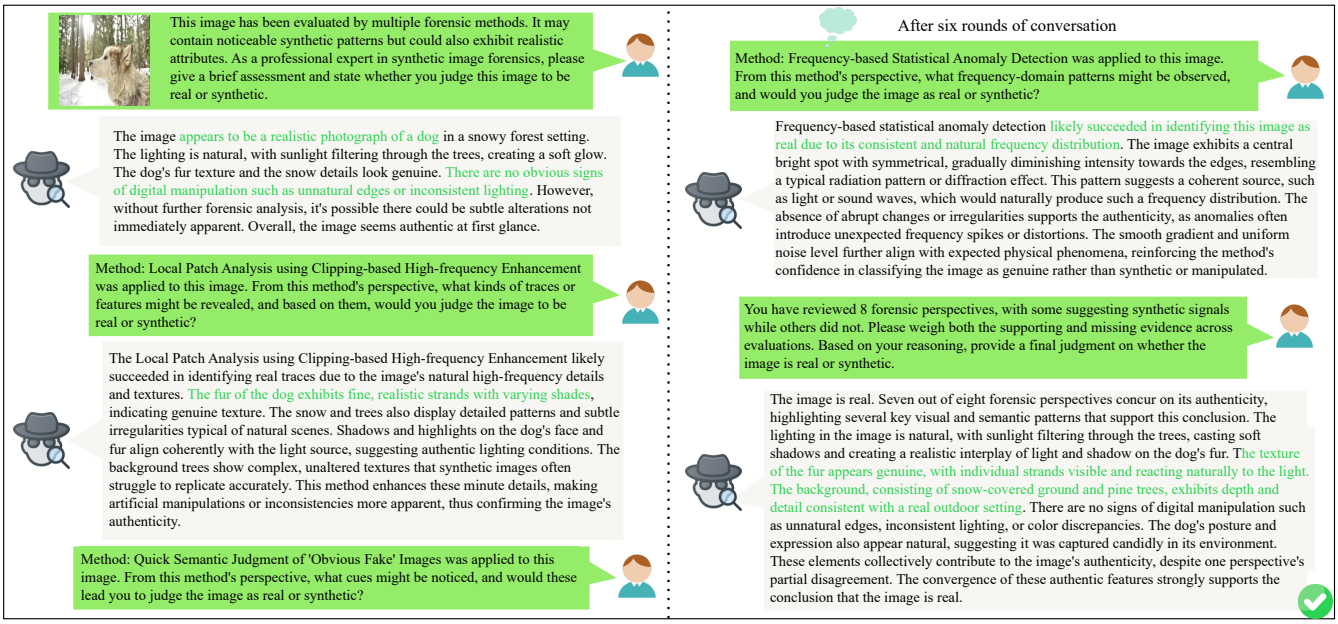
[6] Haifeng Zhang, Qinghui He, Xiuli Bi, Weisheng Li, Bo Liu, and Bin Xiao. 2025. Towards Universal AI-Generated Image Detection by Variational Information Bottleneck Network. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 23828–23837.

[7] Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. 2025. AIGI-Holmes: Towards Explainable and Generalizable AI-Generated Image Detection via Multimodal Large Language Models. *arXiv preprint arXiv:2507.02664* (2025).



a) Closed-source large model detection

b) Single-turn large model detection



c) Multi-turn forensic detection

Figure 2: Representative detection results on real images. Yellow text indicates that the model has insufficient confidence in the current prediction and requires further verification, while green text indicates that the model makes a correct judgment on real images.

