MedForge: Interpretable Medical Deepfake Detection via Forgery-aware Reasoning

Zhihui Chen[1], Kai He[1], Qingyuan Lei[2], Bin Pu[3], Jian Zhang[4], Yuling Xu[5], Mengling Feng[1]*

[1]National University of Singapore [2]The Chinese University of Hong Kong
[3]Hunan University [4]Xi'an Jiaotong University [5]Guangdong Provincial People's Hospital
zhihui.chen@u.nus.edu, {kai_he, ephfm}@nus.edu.sg
qingyuan.lei@link.cuhk.edu.hk, pubin@hnu.edu.cn
zhangjian062422@stu.xjtu.edu.cn, xuyuling@gdph.org.cn

Abstract

Text-guided image editors can now manipulate authentic medical scans with high fidelity, enabling lesion implantation/removal that threatens clinical trust and safety. Existing defenses are inadequate for healthcare. Medical detectors are largely black-box, while MLLM-based explainers are typically post-hoc, lack medical expertise, and may hallucinate evidence on ambiguous cases. We present MedForge, a data-and-method solution for pre-hoc, evidence-grounded medical forgery detection. We introduce MedForge-90K, a large-scale benchmark of realistic lesion edits across 19 pathologies with expert-aligned reasoning supervision via doctor inspection guidelines and gold edit locations. Building on it, MedForge-Reasoner performs localize-then-analyze reasoning, predicting suspicious regions before producing a verdict, and is further aligned with Forgery-aware GSPO to strengthen grounding and reduce hallucinations. Experiments demonstrate state-of-the-art detection accuracy and trustworthy, expert-aligned explanations. †

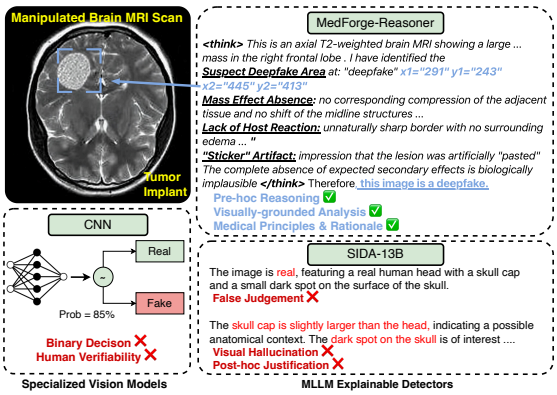

Figure 1: **Framework comparison. Left:** computer vision detectors output binary decision, offering no clinically verifiable evidence. **Right-bottom:** post-hoc MLLM explainers (e.g., SIDA (Huang et al., 2025b)) may produce plausible-sounding but ungrounded rationales, including hallucinated visual details. **Right-top: MedForge-Reasoner** performs *pre-hoc* localized reasoning by first identifying suspicious regions (blue) and then generating medically coherent, visually verifiable rationales grounded in image evidence.

1 Introduction

Recent advances in text-guided image editing have made it feasible to tamper with authentic medical scans with high fidelity. Editors such as Nano-Banana (Comanici et al., 2025; Chen and Feng, 2025) and GPT-Image (Hurst et al., 2024) can implant or remove subtle lesions while largely preserving anatomical structure and acquisition-style cues (Huang et al., 2025a; Alsaheel et al., 2023). Such manipulations are not merely hypothetical. They can distort clinical records for insurance fraud, malpractice disputes, or biased treatment/triage, and may even mislead trained experts (Amiri et al., 2024). This creates an urgent need for medical forgery detection that is reliable under clinically realistic edits.

However, existing defenses fall short of clinical requirements. Medical deepfake detectors (Li et al., 2025; Chen et al., 2025; Albahli and Nawaz, 2024) are often black-box classifiers that provide little interpretable evidence, limiting trust and accountability. General-domain "explainable" detectors (Huang et al., 2025b; Zhou et al., 2025) leverage Multimodal Large Language Models (MLLMs), but typically in a post-hoc manner and without medical expertise, despite the fact that clinically useful rationales must be *medically coherent* and *visually verifiable*. As shown in Figure 1, under unfamiliar or ambiguous cases, their explanations may regress to generic templates or hallucinated evidence, yielding plausible-sounding but non-verifiable rationales. In other words, post-hoc rationalization does not guarantee evidence-based

*Corresponding author

†Code, dataset and model checkpoint are released at https://github.com/richardChenzhihui/ACL2026-MedForge

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 25672–25692
July 2-7, 2026

reasoning, which is precisely the requirement for clinical adoption (Zhang et al., 2026a,b; Lin et al., 2025b; He et al., 2025).

We argue that medical forgery detection should be formulated as *pre-hoc* reasoning grounded in localized evidence. Concretely, a system should first identify suspicious manipulated regions and only then reason toward a verdict. This "localize-then-analyze" constraint makes explanations inspectable and suppresses template reuse and hallucination by anchoring reasoning to verifiable pixels (Zhang et al., 2025; Li et al., 2021). More broadly, we treat localization as a first-class constraint for explanation faithfulness, turning grounding from an afterthought into an explicit objective (He et al., 2022; Lin et al., 2025a).

To enable this paradigm, we introduce **MedForge-90K**, a large-scale benchmark of lesion implant/removal on authentic images across 19 pathologies, generated by 10 SOTA MMDiT/LDM-based editing models (Huang et al., 2025a). Crucially, MedForge-90K provides expert-aligned supervision for grounded explanations: we combine doctor-defined inspection guidelines with gold manipulation locations, and use them to produce medically aligned rationales that are explicitly tied to the edited regions. Building on this resource, we propose **MedForge-Reasoner**, an MLLM-based detector trained with an explicit localization-then-analysis objective to reason before deciding. We further align grounding and explanation quality via a two-stage strategy (SFT cold-start + Forgery-aware GSPO) that directly rewards correct localization and evidence-grounded reasoning. Experiments show that enforcing such grounding improves explanation quality and reduces hallucinations, measured with an MLLM-as-judge protocol. The main contributions are as follows:

- We introduce **MedForge-90K**, the first large-scale medical forgery benchmark of high-quality lesion manipulations with granular explainable annotations, addressing data scarcity in medical deepfake detection.
- We propose **MedForge-Reasoner**, a novel MLLM-based detector that integrates detection with grounded CoT reasoning, and employs a forgery-aware GSPO to anchor reasoning to visual forgery evidence.
- Extensive experiments show that Forgery-aware GSPO aligns the detector with factual visual evidence in forgery reasoning, improving detection accuracy by 7.65% while significantly reducing hallucinations by 16.2% compared to strong baselines.

2 Related Work

Medical Deepfake Benchmarks. Most prior work on medical image generation targets data augmentation and class balancing rather than simulating adversarial forgery scenarios. Early studies (Guo et al., 2025; Motamed et al., 2021) used VAEs/GANs to synthesize CT/MRI scans, which do not reflect the modern threat of editing authentic patient records. While MedForensics (Li et al., 2025) takes a step toward forgery detection, existing benchmarks remain limited in two aspects. (i) *Threat mismatch:* real-world medical deepfakes often involve targeted tampering of authentic scans (e.g., lesion implant/removal) to enable insurance fraud or misdiagnosis (Stroebel et al., 2023; Hsu et al., 2025), rather than generating scans from scratch. (ii) *Supervision gap:* they typically provide only labels and lack localized edit evidence and expert-aligned reasoning signals required for clinically verifiable explanations. **MedForge-90K** addresses these gaps by benchmarking high-fidelity lesion edits on authentic images using modern text-guided editors and by providing guideline- and location-grounded reasoning supervision.

Interpretable Deepfake Detection. Standard medical forgery detectors are predominantly black-box binary classifiers (Li et al., 2025; Tan et al., 2024), offering limited evidence to support clinical trust. Recent general-domain approaches (Huang et al., 2025b; Zhou et al., 2025; Xu et al., 2025) incorporate MLLMs to generate textual explanations, yet they are often post-hoc: a separate module makes the decision and the MLLM rationalizes it afterwards, which can decouple explanations from the actual evidence. Moreover, MLLMs are prone to visual hallucination (Huang et al., 2024), especially on unfamiliar or ambiguous cases, where they may repeat generic templates or describe non-existent artifacts. Although pre-hoc reasoning has been explored in AIGC detection (Tan et al., 2025; Gao et al., 2025), these methods are not designed for subtle medical lesion forgeries and typically lack (i) medical-domain constraints and (ii) explicit localization-grounding objectives to enforce pixel-verifiable rationales. In contrast, our approach unifies detection and reasoning in a pre-hoc manner

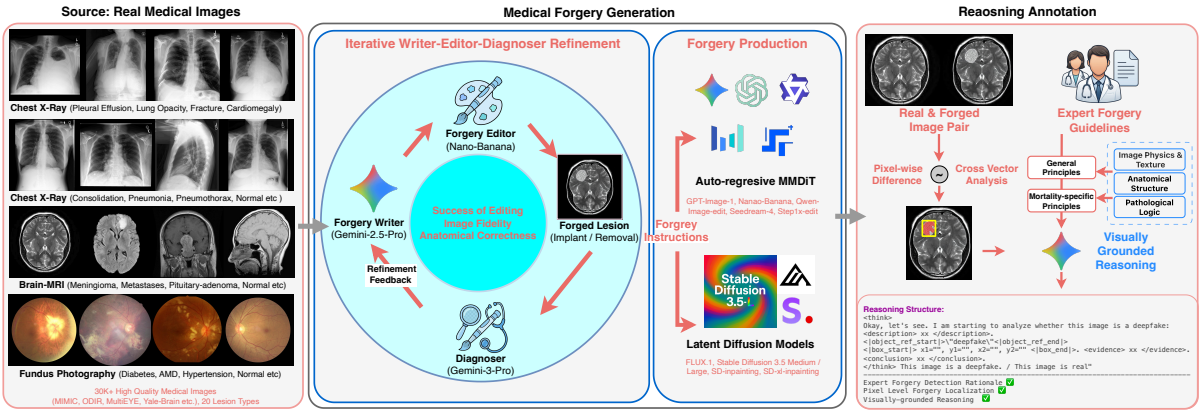

Figure 2: **Overview of the MedForge-90K construction pipeline.** The framework proceeds in three stages: medical image collection across three modalities, forgery generation via a *Writer-Editor-Diagnoser* loop, and expert-aligned annotation utilizing expert guidelines to generate hierarchical diagnostic reasoning.

and explicitly enforces localization-grounded reasoning through Forgery-aware GSPO. Crucially, GSPO makes localization-grounding an optimization objective, coupling the verdict with inspectable regions and curbing hallucinated rationales.

3 MedForge-90K Dataset

We introduce MedForge-90K (Figure 2), the first large-scale medical forgery benchmark with detailed forgery and reasoning annotations. For the source images, we evenly select 30K high-quality medical images across Chest X-Ray, Brain MRI, and Fundus Photography from 5 public datasets, MIMIC (Johnson et al., 2016), ODIR (Li et al., 2020), MultiEYE (Wang et al., 2024), Yale-Brain (Chadha et al., 2025) and Brain-MRI (Nickparvar, 2021). These medical images are classified into 19 types of pathologies and 1 normal status according to their original labels. Forgery manipulations including lesion implant and removal take place within each modality. In summary, MedForge includes: (i) **Real Images** (30K) spanning major 2D modalities with 19 lesion types plus healthy scans; (ii) **Lesion Implant** (30K) healthy scans with implanted lesions, evenly distributed across 10 forgery models; and (iii) **Lesion Removal** (30K) diseased scans with removed lesions, evenly distributed across 10 forgery models.