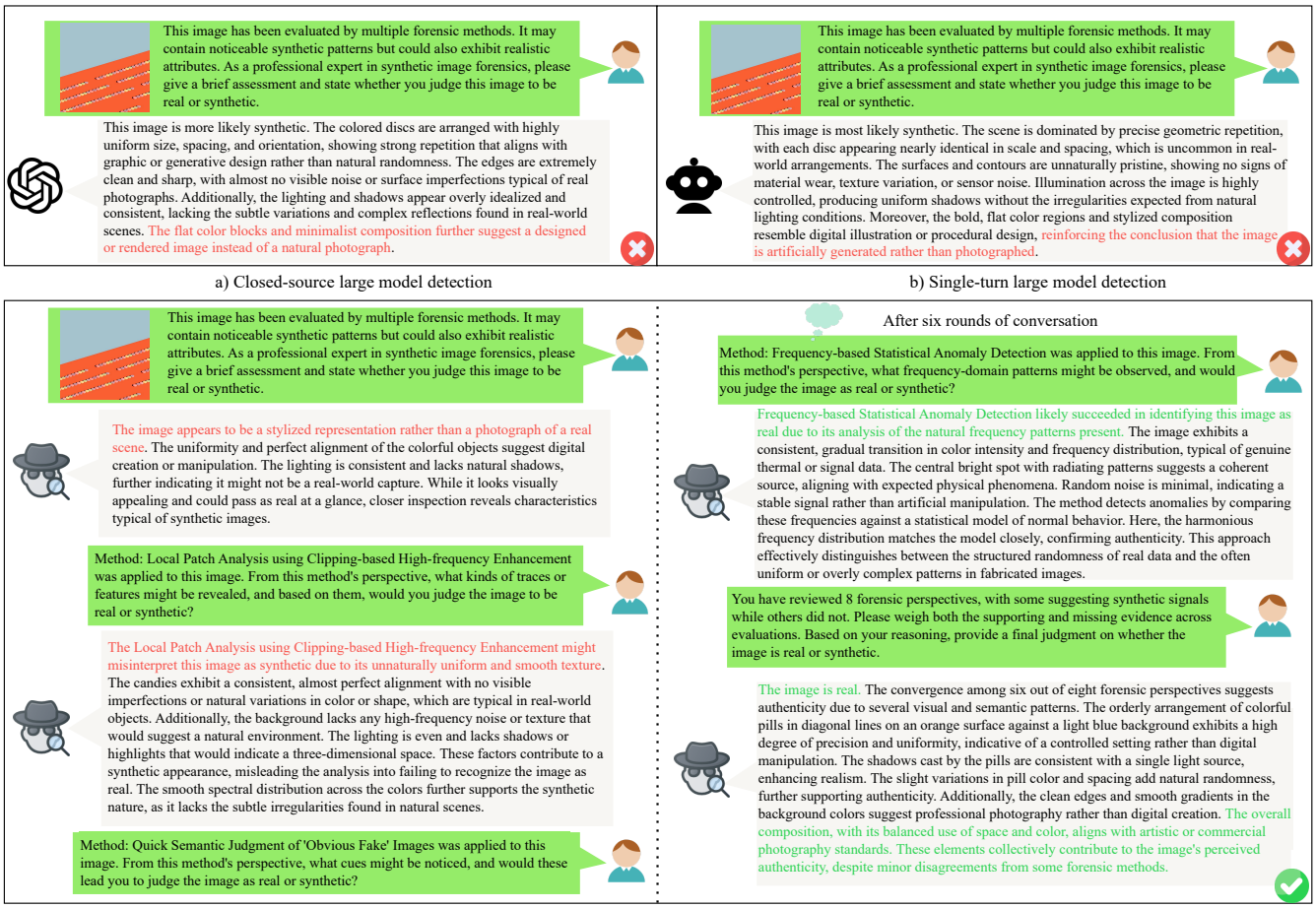
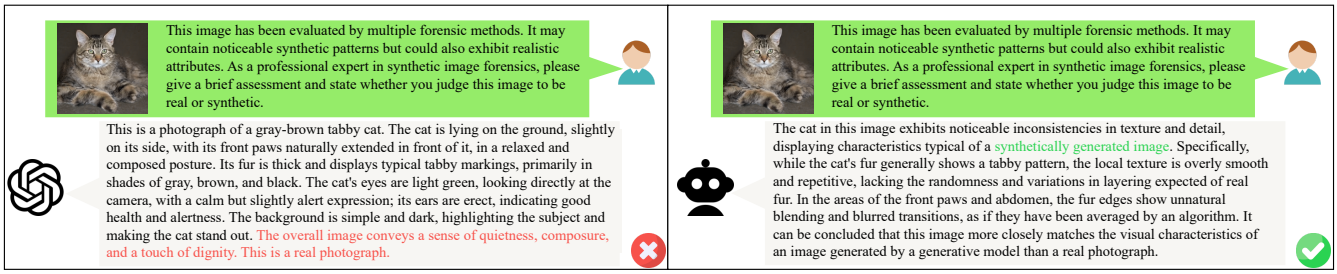