3.1 Forgery Pipeline

We employ 10 state-of-the-art text-guided medical image editing models based on MMDiT/LDM paradigms, including Nano-Banana (Comanici et al., 2025), GPT-Image (Hurst et al., 2024), Qwen-Image-Edit (Wu et al., 2025), Seedream 4.0 (Seedream et al., 2025), Stable Diffusion 3.5 (Esser et al., 2024), and Stable Diffusion Inpainting (Rombach et al., 2022). Text prompts are a critical component of editing, as they specify the medical context and transformation intent. To obtain realistic and anatomically plausible manipulations, we introduce a *writer–editor–diagnoser* refinement loop. Specifically, a *writer* drafts an initial prompt, the *editor* generates an edited image, and a *diagnoser* evaluates whether the result achieves the desired condition while remaining anatomically consistent. If the edit is unsatisfactory, the diagnoser provides targeted feedback and the writer revises the prompt; the loop iterates until success or a maximum number of rounds, after which the sample is discarded. In practice, the writer and diagnoser are implemented with Gemini 2.5/3 Pro, while Nano-Banana serves as the editor during prompt refinement. The refined prompts are applied to all editing models to construct forgeries. For diffusion-based editors requiring inpainting masks, we use Nano-Banana's localized forgery regions as mask inputs.

3.2 Expert-aligned Reasoning Annotation

We aim to annotate forged images with accurate and professional rationales. To achieve this, we engaged medical experts to formulate a comprehensive detection guideline. As shown in the "Expert Forgery Guidelines" in Figure 2, this guideline is structured into two pillars: General Principles (universal biomedical principles) and Modality-Specific Principles (specific constraints for MRI, Fundus, and CXR). During annotation, these guidelines are incorporated into the MLLM prompt, enabling **hierarchical expert-aligned reasoning**

over medical forgeries at three levels:
1. **Image Physics & Texture**: Following the *General Principles*, the model detects low-level anomalies such as inconsistent noise distribution, inpainting traces, and unnatural boundaries.
2. **Anatomical Structure**: Based on the *Modality-Specific Criteria*, the model verifies morphological correctness, such as vascular continuity in fundus photography or gyral symmetry in brain MRI.
3. **Pathological Logic**: Integrating the core philosophy of "Biological Interconnectivity" from the guidelines, the model validates high-level plausibility, rejecting lesions that lack necessary secondary effects (e.g., mass effect, edema) or violate chronological disease evolution.

The above expert-aligned protocol steers the generated rationales toward clinically meaningful diagnostic reasoning. Following recent practice (Zhou et al., 2025; Huang et al., 2025b), we use an MLLM (Gemini 2.5 Pro) to automate annotation. To reduce visual hallucinations and enforce the medical principles described above, we adopt a *forgery-grounded* annotation strategy by applying Change Vector Analysis (CVA) (Malila, 1980) to compute a per-pixel change magnitude, $|\mathbf{I}_{\text{forged}} - \mathbf{I}_{\text{real}}|$. We then threshold high-response regions to obtain a manipulation mask, which is finally converted into bounding box (bbox) coordinates as Eq. 1.

$$M_{\text{bbox}} : \texttt{<bbox } x_1, y_1, x_2, y_2 \texttt{ />} \tag{1}$$

These modified regions serve as the key visual components of forgery signs. To generate high-quality annotations, we integrate these CVA-derived coordinates with the hierarchical expert guidelines to construct a **visually-grounded reasoning prompt**.

This unified prompting strategy explicitly directs the MLLM to anchor its analysis on the provided bounding boxes (or the absence). Guided by the three-tiered criteria (Physics, Anatomy, Pathology), the model scrutinizes the designated regions to expose specific artifacts in forged samples, or validates the preservation of biological logic in real samples. As illustrated in the "Reasoning Structure" of Figure 2, the output is enforced into a structured chain-of-thought format consisting of description, evidence, and conclusion. This ensures that the reasoning is derived from professional medical rationale and grounded with visual evidence.

4 Methodology

In this section, we present the MedForge-Reasoner framework. We first formulate the task of interpretable medical forgery detection. Then, we detail our two-stage training pipeline: the reasoning cold-start via Supervised Fine-tuning (SFT) and the Forgery-aware Group Sequence Policy Optimization (GSPO), designed to align the model with factual visual evidence.

4.1 Task Formulation

Existing MLLMs often suffer from *visual hallucination* (Huang et al., 2024), where the model fabricates details which are not present in the image. In forgery detection, this leads to ungrounded reasoning. To address this, we define the detection task as a unified sequence generation problem that enforces *grounding before reasoning*.

Specifically, given a medical image $\boldsymbol{x}$, the model is trained to generate a sequence S structured as:

$$S = [\hat{M}_{\text{bbox}}, \texttt{<reasoning>}, \hat{y}], \tag{2}$$

where $\hat{M}_{\text{bbox}}$ represents the coordinates of the manipulated region (or a special token for authentic images), followed by the textual reasoning chain, and finally the detection decision $\hat{y}$. By enforcing the prediction of forgery location at the very beginning, we force the model to attend to visual anomalies before hallucinating textual descriptions.

4.2 Stage 1: Reasoning Cold Start

To equip the MLLM with fundamental medical knowledge and the proposed reasoning format, we perform SFT training. As illustrated in Figure 3, the SFT data is derived from the MedForge-90K, incorporating expert-aligned rationales and ground-truth bounding boxes.

We employ LoRA to efficiently fine-tune the model parameters θ on the dataset $\mathcal{D} = \{(\boldsymbol{x}, \boldsymbol{y})\}$. Cross-entropy loss serves as the optimization objective:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}} \sum_{t=1}^{T} \log P_\theta(y_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t}), \tag{3}$$

where x is the input image and user query, y denotes the target output sequence including reasoning and final answer, with t as index of the generated token. This stage allows the model to internalize the format requirements and forgery patterns.

4.3 Stage 2: Forgery-aware GSPO

Although SFT establishes basic capabilities, standard cross-entropy loss is insufficient to penalize subtle hallucinations or enforce strict alignment

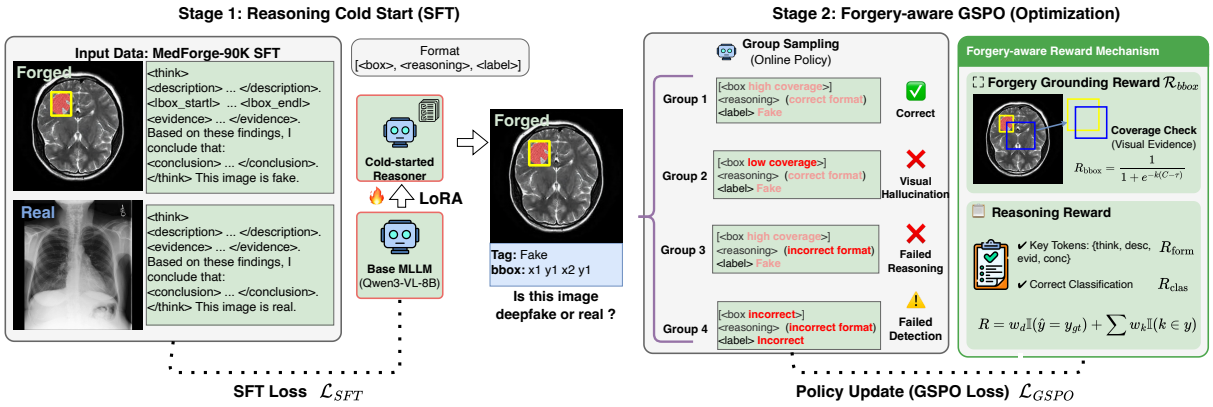

Figure 3: **MedForge-Reasoner Two-stage Training.** SFT for cold-starting the reasoning format, followed by Forgery-aware GSPO. The GSPO stage introduces a reward function balancing visual grounding coverage and reasoning structure compliance to ensure the model localize the correct forgery region before reasoning.

with visual evidence. To further align the detector, we introduce Forgery-aware Group Sequence Policy Optimization (GSPO).

GSPO applies importance sampling at the sequence level, which provides stable updates for reasoning tasks. Given a forgery input $\boldsymbol{x}$, we sample a group of G outputs $\{y_1, y_2, \ldots, y_G\}$ from the current policy π_θ. The objective function maximizes the expected reward of these generations:

$$\mathcal{L}_{\text{GSPO}}(\theta) = -\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|\boldsymbol{x})}\left[\frac{1}{G}\sum_{i=1}^{G}\min\left(s_i(\theta)\hat{A}_i,\ \text{clip}\big(s_i(\theta),\,1-\epsilon,\,1+\epsilon\big)\hat{A}_i\right)\right], \tag{4}$$

where $s_i(\theta)$ is the importance ratio between new and old policies, and $\hat{A}_i$ is the advantage:

$$\hat{A}_i = \frac{R(\boldsymbol{x},y_i) - \text{mean}(\{R(\boldsymbol{x},y_j)\}_{j=1}^{G})}{\text{std}(\{R(\boldsymbol{x},y_j)\}_{j=1}^{G})}. \tag{5}$$

Crucially, to enforce sequence-level stability, we define the importance ratio $s_i(\theta)$ based on the geometric mean of the likelihood ratio over the sequence length $|y_i|$:

$$s_i(\theta) = \exp\left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\log\frac{\pi_\theta(y_{i,t}|\boldsymbol{x},y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|\boldsymbol{x},y_{i,<t})}\right). \tag{6}$$

Then, our reward function $R(\boldsymbol{x}, y_i)$ is composed of two parts to penalize visual hallucination and incorrect analysis in forgery detection as follows.

1. Forgery Grounding Reward (R_{bbox}). Unlike standard object detection tasks that demand precise boundary regression, our goal is to ensure the MLLM's reasoning is grounded in the correct anomaly region. Therefore, instead of strict Intersection over Union (IoU), we adopt a *Mask Coverage* $\mathcal{C}$ to measure the rate of ground truth forgery area captured by the model's prediction:

$$\mathcal{C} = \frac{|M_{\text{bbox}} \cap \hat{M}_{\text{bbox}}|}{|M_{\text{bbox}}|}, \tag{7}$$

To enhance training stability, we map the coverage metric $\mathcal{C}$ into a reward signal using a shaped sigmoid function. This design serves two purposes: (a) it suppresses noise from low-overlap predictions and (b) saturates for high-quality overlaps, thereby prioritizing the robust localization of forgeries over pixel-perfect alignment. The bounding box reward is formulated as:

$$R_{\text{bbox}} = \frac{1}{1+e^{-k(\mathcal{C}-\tau)}}, \tag{8}$$

where k and τ are hyperparameters controlling the reward sensitivity and threshold.

2. Reasoning Rewards. To ensure the model follows a logical reasoning path and arrives at an accurate conclusion, we decompose the task-related reward into two components: the formatting reward (R_{form}) and the classification reward (R_{clas}).

R_{form} incentivizes the model to adhere to the mandated Chain-of-Thought (CoT) structure (Description → Analysis → Conclusion):

$$R_{\text{form}} = \sum_{k\in\mathcal{K}} w_k\mathbb{I}(k \in y), \tag{9}$$

where y denotes the generated text sequence, and $\mathcal{K} = \{\text{“description”}, \text{“analysis”}, \text{“conclusion”}\}$ represents the set of mandatory structural keywords. The indicator function $\mathbb{I}(\cdot)$ assigns a weight w_k for each keyword present in the sequence, penalizing structural deviations.

R_{clas} evaluates the correctness of detection:

$$R_{\text{clas}} = w_d\mathbb{I}(\hat{y} = y_{gt}) \tag{10}$$

where $\hat{y}$ is the predicted label parsed from the generated sequence, y_{gt} is the ground truth, and w_d is the weighting factor for prediction accuracy.

The total reward is then formulated as $R = R_{\text{bbox}} + R_{\text{form}} + R_{\text{clas}}$. This multi-faceted reward strategy explicitly incentivizes the model to "look" at the correct region before "reasoning" and "concluding", thereby minimizing visual hallucinations and improving reasoning quality.