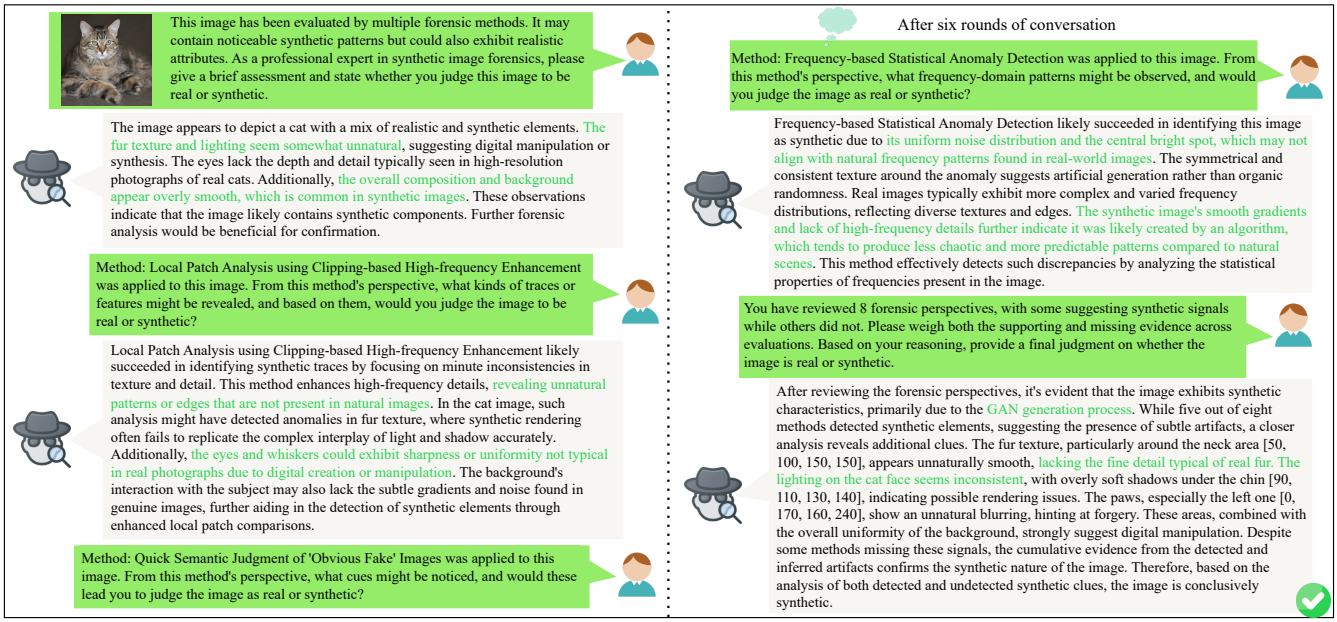


Figure 3: Representative detection results on real images. Red text indicates incorrect predictions in which real images are falsely classified as synthesized with high confidence, whereas green text indicates correct judgments on real images.



a) Closed-source large model detection

b) Single-turn large model detection



c) Multi-turn forensic detection

Figure 4: Representative detection results on fake images. Red text denotes incorrect predictions where synthesized images are misclassified as real, while green text denotes correct identification of synthesized images.

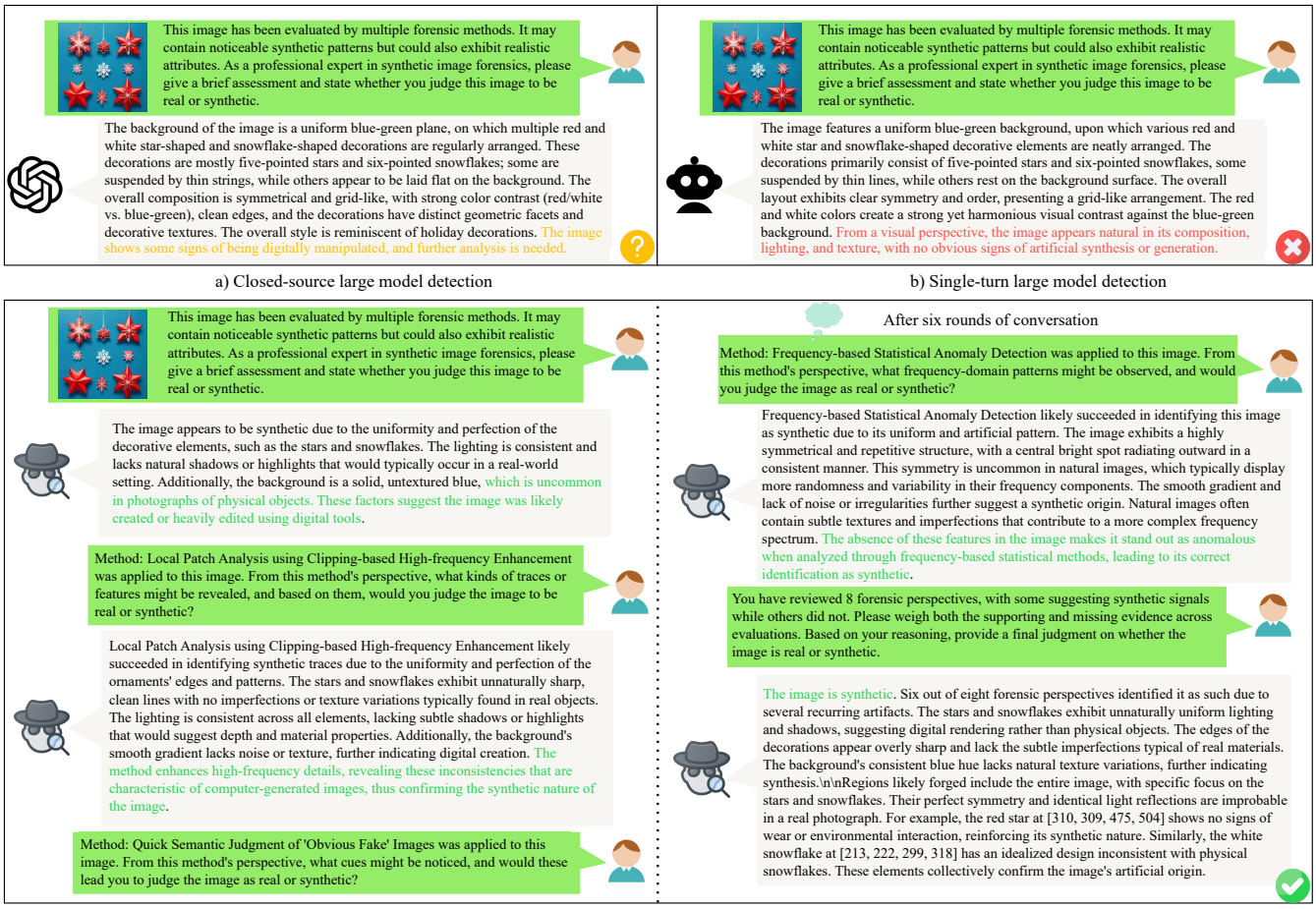


Figure 5: Representative detection results on fake images. Red text denotes incorrect predictions where synthesized images are misclassified as real, yellow text denotes insufficient confidence in the model's current conclusion and thus requires further verification, whereas green text denotes correct identification of synthesized images.