5 Experiments

In this section, we conduct comprehensive empirical evaluations to validate MedForge-Reasoner's detection performance, generalizability and reasoning quality. Then, we perform ablation studies to verify the efficacy of our proposed contributions.

5.1 Experimental Setup

SOTA Baselines. We benchmark our method against SOTA interpretable deepfake detectors, AIGI-Holmes (Zhou et al., 2025), SIDA (Huang et al., 2025b), FakeVLM (Wen et al., 2025). We also assess four SOTA generic MLLMs including Qwen3-VL-Flash (30B), Qwen3-VL-Plus (235B) (Bai et al., 2025) and Gemini 3 Flash, Gemini 3 Pro (Google DeepMind, 2025).

Evaluation Metrics. Forgery Detection performance is evaluated via **Accuracy** and **F1** score, and presented in "Real", "Forgery Implant" and "Forgery Removal". To assess reasoning quality and visual hallucinations, we introduce *MLLM-as-Judge* metric using SOTA MLLMs. The judge scores generated reasoning output on a scale of 0-100% based on three criteria: (1) *Logical Correctness*: Whether the judgement is derived from the visual evidence. (2) *Visual Hallucination*: Whether the analysis matches the ground truth anomalies (e.g., matching the bbox) or fabricates. (3) *Medical Professionalism*: Whether the terminology aligns with the expert guidelines. Detailed metric definitions are shown in Appendix D.3.

Implementation Details. We utilize the constructed MedForge-90K dataset for experiments. We randomly split the data into SFT, GSPO training, and testing sets with a ratio of 5:1:3. Specifically, 50K samples are used for SFT cold-start, 10K for GSPO training, and 30K for testing. To ensure balanced evaluation, each split maintains a 1:1:1 ratio of Real, Lesion Implant, and Lesion Removal images. Training details of model and baseline are elaborated in Appendix D.

5.2 Main Results

We report the main detection performance in Table 1 with additional results in Appendix B . To assess performance under both in-domain and out-of-distribution (OOD) conditions, we consider three evaluation settings in Table 1 as follows:

(a) In-Domain: The detector is trained and tested on the full dataset, covering all forgery types and generator models.

(b) Cross-Model: To test robustness to unseen generators, we exclude four advanced models from training, Nano-Banana, GPT-Image, Stable Diffusion 3.5 Medium, and Stable Diffusion XL-Inpainting, while evaluating on the default test set.

(c) Cross-Forgery: To evaluate generalization to unseen manipulations, the training set excludes lesion implant samples (for both OOD cases, the test data follows the default setting).

Targeting untrainable SOTA commercial MLLM baselines, we simulate the above settings by In Context Learning (ICL) (Dong et al., 2024). We design ICL prompts to provide MLLMs with different levels of forgery detection knowledge. Three levels of ICL prompts are customized to match the *In-Domain*, *Cross-Forgery* and *Cross-Model* setting. The *In-Domain* ICL provides detection clues for all manipulation types. The *Cross-Forgery* ICL covers only Lesion Removal Forgeries, and the *Cross-Model* ICL excludes unseen forgery models. See details in Appendix D.1.

As illustrated in Table 1, MedForge-Reasoner achieves SOTA performance across all settings. For the In-Domain setting, our method achieves near-perfect detection, outperforming the strongest specialized detector (SIDA-13B) by over 7.65% in average accuracy. Notably, MedForge-Reasoner demonstrates significant robustness in OOD scenarios. While specialized detectors and generic MLLMs suffer noticeable performance degradation when facing unseen forgeries or models, our method maintains a substantial lead, surpassing the strongest baselines by 8.2% in Cross-Forgery and 10.0% in Cross-Model settings. This suggests that by explicitly training the model to ground its reasoning in visual anomalies (via GSPO), MedForge-Reasoner learns generic traces of tampering (e.g., edge inconsistencies, noise artifacts) rather than overfitting to specific lesion patterns or generator fingerprints. This generalizability suggests that MedForge-Reasoner is applicable in real-world forgery defence.

Methods	Real			Forgery: Implant			Forgery: Remove			Average		
	In-Domain	Cross-Forgery	Cross-Model	In-Domain	Cross-Forgery	Cross-Model	In-Domain	Cross-Forgery	Cross-Model	In-Domain	Cross-Forgery	Cross-Model
Accuracy												
Specialized Detectors												
SIDA-7B	77.56$_{(\uparrow 10.27)}$	73.14$_{(\uparrow 8.03)}$	73.42$_{(\uparrow 8.93)}$	79.83$_{(\uparrow 4.49)}$	76.18$_{(\uparrow 0.94)}$	76.57$_{(\uparrow 1.77)}$	82.31$_{(\uparrow 17.26)}$	77.92$_{(\uparrow 11.43)}$	75.68$_{(\uparrow 9.98)}$	79.90$_{(\uparrow 10.67)}$	75.75$_{(\uparrow 6.80)}$	75.22$_{(\uparrow 6.89)}$
SIDA-13B	89.37$_{(\uparrow 22.08)}$	85.68$_{(\uparrow 20.57)}$	84.25$_{(\uparrow 19.76)}$	91.52$_{(\uparrow 16.18)}$	87.24$_{(\uparrow 12.00)}$	85.86$_{(\uparrow 11.06)}$	93.84$_{(\uparrow 28.79)}$	90.17$_{(\uparrow 23.68)}$	86.53$_{(\uparrow 20.83)}$	91.58$_{(\uparrow 22.35)}$	87.70$_{(\uparrow 18.75)}$	85.55$_{(\uparrow 17.22)}$
FakeVLM	86.48$_{(\uparrow 19.19)}$	81.86$_{(\uparrow 16.75)}$	76.94$_{(\uparrow 12.45)}$	88.93$_{(\uparrow 13.59)}$	83.57$_{(\uparrow 8.33)}$	79.82$_{(\uparrow 5.02)}$	91.27$_{(\uparrow 26.22)}$	86.35$_{(\uparrow 19.86)}$	82.51$_{(\uparrow 16.81)}$	88.89$_{(\uparrow 19.66)}$	83.93$_{(\uparrow 14.98)}$	79.76$_{(\uparrow 11.43)}$
AIGI-Holmes	90.24$_{(\uparrow 22.95)}$	86.53$_{(\uparrow 21.42)}$	84.37$_{(\uparrow 19.88)}$	88.12$_{(\uparrow 12.78)}$	84.46$_{(\uparrow 9.22)}$	80.91$_{(\uparrow 6.11)}$	90.58$_{(\uparrow 25.53)}$	90.64$_{(\uparrow 24.15)}$	87.23$_{(\uparrow 21.53)}$	89.65$_{(\uparrow 20.42)}$	87.21$_{(\uparrow 18.26)}$	84.17$_{(\uparrow 15.84)}$
Generic MLLMs												
Qwen3-VL-Flash	57.60$_{(\downarrow 9.69)}$	55.50$_{(\downarrow 9.61)}$	54.90$_{(\downarrow 9.59)}$	47.83$_{(\downarrow 27.51)}$	50.47$_{(\downarrow 24.77)}$	50.93$_{(\downarrow 23.87)}$	54.17$_{(\downarrow 10.88)}$	54.17$_{(\downarrow 12.32)}$	54.61$_{(\downarrow 11.09)}$	53.20$_{(\downarrow 16.03)}$	53.38$_{(\downarrow 15.57)}$	53.48$_{(\downarrow 14.85)}$
Qwen3-VL-Plus	54.14$_{(\downarrow 13.15)}$	54.28$_{(\downarrow 10.83)}$	55.10$_{(\downarrow 9.39)}$	57.42$_{(\downarrow 17.92)}$	55.78$_{(\downarrow 19.46)}$	56.03$_{(\downarrow 18.77)}$	55.80$_{(\downarrow 9.25)}$	55.86$_{(\downarrow 10.63)}$	56.10$_{(\downarrow 9.60)}$	55.79$_{(\downarrow 13.44)}$	55.31$_{(\downarrow 13.64)}$	55.74$_{(\downarrow 12.59)}$
Gemini 3 Flash	71.26$_{(\uparrow 3.97)}$	73.39$_{(\uparrow 8.28)}$	72.57$_{(\uparrow 8.08)}$	57.33$_{(\downarrow 18.01)}$	62.46$_{(\downarrow 12.78)}$	60.14$_{(\downarrow 14.66)}$	57.14$_{(\downarrow 7.91)}$	62.27$_{(\downarrow 4.22)}$	60.34$_{(\downarrow 5.36)}$	61.91$_{(\downarrow 7.32)}$	66.04$_{(\downarrow 2.91)}$	64.35$_{(\downarrow 3.98)}$
Gemini 3 Pro	67.29	65.11	64.49	75.34	75.24	74.80	65.05	66.49	65.70	69.23	68.95	68.33
MedForge-Reasoner	**99.24**$_{(\uparrow 31.95)}$	**95.24**$_{(\uparrow 30.13)}$	**92.86**$_{(\uparrow 28.37)}$	**99.24**$_{(\uparrow 23.90)}$	**93.39**$_{(\uparrow 18.15)}$	**94.86**$_{(\uparrow 20.06)}$	**99.21**$_{(\uparrow 34.16)}$	**99.15**$_{(\uparrow 32.66)}$	**94.09**$_{(\uparrow 28.39)}$	**99.23**$_{(\uparrow 30.00)}$	**95.93**$_{(\uparrow 26.98)}$	**93.94**$_{(\uparrow 25.61)}$
F1 Score												
Specialized Detectors												
SIDA-7B	76.84$_{(\uparrow 15.02)}$	72.47$_{(\uparrow 9.83)}$	67.73$_{(\uparrow 5.51)}$	86.47$_{(\uparrow 12.25)}$	74.06$_{(\uparrow 2.53)}$	69.85$_{(\downarrow 1.17)}$	81.64$_{(\uparrow 22.17)}$	77.28$_{(\uparrow 19.66)}$	72.94$_{(\uparrow 16.75)}$	81.65$_{(\uparrow 16.48)}$	74.60$_{(\uparrow 10.67)}$	70.17$_{(\uparrow 7.03)}$
SIDA-13B	88.69$_{(\uparrow 26.87)}$	84.95$_{(\uparrow 22.31)}$	80.58$_{(\uparrow 18.36)}$	90.86$_{(\uparrow 16.64)}$	86.57$_{(\uparrow 15.04)}$	83.17$_{(\uparrow 12.15)}$	93.18$_{(\uparrow 33.71)}$	89.46$_{(\uparrow 31.84)}$	85.87$_{(\uparrow 29.68)}$	90.91$_{(\uparrow 25.74)}$	86.99$_{(\uparrow 23.06)}$	83.21$_{(\uparrow 20.07)}$
FakeVLM	85.73$_{(\uparrow 23.91)}$	81.18$_{(\uparrow 18.54)}$	76.29$_{(\uparrow 14.07)}$	88.27$_{(\uparrow 14.05)}$	82.86$_{(\uparrow 11.33)}$	79.16$_{(\uparrow 8.14)}$	90.58$_{(\uparrow 31.11)}$	85.67$_{(\uparrow 28.05)}$	81.84$_{(\uparrow 25.65)}$	88.19$_{(\uparrow 23.02)}$	83.24$_{(\uparrow 19.31)}$	79.10$_{(\uparrow 15.96)}$
AIGI-Holmes	88.57$_{(\uparrow 26.75)}$	85.84$_{(\uparrow 23.20)}$	86.68$_{(\uparrow 24.46)}$	90.46$_{(\uparrow 16.24)}$	88.73$_{(\uparrow 17.20)}$	84.25$_{(\uparrow 13.23)}$	89.82$_{(\uparrow 30.35)}$	85.97$_{(\uparrow 28.35)}$	83.58$_{(\uparrow 27.39)}$	89.62$_{(\uparrow 24.45)}$	86.85$_{(\uparrow 22.92)}$	84.84$_{(\uparrow 21.70)}$
Generic MLLMs												
Qwen3-VL-Flash	32.48$_{(\downarrow 29.34)}$	38.62$_{(\downarrow 24.02)}$	40.26$_{(\downarrow 21.96)}$	55.08$_{(\downarrow 19.14)}$	53.70$_{(\downarrow 17.83)}$	52.69$_{(\downarrow 18.33)}$	63.13$_{(\uparrow 3.66)}$	59.51$_{(\uparrow 1.89)}$	58.78$_{(\uparrow 2.59)}$	50.23$_{(\downarrow 14.94)}$	50.61$_{(\downarrow 13.32)}$	50.58$_{(\downarrow 12.56)}$
Qwen3-VL-Plus	47.73$_{(\downarrow 14.09)}$	46.36$_{(\downarrow 16.28)}$	46.18$_{(\downarrow 16.04)}$	54.70$_{(\downarrow 19.52)}$	53.91$_{(\downarrow 17.62)}$	54.99$_{(\downarrow 16.03)}$	52.42$_{(\downarrow 7.05)}$	54.26$_{(\downarrow 3.36)}$	55.32$_{(\downarrow 0.87)}$	51.62$_{(\downarrow 13.55)}$	51.51$_{(\downarrow 12.42)}$	52.16$_{(\downarrow 10.98)}$
Gemini 3 Flash	24.42$_{(\downarrow 37.40)}$	41.14$_{(\downarrow 21.50)}$	34.78$_{(\downarrow 27.44)}$	70.22$_{(\downarrow 4.00)}$	72.05$_{(\uparrow 0.52)}$	71.11$_{(\uparrow 0.09)}$	70.07$_{(\uparrow 10.60)}$	71.91$_{(\uparrow 14.29)}$	71.34$_{(\uparrow 15.15)}$	54.90$_{(\downarrow 10.27)}$	61.70$_{(\downarrow 2.23)}$	59.08$_{(\downarrow 4.06)}$
Gemini 3 Pro	61.82	62.64	62.22	74.22	71.53	71.02	59.47	57.62	56.19	65.17	63.93	63.14
MedForge-Reasoner	**98.86**$_{(\uparrow 37.04)}$	**93.29**$_{(\uparrow 30.65)}$	**92.07**$_{(\uparrow 29.85)}$	**98.86**$_{(\uparrow 24.64)}$	**92.97**$_{(\uparrow 21.44)}$	**94.63**$_{(\uparrow 23.61)}$	**99.21**$_{(\uparrow 39.74)}$	**99.14**$_{(\uparrow 41.52)}$	**93.76**$_{(\uparrow 37.57)}$	**98.98**$_{(\uparrow 33.81)}$	**95.13**$_{(\uparrow 31.20)}$	**93.49**$_{(\uparrow 30.35)}$

Table 1: **Main Experiment** - Forgery detection on MedForge-90K dataset. Methods are benchmarked against the Gemini 3 Pro, with arrows indicating performance differences (↑/↓) relative to it. **Bold** indicates the best result, and underline denotes the second-best.

MLLM-as-Judge	**Gemini 3 Pro (%)**				**Qwen3-VL-Plus (%)**			
	LC	VH	MP	**Avg.**	LC	VH	MP	**Avg.**
SIDA-7B	34.6	27.9	54.3	38.9	71.3	53.4	78.5	67.7
SIDA-13B	40.2	32.6	61.6	44.8	75.6	57.7	82.1	71.8
FakeVLM	48.0	53.0	45.0	48.7	50.0	55.0	48.0	51.0
AIGI-Holmes	55.0	60.0	58.0	57.7	58.0	62.0	57.0	59.0
Qwen3-VL-Flash	68.2	57.3	85.9	70.5	87.6	73.9	91.0	84.2
Qwen3-VL-Plus	69.7	59.1	84.1	71.0	91.9	**81.1**	95.9	89.6
Gemini 3 Flash	71.2	59.0	85.0	71.7	87.4	73.7	91.3	84.1
Gemini 3 Pro	**75.7**	66.4	**86.6**	**76.2**	91.9	**81.1**	95.9	89.6
Proposed (w/o GSPO)	71.6	66.2	83.1	73.6	91.1	77.7	94.2	87.7
Proposed (w/ GSPO)	71.1	**67.4**	83.2	73.9	**93.3**	79.9	**97.5**	**90.2**

Table 2: **Evaluation of reasoning quality** via MLLM-as-Judge. We report *Logical Correctness* (LC), *Visual Hallucination* (VH), *Medical Professionalism* (MP), and their **Average** score in percentage (%). Gray rows highlight the contribution of Forgery-aware GSPO.

5.3 Reasoning Quality

MedForge-Reasoner provides visually grounded reasoning for detection judgements. Figure 4 shows a comparison of reasoning outcomes of MedForge-Reasoner and SOTA baselines where MedForge-Reasoner achieves a clear advantage in providing hallucination-free and professional forgery explanations. We further quantitatively evaluate the quality of forgery explanations of all baselines in Table 2. For fair comparison, we randomly select 100 forgery samples where all models provide correct detections in the *In-Domain* setting.

The evaluation reveals that, in terms of reasoning quality, MedForge-Reasoner outperforms the best forgery detectors (AIGI-Holmes) by 16.2% and 31.2% as measured by the Gemini and Qwen judge. By incorporating the proposed GSPO, our model achieves a leading average Judge Score of 90.2% under Qwen3-VL-Plus and a competitive 73.9% under Gemini 3 Pro, outperforming the strongest baseline in the former case. Notably, the GSPO module provides a substantial boost to reasoning quality, increasing the average score by up to 2.5 percentage points compared to the version without GSPO. MedForge-Reasoner demonstrates superior performance in Logical Correctness and Medical Professionalism, while achieving a significant reduction in Visual Hallucination, with scores reaching 79.9% and 67.4% under the two judges respectively. This confirms that Forgery-aware GSPO effectively enforces visually grounded reasoning, ensuring the textual output is grounded in visual reality and aligns with medical expertise.

5.4 Ablation Studies

In this section, we conduct extensive ablation studies to validate the effectiveness of the proposed architecture and training strategies. To quantify the precision of forgery localization, we additionally report the Intersection over Union (IoU) between the predicted and ground-truth bounding boxes. The ablation studies consist of three parts: **Part (A)**

Setting	Acc (%)	F1 (%)	IoU	Judge Score
(A) Contribution of Reasoning Components				
Binary Classification	**99.42**	**99.31**	-	-
w/o Reasoning	99.31	99.10	0.30	-
w/o Bbox Grounding	99.31	98.97	-	53.9
Proposed	99.23	98.98	**0.31**	**90.2**
(B) Efficacy of Optimization Strategies				
SFT Cold-Start Only	98.20	98.40	0.30	87.4
GSPO w/o R_{bbox}	99.01	98.84	0.29	89.6
GSPO w/o R_{form}	99.13	98.98	0.30	90.1
Proposed	**99.23**	**98.98**	**0.31**	**90.2**
(C) Scalability across MLLM Backbones				
InternVL3.5-8B	96.92	97.66	0.32	85.8
Qwen2.5-VL-7B	93.17	94.69	**0.33**	80.4
MimoVL-7B	92.26	93.91	0.32	79.1
Qwen3-VL-8B (Ours)	**99.23**	**98.98**	0.31	**90.2**

Table 3: Ablation studies on model components, optimization strategies, and backbone architectures.

decomposes the model's response components to assess the necessity of localization and textual rationale; **Part (B)** isolates the benefits of the specific GSPO training objectives; and **Part (C)** tests the scalability and robustness of our method across different model architectures. Note that the reasoning quality is evaluated using the Qwen3-VL-Plus judge as described in Section 5.3.

Impact of Response Components. As shown in Table 3, the non-interpretable *Binary Classification* and *w/o Reasoning* have the highest detection performance, suggesting that the forgery bounding boxes and textual rationales might slightly interfere with the model's pure decision performance. However, as discussed, black box classification is insufficient for clinical reliability and trustworthy judgment. Crucially, while the *w/o Bbox Grounding* setting achieves marginally higher accuracy (+0.08%) than the proposed method, its Judge Score collapses to 53.9%. This discrepancy reveals that without explicit spatial supervision, the model tends to "hallucinate" justifications, correctly classifying images but for incorrect or non-verifiable reasons. The *Proposed* method achieves the highest Judge Score and IoU with a small accuracy trade-off (<0.2%), showing that MedForge-Reasoner successfully formulates a black-box detection task into interpretable reasoning grounded with factual visual evidence.

Efficacy of GSPO Optimization. Part B disentangles the contributions of our training objectives. Although the *SFT Cold-Start* establishes a strong baseline with 98.20% accuracy, it lags in reasoning quality. Incorporating the proposed GSPO significantly boosts performance. Specifically, removing the spatial reward (*GSPO w/o* R_{bbox}) results in a 0.02 decrease in IoU, verifying that R_{bbox} is essential for forgery localization. Similarly, removing the format reward (*GSPO w/o* R_{form}) leads to a slight degradation in accuracy (99.13% vs 99.23%), affecting the logical coherence of the output. The full GSPO framework achieves the best balance, yielding the highest Judge Score of *90.2* and Accuracy of *99.23%*.

Scalability across MLLM Backbones. In Part C, we assess the robustness of our method across different architectures. While *InternVL3.5-8B* shows competitive performance (96.92% Acc), our *Qwen3-VL-8B* based model outperforms it by over 2.3%. Interestingly, although *Qwen2.5-VL-7B* achieves the highest raw IoU (0.33), its reasoning capability is significantly weaker, evidenced by a low Judge Score of 80.4% and Accuracy of 93.17%. Our proposed method, leveraging the Qwen3-VL backbone, successfully bridges this gap, offering the optimal trade-off between geometric precision and semantic reasoning.

6 Conclusion

In this work, we presented a framework to safeguard the trustworthiness of medical imaging against the evolving threat of advanced deepfakes. We established **MedForge-90K**, the first large-scale medical forgery benchmark with high-fidelity lesion manipulations granularly annotated with expert-aligned reasoning. Addressing the limitations of black-box detectors and hallucination-prone MLLMs, we proposed **MedForge-Reasoner**, a novel detector capable of *pre-hoc* reasoning. By introducing the Forgery-aware GSPO, we successfully aligned the model's textual outputs with factual visual evidence, explicitly enforcing the detector to localize anomalies before reasoning. Extensive experiments demonstrate that our approach not only achieves state-of-the-art detection performance across unseen forgeries and architectures but also provides clinically rigorous, hallucination-free explanations. We hope this work bridges the gap between AI-driven forgery detection and clinical interpretability, offering a trustworthy solution for high-stakes healthcare environments.

7 Limitations

We discuss three main limitations of our work. First, MedForge-90K currently focuses on three common 2D imaging modalities: chest X-ray, brain MRI, and fundus photography. Although our framework is not modality-specific in principle, extending the benchmark to additional modalities (e.g., CT and ultrasound) and their corresponding forgery patterns would improve coverage of real-world clinical settings. Second, our reasoning and explanations are generated in English, consistent with most prior work. This choice limits the usability of MedForge-Reasoner in non-English clinical environments. A natural direction for future work is to support multilingual explanations, enabling broader deployment across global healthcare contexts. Third, while MedForge-Reasoner is designed as a trustworthy medical deepfake detector, it could potentially be misused for malicious purposes, such as improving forgery techniques to evade detection. It is therefore necessary to enforce responsible usage for our released models.

8 Acknowledgement

This research is supported by A*STAR, CISCO Systems (USA) Pte. Ltd., and the National University of Singapore under its Cisco-NUS Accelerated Digital Economy Corporate Laboratory (Award I21001E0002), the National Medical Research Council (NMRC) Healthy And Meaningful Longevity - Cognition Grant (NICCOG2024-0028) and the Talent Development Award Project of SSH-SPH (26-0065-A0001).

References

Saleh Albahli and Marriam Nawaz. 2024. Mednet: Medical deepfakes detection using an improved deep learning approach. *Multimedia Tools and Applications*, 83(16):48357–48375.

Alaa Alsaheel, Reem Alhassoun, Reema Alrashed, Noura Almatrafi, Noura Almallouhi, and Saleh Albahli. 2023. Deep fakes in healthcare: How deep learning can help to detect forgeries. *Computers, Materials Continua*, 76:2461–2482.

Ehsan Amiri, Ahmad Mosallanejad, and Amir Sheikhahmadi. 2024. The optimal model for copy-move forgery detection in medical images. *Journal of Medical Signals Sensors*, 14(2):5.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Saahil Chadha, David Weiss, Anastasia Janas, Divya Ramakrishnan, Thomas Hager, Klara Osenberg, Klara Willms, Joshua Zhu, Veronica Chiang, Spyridon Bakas, et al. 2025. An 11,000-study open-access dataset of longitudinal magnetic resonance images of brain metastases. *arXiv preprint arXiv:2506.14021*.

Zhihui Chen and Mengling Feng. 2025. Med-banana-50k: A cross-modality large-scale dataset for text-guided medical image editing. *arXiv preprint arXiv:2511.00801*.

Zhihui Chen, Kai He, Yucheng Huang, Yunxiao Zhu, and Mengling Feng. 2025. DivScore: Zero-shot detection of LLM-generated text in specialized domains. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19231–19253, Suzhou, China. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. 2024. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 1107–1128.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Virginia Fernandez, Pedro Sanchez, Walter Hugo Lopez Pinaya, Grzegorz Jacenków, Sotirios A Tsaftaris, and Jorge Cardoso. 2023. Privacy distillation: reducing re-identification risk of multimodal diffusion models. *arXiv preprint arXiv:2306.01322*.

Yueying Gao, Dongliang Chang, Bingyao Yu, Haotian Qin, Lei Chen, Kongming Liang, and Zhanyu Ma. 2025. Fakereasoning: Towards generalizable forgery detection and reasoning. *arXiv preprint arXiv:2503.21210*.

Google DeepMind. 2025. Gemini 3 Pro Model. urlhttps://deepmind.google/models/gemini/pro/. Accessed on 26 December 2025.

Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, et al. 2025. Maisi: Medical ai for synthetic imaging. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4430–4441. IEEE.

Kai He, Rui Mao, Tieliang Gong, Erik Cambria, and Chen Li. 2022. Jcbie: A joint continual learning neural network for biomedical information extraction. *BMC bioinformatics*, 23(1):549.

Kai He, Rui Mao, Qika Lin, et al. 2025. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Information Fusion*, 118:102963.

Abdullah Hosseini and Ahmed Serag. 2025. Is synthetic data generation effective in maintaining clinical biomarkers? investigating diffusion models across diverse imaging modalities. *Frontiers in Artificial Intelligence*, 7:1454441.

Chia-Chi Hsu, Min-Yan Tsai, and Chia-Mu Yu. 2025. Securing healthcare data integrity: Deepfake detection using autonomous ai approaches. *IEEE journal of biomedical and health informatics*.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Gong. 2024. Visual hallucinations of multi-modal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9614–9631.

Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Liangliang Cao, and Shifeng Chen. 2025a. Diffusion model-based image editing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. 2025b. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28831–28841.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. 2025. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *Preprint*, arXiv:2506.15742.

Ning Li, Tao Li, Chunyu Hu, Kai Wang, and Hong Kang. 2020. A benchmark of ocular disease intelligent recognition: One shot for multi-disease detection. In *International symposium on benchmarking, measuring and optimization*, pages 177–193. Springer.

Shuaibo Li, Zhaohu Xing, Hongqiu Wang, Pengfei Hao, Xingyu Li, Zekai Liu, and Lei Zhu. 2025. Toward medical deepfake detection: A comprehensive dataset and novel method. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 626–637. Springer.

Yufei Li, Xiaoyong Ma, Xiangyu Zhou, Pengzhen Cheng, Kai He, and Chen Li. 2021. Knowledge enhanced lstm for coreference resolution on biomedical texts. *Bioinformatics*, 37(17):2699–2705.

Qika Lin, Tianzhe Zhao, Kai He, Zhen Peng, Fangzhi Xu, Ling Huang, Jingying Ma, and Mengling Feng. 2025a. Self-supervised quantized representation for seamlessly integrating knowledge graphs with large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13587–13602.

Qika Lin, Yifan Zhu, Xin Mei, et al. 2025b. Has multimodal learning delivered universal intelligence in healthcare? a comprehensive survey. *Information Fusion*, 116:102795.

William A Malila. 1980. Change vector analysis: An approach for detecting forest changes with landsat. In *LARS symposia*, page 385.

Saman Motamed, Patrik Rogalla, and Farzad Khalvati. 2021. Data augmentation using generative adversarial networks (gans) for gan-based detection of pneumonia and covid-19 in chest x-ray images. *Informatics in medicine unlocked*, 27:100779.

Msoud Nickparvar. 2021. Brain tumor mri dataset.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Daniel Schaudt, Christian Späte, Reinhold von Schwerin, Manfred Reichert, Marianne von Schwerin, Meinrad Beer, and Christopher Kloth. 2023. A critical assessment of generative models for synthetic data augmentation on limited pneumonia x-ray data. *Bioengineering*, 10(12):1421.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. 2025. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*.

Laura Stroebel, Mark Llewellyn, Tricia Hartley, Tsui Shan Ip, and Mohiuddin Ahmed. 2023. A systematic literature review on the effectiveness of deepfake detection techniques. *Journal of Cyber Security Technology*, 7(2):83–113.

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*, pages 28130–28139.

Hao Tan, Jun Lan, Zichang Tan, Ajian Liu, Chuanbiao Song, Senyuan Shi, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. 2025. Veritas: Generalizable deepfake detection via pattern-aware reasoning. *arXiv preprint arXiv:2508.21048*.

Lehan Wang, Chongchong Qi, Chubin Ou, Lin An, Mei Jin, Xiangbin Kong, and Xiaomeng Li. 2024. Multieye: Dataset and benchmark for oct-enhanced retinal disease recognition from fundus images. *IEEE Transactions on Medical Imaging*.

Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Wenjun Wu, Conghui He, and Weijia Li. 2025. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. *arXiv preprint arXiv:2503.14905*.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. 2025. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*.

Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. 2025. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. In *International Conference on Learning Representations*.

Jian Zhang, Zhangqi Wang, Haiping Zhu, Kangda Cheng, Kai He, Bo Li, Qika Lin, Jun Liu, and Erik Cambria. 2026a. Mars: Multi-agent adaptive reasoning with socratic guidance for automated prompt optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(19):16307–16315.

Jian Zhang, Zhiyuan Wang, Zhangqi Wang, Fangzhi Xu, Qika Lin, Lingling Zhang, Rui Mao, Erik Cambria, and Jun Liu. 2026b. Maps: Multi-agent personality shaping for collaborative reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(19):16316–16324.

Yuchen Zhang, Zeyu Gao, Kai He, Chen Li, and Rui Mao. 2025. From patches to wsis: A systematic review of deep multiple instance learning in computational pathology. *Information Fusion*, 119:103027.

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*.

A Expert Acknowledgements

We would like to express our sincere gratitude to a panel of medical experts for their generous support, valuable clinical insights, and important contributions to the formulation of the Medical Expert Deepfake Guideline. Their expertise greatly strengthened both the medical rigor of this work and the development of the MedForge-90K benchmark. In particular, we would like to acknowledge the following experts:

- **Dr. Yuling Xu** (Ophthalmologist, Guangdong Provincial People's Hospital) offered invaluable clinical insights into fundus image authentication, particularly regarding vascular network logic, the precise morphological and spatial distribution of specific lesions (e.g., macular drusen in AMD and diabetic microaneurysms), and the identification of texture smudging or repetitive artifacts during lesion removal. Dr. Xu's guidance was also instrumental in the construction and validation of the MedForge-90K dataset.
- **Dr. Haonan Cai** (Resident Physician, Department of Neurology, Sun Yat-sen Memorial Hospital, Sun Yat-sen University) contributed detailed diagnostic criteria for brain MRI forgeries, with particular emphasis on multi-sequence signal characteristics (T1, T2, and DWI) and anatomical mass effects such as midline shift.
- **Dr. Yongzhen Huang** (Resident Physician, Department of General Surgery, Nanfang Hospital of Southern Medical University) provided valuable perspectives on evaluating 3D projection logic in chest X-rays and the biological interconnectivity between lesions and surrounding tissues.
- **Dr. Minjing Zhuang** (Ophthalmologist, Guangdong Provincial People's Hospital) contributed important insights into pathological logic in fundus imaging, especially regarding the consistency between primary lesion characteristics and secondary imaging manifestations.
- **Dr. Yuhe Lan** (Resident Physician) highlighted the importance of chronological disease progression and emphasized that isolated lesions should remain consistent with secondary clinical and imaging findings.

We also extend our heartfelt thanks to the anonymous medical experts whose thoughtful feedback further refined our understanding of medical forgeries and helped shape the foundation of this interpretable benchmark.

B Additional Results

B.1 Human Evaluation on Reasoning Quality

To mitigate potential bias in MLLM-based automatic judges (e.g., stylistic or model-specific preferences), we additionally conducted a single-blind human evaluation. Two annotators with training in medical image annotation single-blindly evaluated the 100 randomly sampled cases using the same rating protocol as the MLLM-as-Judge evaluation in Section 5.3.

As shown in Table 4, our MedForge-Reasoner ranks 3rd best overall, particularly surpassing all specialized deepfake detectors by a large margin. Notably, MedForge-Reasoner achieves the highest score in avoiding Visual Hallucination (76.2%). While the benchmarked commercial MLLMs (such as the Qwen3-VL series) provide high explanation quality on correctly detected cases, they suffer from poor overall detection accuracy (as previously shown in Table 1). MedForge-Reasoner performs as a balanced solution combining robust detection performance with highly interpretable, hallucination-free reasoning.

Method	**Human Evaluation (%)**			
	LC	VH	MP	**Avg.**
SIDA-7B	52.9	52.3	53.4	52.9
FakeVLM	55.2	57.2	55.3	55.9
AIGI-Holmes	47.8	48.1	48.7	48.2
Qwen3-VL-Flash	75.7	75.8	75.6	75.7
Qwen3-VL-Plus	**77.6**	75.6	**75.8**	**76.3**
Gemini 3 Flash	64.4	63.8	64.4	64.2
Gemini 3 Pro	72.1	66.8	70.9	69.9
Proposed (w/o GSPO)	70.2	70.5	71.4	70.7
Proposed (w/ GSPO)	70.3	**76.2**	71.5	72.7

Table 4: **Human Evaluation on Reasoning Quality.** We report *Logical Correctness* (LC), *Visual Hallucination* (VH), *Medical Professionalism* (MP), and their **Average** score in percentage (%). Evaluations were conducted by trained human annotators in a single-blind setting following the same protocol as the MLLM judges.

B.2 OOD Generalization

To further evaluate whether detectors generalize beyond the training distribution, rather than relying on shortcut learning or generator-specific artifacts, we further benchmarked MedForge-Reasoner and major baselines on four unseen medical deepfake datasets. These datasets contain manipulations produced by unseen deepfake generators and exhibit artifact patterns that differ from MedForge-90K.

Specifically, we evaluate performance on Med-Diffu (Fernandez et al., 2023), Synth-Pneumonia (Schaudt et al., 2023), Synth-Biomarker (Hosseini and Serag, 2025), and Med-Banana (Chen and Feng, 2025). Among them, Synth-Pneumonia and Synth-Biomarker are fake-only datasets, we randomly sample real images from MIMIC-CXR to construct a balanced test set with a 1:1 real-to-fake ratio.

As shown in Table 5, MedForge-Reasoner achieves the best overall detection performance, maintaining both Accuracy and F1 score above 89% across all four OOD datasets. In contrast, existing baselines exhibit substantial performance degradation under unseen manipulations. Particularly, FakeVLM collapses to an all-real prediction regime, resulting in a near-zero F1 score.

These results suggest that MedForge-Reasoner captures more generalizable forensic cues of medical image tampering, instead of overfitting to dataset or generator-specific characteristics.

Method	Med-Diffu	Synth-Pneu	Synth-Bio	Med-Bana
Accuracy (%)				
FakeVLM	50.0	50.0	50.0	54.1
AIGI-Holmes	51.6	51.2	49.8	49.6
Qwen3-VL-Plus	52.5	50.5	50.2	51.0
Gemini 3 Pro	<u>90.5</u>	<u>80.7</u>	<u>74.9</u>	<u>70.8</u>
Proposed	**90.7**	**94.6**	**92.3**	**99.6**
F1 Score (%)				
FakeVLM	0.0	0.0	0.0	15.0
AIGI-Holmes	63.3	63.3	62.3	61.2
Qwen3-VL-Plus	7.7	1.0	0.4	4.2
Gemini 3 Pro	**90.1**	<u>77.5</u>	<u>69.3</u>	<u>63.6</u>
Proposed	<u>89.8</u>	**94.3**	**91.7**	**99.6**

Table 5: **OOD Evaluation on Detection Performance.** We assess generalizability across four unseen medical deepfake datasets. **Bold** indicates the best result, and <u>underline</u> denotes the second-best.

B.3 Failure Mode Analysis

To better understand the remaining errors of MedForge-Reasoner, we conduct a detailed failure mode analysis on the 30,000-sample test set, focusing on three aspects: pathology difficulty, lesion implantation versus removal, and lesion size. Overall, the model exhibits highly consistent performance across edit types and modalities. The accuracy gap between implant and removal forgeries is below 0.2% in all modalities, suggesting that the detector does not exhibit a strong bias toward either insertion or deletion manipulations. In terms of lesion scale, large lesions ($\geq$8% of image area) are slightly easier to detect than small/medium lesions ($\leq$8%), achieving 99.77% versus 99.49% accuracy. Across the whole test set, we observe 140 deepfakes misclassified as real (71 implant, 69 removal) and 89 real images misclassified as deepfake, corresponding to an overall error rate of 0.76%.

Table 6 breaks down fake-image accuracy by modality and edit type. The results confirm that the implant–removal gap is consistently small, while the most challenging fake pathology varies across modalities. In particular, MRI remains the most difficult modality, with *Glioma* being the hardest fake category and also having the smallest average GT forgery area ratio.

Modality	Implant	Removal	Area
CXR	99.37	99.47	6.05
MRI	98.65	98.63	4.29
Fundus	99.82	99.78	5.67

Table 6: **Failure-mode breakdown by modality and edit type.** Values denote fake-image accuracy (%) for implant/removal and average ground truth forged-area ratio (%).

We further examine the hardest authentic categories that cause false alarms. As shown in Table 7, the dominant source of false positives comes from *Fundus Normal*, which alone contributes 55 errors. This suggests that the detector can occasionally overreact to naturally occurring retinal textures or subtle acquisition variations in healthy fundus images. MRI *Meningioma* and *Pituitary Adenoma* are also among the more difficult authentic categories.

Table 8 reports the hardest fake categories that are missed by the detector. These errors are concentrated in subtle or anatomically ambiguous pathologies, especially MRI *Glioma*, *Meningioma*, and *Pituitary Adenoma*. We conjecture that such cases are

Rank	Modality	Pathology	Accuracy (%)	Errors
1	Fundus	Normal	97.60	55
2	Fundus	Myopia	98.68	1
3	MRI	Meningioma	98.73	7
4	MRI	Pituitary Adenoma	98.83	7
5	CXR	Pneumothorax	99.07	3
6	CXR	Pleural Effusion	99.10	3
7	CXR	Lung Lesion	99.32	2
8	CXR	Fracture	99.34	2
9	CXR	Enlarged Cardiome-diastinum	99.34	2
10	Fundus	Glaucoma	99.40	1

Table 7: **Top-10 hardest authentic categories** that are misclassified as deepfake (false positive).

more difficult because their forgeries may preserve coarse anatomical structure while only weakly violating secondary pathological effects, making them challenging even under the proposed localize-then-analyze protocol.

Rank	Modality	Pathology	Accuracy (%)	Errors
1	MRI	Glioma	98.34	34
2	CXR	Fracture	98.39	10
3	MRI	Meningioma	98.68	29
4	CXR	Pneumothorax	98.87	7
5	MRI	Pituitary Adenoma	98.89	23
6	CXR	Atelectasis	99.10	6
7	CXR	Pneumonia	99.38	4
8	CXR	Consolidation	99.39	4
9	Fundus	Glaucoma	99.54	6
10	CXR	Lung Opacity	99.54	3

Table 8: **Top-10 hardest fake categories** that are misclassified as real (false negative).

Overall, these results show that the dominant residual errors are not caused by a systematic failure on one manipulation direction, but are instead concentrated on a small number of subtle authentic fundus images and anatomically challenging MRI forgeries. This also explains why the overall performance remains near-ceiling while still leaving room for improvement on hard boundary cases.

B.4 Localization IoU Analysis and BBox Diagnostics

Because our GSPO reward uses one-sided forgery coverage rather than strict IoU, a natural concern is whether the model could exploit the objective by predicting overly large bounding boxes. To directly assess this issue, we report the localization IoU statistics on true-positive forgery test samples in Table 9. Importantly, GSPO improves localization quality rather than collapsing to trivial large-box solutions, which suggests that reward hacking does not dominate in our current setting.

Mean	Std	Median	95th	> 0.25	$> 0.5/0.75$
31.55	29.22	25.06	87.33	50.0%	27.2% / 11.7%

Table 9: **Localization IoU statistics** on true-positive forgery test samples ($n = 20{,}000$). IoU summary values are reported in percentage.

The IoU distribution reveals two important observations. First, localization is often coarse rather than pixel-tight: while the mean IoU is 31.55% and the median is 25.06%, half of the correctly detected fake samples still achieve IoU above 0.25. Second, a non-trivial subset of cases is localized very accurately, as reflected by the 95th percentile of 87.33%. This behavior is consistent with the design of our method: the reward primarily encourages the model to attend to the correct forensic evidence region before reasoning, rather than optimizing for object-detection-style boundary precision.

Finally, although the current results do not indicate systematic reward hacking, the one-sided coverage design may still admit oversized-box solutions in principle. A natural direction for future improvement is to combine the current grounding reward with explicit size-aware regularization, such as area penalty or area-normalized overlap, in order to better suppress trivial large-box predictions while preserving the stability advantages of coverage-based optimization.

C MedForge-90K Implementation Details

To construct high-fidelity and anatomically plausible medical forgeries, we implemented a rigorous pipeline involving automated prompt engineering and diverse image generation models. This section details the specific implementations of the prompt generation, the iterative refinement loop, and the generator models used.

C.1 Data Collection

To ensure the authenticity of the source material and the clinical relevance of the forgeries, we curated a diverse collection of 31,990 high-resolution medical images from public benchmarks. As detailed below, our collection spans three imaging modalities and covers 19 distinct pathologies, along with healthy images for each modality.

Chest X-Ray (CXR) We sourced frontal-view radiographs from the MIMIC-CXR dataset (Johnson et al., 2016). To facilitate precise lesion removal and implantation, we specifically filtered for scans annotated with exactly one positive pathology. The subset includes 11 thoracic conditions: *Atelectasis*, *Cardiomegaly*, *Consolidation*, *Pulmonary Edema*, *Enlarged Cardiomediastinum*, *Rib Fracture*, *Lung Lesion*, *Lung Opacity*, *Pleural Effusion*, *Pneumonia*, and *Pneumothorax*. Healthy control images were selected from the *No Finding* category.

Brain MRI Magnetic Resonance Imaging data was sourced from the Brain Tumor Classification dataset (Nickparvar, 2021) and Yale-Brain (Chadha et al., 2025). We focused on contrast-enhanced MRI scans, organizing them into 3 specific tumor typologies: *Glioma*, *Meningioma*, and *Pituitary Tumor*. A corresponding set of healthy brain scans was collected under the *Healthy Control* (No Tumor) category to serve as the baseline for tumor implantation tasks.

Fundus Photography Retinal images were collected from the ODIR-5K (Ocular Disease Intelligent Recognition) dataset (Li et al., 2020) and MultiEYE (Wang et al., 2024). We categorized the data into 5 major ocular pathologies based on the diagnostic labels: *Age-related Macular Degeneration (AMD)*, *Diabetic Retinopathy*, *Glaucoma*, *Hypertensive Retinopathy*, and *Pathological Myopia*. The *Normal* category was used for healthy reference images.

Preprocessing To ensure compatibility with high-fidelity diffusion models, all raw images underwent a standardization pipeline. Images were resized and padded to a uniform resolution of 1024×1024 pixels, strictly preserving the original aspect ratio to maintain anatomical integrity before being fed into the forgery generation loop.

Dataset inventory. Table 10 summarizes verified RGB image counts from the MedForge-90K release manifest. Across authentic and forged splits, the benchmark comprises **95,276** images in total (31,990 real & 63,286 deepfake).

C.2 Forgery Prompt Generation

To guide the image editing models in performing precise lesion manipulation, we utilize a Large Language Model (Gemini 2.5 Pro) acting as the *Writer*. The goal is to translate medical tasks (e.g., "Implant Pleural Effusion" or "Remove Brain Tumor") into natural language instructions understandable by text-guided image editing models.

The generation process adheres to three critical constraints to ensure the output is realistic and undetectable as a deepfake:

1. **Fidelity Preservation:** The prompt must explicitly instruct the editor to preserve original image noise, grain texture, and contrast, avoiding alterations to image.

2. **Negative Rules:** We enforce strict negative constraints, forbidding the addition of text, labels, or unnatural sharp boundaries.

3. **Minimal Change Principle (Counterfactual Minimality):** The prompt emphasizes modifying *only* the pixels necessary for the pathology, leaving the background and surrounding anatomy untouched.

For **Lesion Implant**, the system instruction provided to the *Writer* is:

Lesion Implant Prompt Generation

"You are a medical image editing expert. Generate a clear, concise prompt to edit a normal [Modality] image to show [Disease]. [...] Critical Constraints - Fidelity Preservation: Preserve original image noise, grain texture, and contrast. Do not alter device artifacts... Critical Constraints - Minimal Change Principle: Only modify areas directly related to [Disease]. Keep all other anatomical structures unchanged... Add a clear warning: do not edit any element other than adding the disease feature. Keep everything else in the image exactly the same."

For **Lesion Removal**, the instruction shifts to describing the removal of specific anomalies without leaving inpainting traces:

Lesion Removal Prompt Generation

"Generate a clear, concise prompt to edit a [Modality] image showing [Disease] to make it appear normal (healthy). [...] The prompt should not include any medical-related terms... and only describe the direct modifications in the simplest way (e.g., 'delete the white rounded shape'). The prompt should focus on locating the lesion and describing the boundary... and not specify the replacement content."

C.3 Forgery Prompt Refinement

Forgery generator could fail to produce medically accurate or visually seamless results. To address this, we implement a *Writer-Editor-Diagnoser* refinement loop.

Deepfake Model	Dataset ID	Total	Implant	Remove
Stable Diffusion Inpainting (v1.5)	`stable-diffusion-inpainting`	6,303	3,239	3,064
Stable Diffusion XL Inpainting 0.1	`stable-diffusion-xl-1.0-inpainting-0.1`	6,380	3,273	3,107
FLUX.1-dev	`flux.1-dev`	6,399	3,273	3,126
Stable Diffusion 3.5 Large	`stable-diffusion-3.5-large`	6,383	3,273	3,110
Stable Diffusion 3.5 Medium	`stable-diffusion-3.5-medium`	6,383	3,273	3,110
GPT-Image	`gpt`	6,340	3,223	3,117
Gemini 2.5 Flash Image	`gemini`	6,413	3,273	3,140
Qwen-Image-Edit	`qwen`	6,347	3,213	3,134
Seedream 4.0	`seeddream`	6,155	3,131	3,024
Step1X-Edit	`step1x-edit`	6,183	3,223	2,960
All Models (forged)	—	**63,286**	**32,394**	**30,892**
Real (authentic)	—	**31,990**	—	—
Total	—	**95,276**	—	—

Table 10: **MedForge-90K image inventory.** Per-editor RGB forgeries after the Writer–Editor–Diagnoser loop; *Implant* and *Remove* match the `*-edit` and `*-remove` splits. Deepfake models follow the presentation order in Appendix C.4 (inpainting → advanced diffusion → proprietary APIs), with Step1X-Edit listed last as an additional open instruction editor. Footer rows report forged aggregates, the authentic split total, and the full RGB inventory (*Implant*/*Remove* are undefined for real and for the dataset-wide total).

The Verification Loop (Diagnoser) In each iteration, the *Editor* generates a candidate image. A *Diagnoser* (Gemini 3 Pro) then performs a pixel-level side-by-side comparison between the original and the forged image to ensure the pathology is added/removed correctly without affecting the background. The specific instruction used for this verification is:

Forgery Verification Instruction

"You are a medical image verification expert. You are given two images: 1. Original image ... 2. Edited image ... Your task is to verify the editing quality by comparing the two images side-by-side.
Critical Verification - Minimal Change Principle: Compare the original and edited images carefully. The editing should only modify the disease-related regions. Check: - Are non-disease areas (background, other anatomical structures, imaging artifacts) identical? - Does the edited version preserve the exact same imaging characteristics (noise, grain, contrast)? Check these aspects: 1. Has disease: Does the edited image show signs of [Disease]? 2. Structure reasonable: Are the anatomical structures reasonable and correct? 3. Looks realistic: Does the edited image look like a real medical image? 4. Minimal changes preserved: Are changes limited only to disease areas? Return your evaluation in this JSON format: { "qualified": true/false, "reason": "..." }"

The Prompt Refinement (Writer) If the verification fails (e.g., due to artifacts or incorrect anatomy), the execution history and failure reasons are fed back to the *Writer*. The *Writer* is then prompted to analyze the previous failures and generate an improved prompt:

Forgery Prompt Refinement Instruction

"You are a medical image editing expert. Multiple previous editing attempts have failed. You need to analyze ALL previous attempts and generate a BETTER prompt.
History of all previous attempts: [History Log]
Looking at the ORIGINAL image and analyzing the patterns of failures above, generate an IMPROVED editing prompt. ANALYSIS REQUIREMENTS: 1. Identify common issues across multiple attempts 2. Learn from what didn't work in previous rounds 3. Avoid repeating the same mistakes
[...] (Standard constraints on Fidelity Preservation and Minimal Change Principle are repeated here)
Return ONLY the editing prompt in English, no explanations."

This loop repeats for up to 5 rounds. Only images that pass the strict verification criteria ("qualified": true) are included in the MedForge-90K.

C.4 Forgery Generation

Once the prompts are refined and validated, we employ a diverse ensemble of 10 state-of-the-art image editing and generation models to construct the final MedForge-90K dataset. Using a wide range of architectures prevents the detector from overfitting to specific generator artifacts (e.g., specific noise patterns of a single diffusion model). The models utilized are categorized as follows:

Diffusion-based Inpainting Models These models require a mask (derived from the Nano-Banana coordinates) and the refined text prompt to regenerate specific regions.

- **Stable Diffusion Inpainting (SD-v1.5)**: A baseline latent diffusion model specialized for mask-based editing (Rombach et al., 2022).
- **Stable Diffusion XL (SDXL) Inpainting 0.1**: A larger scale model (2.6B parameters) capable of generating higher resolution details and better texture matching in medical scans (Podell et al., 2023).

Advanced Diffusion-based Image Editing Models These models perform instruction-based editing without needing explicit masks, relying on the refined prompts to localize and modify content.

- **FLUX.1-dev**: A 12B parameter rectified flow transformer model. It is chosen for its superior prompt adherence and ability to generate high-frequency details (noise/grain) crucial for medical realism (Labs et al., 2025).
- **Stable Diffusion 3.5 Large**: The latest Multimodal Diffusion Transformer (MMDiT) from Stability AI, offering state-of-the-art conceptual understanding of complex prompts (Esser et al., 2024).
- **Stable Diffusion 3.5 Medium**: A distilled version of SD3.5, providing a variation in generation artifacts to test detector robustness against model compression traces (Esser et al., 2024).

MMDiT Image Editing Models We also utilize closed-source or specialized APIs to capture the distribution of commercial deepfake tools.

- **GPT-Image**: Accessed via OpenAI API. Known for high semantic understanding, used primarily for complex lesion removal tasks where context reasoning is required (Hurst et al., 2024).
- **Gemini-2.5-Flash-Image (Nano-Banana)**: Accessed via Google GenAI API. Utilized for its strong instruction-following capabilities in medical contexts (Comanici et al., 2025).
- **Qwen-Image-Edit**: Based on the Qwen-Image architecture, this model integrates visual understanding with generation, allowing for precise editing based on visual cues (Wu et al., 2025). It stands out as the SOTA open-source image editor currently, making it crucial to evaluate on medical forgery detection.
- **Seedream 4.0**: A high-performance multimodal image generation model designed for high-consistency semantic editing, minimizing changes to the background (Seedream et al., 2025). Seedream 4.0 is pretrained on billions of text-image pairs spanning diverse taxonomies and knowledge-centric concepts, suitable for high-quality medical forgeries.

This ensemble ensures that MedForge-90K covers open-source latent diffusion models and proprietary MMDiT-based generators, representing a comprehensive threat landscape.

C.5 Reasoning Annotation

To equip MedForge-90K with granular and clinically grounded explanations, we developed an automated annotation pipeline utilizing advanced MLLMs (Gemini 2.5 Pro). Unlike standard captioning tasks, our pipeline employs a **Hierarchical Guideline-Driven Reasoning** strategy. This mechanism enforces the model to scrutinize images through a three-tiered cognitive framework derived directly from our expert guidelines (detailed in Section C.6):

- **Level 1: Image Physics & Texture.** Detecting low-level anomalies such as "sticker" artifacts, unnatural noise distribution, or inpainting smudges that violate the physical properties of medical imaging.
- **Level 2: Anatomical Structure.** Verifying morphological correctness, such as the continuity of vascular networks in fundus photography or the symmetry of gyri in brain MRI.
- **Level 3: Pathological Logic.** Checking high-level biological interconnectivity to ensure lesions exhibit necessary secondary effects (e.g., mass effect, edema, chronological progression) rather than in biological isolation.

Annotation for Authentic Images For real images, the pipeline shifts to validating *Biological Consistency*. The prompt directs the MLLM to confirm the *satisfaction* of the hierarchical logic—verifying that noise patterns are stochastic, anatomy is continuous, and pathological signs follow a natural progression. This ensures the detector learns the logic of authenticity, distinct from the features of forgery.

Annotation for Forged Images For images in the *Lesion Implant* and *Removal* categories, the annotation is spatially grounded using the ground-truth manipulation mask for hierarchical analysis:

1. **Bbox Extraction:** Bounding box coordinates $\mathbf{b} = [x_1, y_1, x_2, y_2]$ are extracted from the binary manipulation mask.

2. **Hierarchical Prompting:** The prompt that explicitly informs the MLLM of the forgery location. Crucially, we inject the specific *Expert Forgery Guidelines* into the prompt context. The model is instructed to analyze the image area within $\mathbf{b}$ specifically for violations across the previously defined logic levels.

Hierarchical Forgery Reasoning Prompt Template

System: This is a medical deepfake image. The bounding box $[y_{min}, x_{min}, y_{max}, x_{max}]$ indicates the location of the deepfake region. **Task:** Analyze why this image is a deepfake by systematically applying the *Medical Deepfake Detection Guidelines* provided below. **Requirements:** 1. **Location:** Output the coordinates in <box> format. 2. **Description:** Briefly describe the image modality and features. 3. **Key Explanation:** Identify anomalies following the hierarchical logic: - *Physics:* Are there noise/texture artifacts or sharp boundaries? - *Anatomy:* Are structures morphologically incorrect? - *Pathology:* Are biological secondary signs missing (e.g., lack of mass effect)? 4. **Conclusion:** Definitive statement of forgery. **Context:** [Injected Modality-Specific Guidelines from Section C.6]

C.6 Medical Deepfake Detection Guidelines

To ensure the reasoning annotations described above align with clinical expertise, we formulated a comprehensive set of detection criteria. These guidelines serve as the "ground truth logic" injected into the annotation prompts.

C.6.1 General Principles

This section applies to all image modalities, focusing on the failure of AI forgery to replicate biological interconnectivity and physical consistency.

Biological Plausibility & Secondary Effects

- **Mass Effect Absence:** Real lesions are physical objects that displace tissue. Reject if a space-occupying lesion exists without corresponding compression, displacement, or midline shift.
- **Lack of Host Reaction:** The body reacts to pathology. Reject if an aggressive lesion appears "isolated" with a sharp boundary and no surrounding edema or infiltration.
- **Chronological Inconsistency:** Diseases follow a timeline. Reject if late-stage features appear without precursor signs (e.g., neovascularization without ischemia).

Image Physics & Texture Consistency

- **The "Sticker" Artifact:** Reject if the lesion-background interface is unnaturally sharp, lacking the gradual transition zone of biological tissues.
- **Noise Distribution Analysis:** Reject if the noise pattern (grain) within the lesion is significantly smoother or different in texture compared to the surrounding unaffected tissue.
- **Inpainting Artifacts:** In removal cases, look for "smudging," blurring, or repetitive cloning patterns that disrupt natural stochastic texture.

C.6.2 Modality-Specific Principles

These criteria address the specific anatomical and structural logic required for each imaging type.

Brain MRI

- **Anatomical Logic:** Sulci adjacent to a mass should be effaced. The ventricular system must be symmetrical unless physically displaced. Large unilateral masses must cause a contralateral midline shift.
- **Signal Intensity:** Peritumoral edema must follow correct signal intensity (e.g., Hyperintense on T2/FLAIR). Lesions must match specific signatures (e.g., Meningiomas require a "Dural Tail").
- **Multi-Sequence Consistency:** Lesion appearance must logically translate across sequences (e.g., fluid is bright on T2, dark on T1).

Fundus Photography

- **Vascular Logic:** Vessels must taper gradually from the optic disc to the periphery without discontinuities. The Artery/Vein (A/V) ratio must be consistent.
- **Lesion Distribution:** Diabetic lesions usually spare the extreme periphery initially. Drusen must be concentrated in the Macula. Macular exudates should form a "Star" pattern due to Henle's fiber layer.
- **Global Physics:** The image must exhibit natural vignetting (e.g., posterior pole brighter than periphery).

Chest X-Ray (CXR)

- **3D Projection Logic:** Lung markings must correctly overlap with ribs/heart. Skeletal structures (rib count, clavicle shape) must be anatomically correct.

- **Density Gradient:** Adherence to the density ladder (Air $<$ Fat $<$ Bone). Vascular markings should be more prominent in lower zones.
- **Secondary Signs:** Atelectasis must show volume loss (elevated diaphragm). Cardiomegaly should manifest with pulmonary congestion.

D Experiment Settings

D.1 Baselines: Generic MLLM

To evaluate the zero-shot and in-context reasoning capabilities of state-of-the-art models in medical deepfake detection, we employ four representative Multi-modal Large Language Models (MLLMs). These include the **Qwen3-VL** series and the **Gemini 3** series, known for SOTA image understanding and reasoning abilities.

D.1.1 Model Settings

All models are accessed via their respective official APIs to ensure reproducibility. The specific models and their configurations are as follows:

- **Qwen3-VL-Flash & Qwen3-VL-Plus**: Accessed via Qwen API.
- **Gemini 3 Flash & Gemini 3 Pro**: Accessed via the Google Generative AI (GenAI) SDK.

For all API calls, we set the `temperature` to 0.1 to minimize stochasticity and encourage deterministic, logical outputs. The maximal number of output tokens is set to 1024 to accommodate the detailed judgement explanations.

D.1.2 In-Context Learning Prompts

We design three distinct levels of In-Context Learning (ICL) prompts to evaluate the model's generalization capability across different forensic scenarios. These prompts are generated using a "Forensics Expert" agent (powered by Gemini 3 Pro) based on a selected set of real and manipulated medical examples.

1. **In-Domain ICL Prompt**: Contains comprehensive guidance covering all available modalities (CXR, MRI, Fundus) and all generator architectures (SD, Flux, GANs). It serves as the upper bound for model performance when full forensic knowledge is available.
2. **Cross-Model ICL Prompt**: Excludes specific generative models (e.g., Stable Diffusion, GPT-based generators) from the context to test if the MLLM can generalize forensic principles to "unseen" generator artifacts.
3. **Cross-Forgery ICL Prompt**: Focuses primarily on one type of manipulation (e.g., lesion removal) while excluding others (e.g., implants/edits), evaluating the model's ability to identify fundamental biological inconsistencies regardless of the forgery task.

D.2 Baselines: Specialized Detectors

To ensure a fair comparison, all specialized baseline detectors were trained on the MedForge-90K training set. We followed the official implementations and recommended hyper-parameters provided by the respective authors, adapting them to medical forgery detection. All models were trained on 8 NVIDIA H100 80G GPUs, requiring around 10-15 hours per model.

SIDA-7B & 13B SIDA (Huang et al., 2025b) is an MLLM-based detector designed for forgery detection and localization. We utilized the LLaVA-v1.5 [7B/13B] as the backbone.

- **Training Stage:** We performed default LoRA training (rank=128 alpha256) on the MedForge-90K SFT split.
- **Hyper-parameters:** Following the same setting as proposed detector, SIDA 7/13B were trained for 10 epochs with a total batch size of 8. We used the AdamW optimizer with a learning rate of 2e-5.
- **Original Setting:** Following the original implementation, the input resolution was set to 336×336, and the prompt followed the "instruction-reasoning-label" format as described in the original paper.

FakeVLM FakeVLM (Wen et al., 2025), a specialized large multimodal model designed for both general synthetic image and DeepFake detection.

- **Backbone:** We employed llava-1.5-7b as the detector backbone, following the original implementation
- **Training Stage:** The model underwent 10 stages of LoRA training (rank=128 alpha256), based on forgery label and textual descriptions formulated from the MedForge-90K SFT set.

AIGI-Holmes AIGI-Holmes (Zhou et al., 2025) utilizes a multi-stage framework consisting of a CLIP-based forgery Visual Expert and subsequent LLM reasoning explainer.

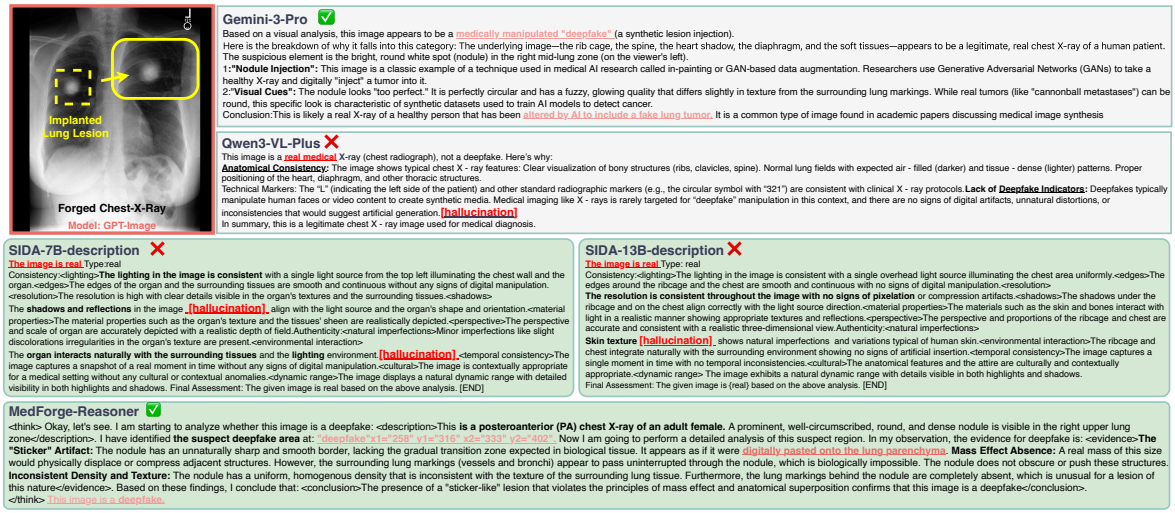

Figure 4: **Qualitative Forgery Explanation Comparison.** Baselines fail due to severe hallucinations (SIDA citing "skin texture") or missed diagnoses. While Gemini-3-Pro correctly detects the forgery using general visual clues, MedForge-Reasoner delivers superior *clinically rigorous rationale*, explicitly grounding the verdict in anatomical logic (e.g., "absence of mass effect") rather than generic visual analysis.

- **Backbone:** We employed CLIP and NPR network as the Visual Expert, and llava-v1.6-mistral-7b-hf as the LLM backbone following the original setting.
- **Training Stage:** The visual experts are trained following default configuration. The LLM module underwent 10 stages of LoRA training (rank=128 alpha256), based on the MedForge-90K SFT set.

D.3 Evaluation Metrics

We report the detection performance using Accuracy and F1 Score metrics. These are standard metrics calculated based on True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The formulas are defined below:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$\text{F1} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{12}$$

In the main experiment (Table 1), we report performance broken down by category. To ensure clarity, we define the specific positive and negative classes used for calculating metrics in each column:

- **Real:** This measures the model's ability to identify authentic images. Here, the positive class is the *Real* image, and the negative class includes all *Fake* images (comprising both Lesion Implant and Lesion Removal).
- **Forgery Implant:** This measures the model's ability to distinguish implanted lesions from healthy tissue. Here, the positive class is the *Lesion Implant* forgery, and the negative class is the *Real* image. Lesion Removal samples are excluded from this calculation to isolate the performance on implantation.
- **Forgery Remove:** This measures the model's ability to detect erased lesions. Here, the positive class is the *Lesion Removal* forgery, and the negative class is the *Real* image. Lesion Implant samples are excluded.

To quantitatively assess the quality of the generated forensic reasoning, we employ a reference-based evaluation protocol using state-of-the-art MLLMs (Qwen3-VL-Plus and Gemini 3 Pro) as impartial judges. Unlike standard n-gram metrics (e.g., BLEU, ROUGE) which fail to capture semantic consistency in medical diagnostics, our MLLM-as-Judge approach evaluates the *factual consistency* between the model's generated rationale and the Ground Truth forgery reasoning. As defined in our evaluation script, the judge scores each response on a scale of 1 to 10, which is then converted to a 100% scale for reporting. The MLLM-as-Judge is based on three distinct criteria:

1. **Logical Correctness:** Evaluates whether the assistant's reasoning follows a sound forensic process. It rewards responses that arrive at the correct conclusion through valid deduction, rather than lucky guesses.

2. **Visual Hallucination:** Measures the faithfulness of the description to the visual reality. A high score indicates the model describes only features present in the Ground Truth (e.g., specific bbox locations, noise patterns), while a low score indicates the fabrication of nonexistent features.
3. **Medical Professionalism:** Assesses whether the terminology (e.g., "mass effect," "vascular continuity") and diagnostic logic align with the provided expert medical guidelines.

Judge Prompt To ensure objectivity, the judge is provided with the specific role of a "Medical Image Forensics Expert." The exact prompt used in our evaluation pipeline is presented below:

MLLM-as-Judge System Prompt

Role: You are a Medical Image Forensics Expert acting as an impartial judge. Your expertise covers Radiology (MRI, CXR) and Ophthalmology (Fundus), specifically in identifying AI-generated (Deepfake) anomalies versus real pathological features.
Task: Please examine the provided text responses and serve as an unbiased judge in assessing the quality of a forensic analysis from an AI assistant. You will evaluate how well the assistant identifies and explains the forensic nature of the image manipulation based on professional medical imaging standards, compared to a Ground Truth reference.
Input Data:
Assistant Response: The forensic analysis provided by the AI assistant for evaluation. **Ground Truth Information:** The definitive expert reference explanation for the manipulations present in the image.
Evaluation Focus: Your evaluation should focus exclusively on the content and factual correctness of the assistant's response compared to the Ground Truth.
DO NOT reward for tedious and verbose responses.
DO focus on whether the assistant correctly identified the same forensic anomalies, biological evidence, and medical logic as described in the Ground Truth. Reward the response outputting correct bbox coordinate.
Evaluation Criteria:
Logical Correctness: Whether the assistant's reasoning follows a sound forensic process and arrives at the correct conclusion. **Visual Hallucination:** Whether the assistant's verbal description of the anomalies matches the ground truth or fabricates nonexistent features/locations. **Medical Professionalism:** Whether the terminology and medical logic used in the text align with expert guidelines.

D.4 MedForge-Reasoner Training

SFT Cold-Start Stage. We utilize the Qwen3-VL-8B-Instruct as the backbone model. We employ LoRA (Low-Rank Adaptation) for parameter-efficient fine-tuning, targeting all linear modules with a rank $r = 128$ and alpha $\alpha = 256$. The model is trained for 10 epochs using the AdamW optimizer with a learning rate of 1×10^{-4} and a cosine decay scheduler (warmup ratio set to 0.05). The training uses a global batch size of 512 (per-device batch size 16 with gradient accumulation) and bfloat16 precision. The maximum sequence length is set to 2048 to accommodate detailed reasoning chains.

Forgery-aware GSPO Stage. We initialize the model with the SFT checkpoint and align it using the Group Sequence Policy Optimization (GSPO) framework. The model is trained for 1 epoch with a reduced learning rate of 1×10^{-6}. We set the group size $G = 8$ to sample diverse reasoning paths for importance sampling. The KL-divergence penalty coefficient β is set to 0.001, and the sampling temperature is 1.0.

Reward Function Details. As implemented in our plugin, the total reward R is a weighted sum of four specific components designed to enforce structure, accuracy, and grounding:

- **Classification Reward (R_{clas}):** A dominant reward to ensure decision correctness. We assign $+4.0$ for correct predictions and -4.0 for incorrect ones.
- **Formatting Reward (R_{form}):** Capped at 1.0, this component rewards the presence of mandatory XML tags (e.g., `<think>`, `<evidence>`) and valid bounding box syntax (e.g., `<|box_start|>`).
- **Formatting Reward (R_{form}):** We also apply a strict format penalty of -1.0 if the detection verdict contradicts the localization output (e.g., predicting "Real" but generating bounding box coordinates, or predicting "Forgery" without coordinates).
- **Grounding Coverage Reward (R_{bbox}):** For correctly classified forgery samples, we reward the Mask Coverage $\mathcal{C}$ using a shaped sigmoid function:

$$R_{\text{bbox}} = \frac{0.25}{1 + e^{-10(\mathcal{C}-0.5)}} \tag{13}$$

This function scales the reward up to a maximum of 0.25, effectively penalizing low overlaps while saturating for high coverage.