Attacking Medical Vision–Language Models with Query-Based Zero-Order Optimization

Maxime Griot [1]

Abstract

Medical vision–language models (VLMs) are increasingly integrated into clinical workflows, yet their robustness to adversarial image manipulation remains poorly understood. Unlike text-based attacks that can often be detected through input logging, image-based perturbations introduce a distinct threat surface: bounded manipulations in the model's image input space that remain clinically plausible and resemble benign preprocessing or encoding artifacts can alter model predictions with direct implications for patient care. We show that adversarial attacks on medical VLMs are feasible under a black-box threat model with query access to next-token scores. Using zero-order optimization with a low-frequency prior on the MedXpertQA benchmark, we induce answer flips in restricted multiple-choice settings on two medical VLMs (`MedGemma 4B` and `MedGemma 27B`) and a generic VLM (`Qwen3-VL-Instruct 4B`). Under fixed perturbation budgets ($\epsilon \leq 0.02$), we achieve attack success rates exceeding 40% on initially correct examples across diverse imaging modalities, including radiology, pathology, and electrocardiography. Attacks are consistently more effective on color images than under luminance-constrained perturbations, reflecting reduced degrees of freedom in the latter setting. These results indicate that restricting access to model weights alone may be insufficient when deployments expose token-level scores, and motivate defenses that limit exposed inference signals and add safeguards for high-stakes decision support.

1. Introduction

Large language models (LLMs) have achieved expert-level performance on a wide range of medical benchmarks, positioning them as promising tools for extracting value from the large volumes of unstructured data contained in Electronic Health Records (EHRs) (Singhal et al., 2022; Saab et al., 2024). Early deployments of LLM-powered systems in clinical environments have shown rapid user adoption (Griot et al., 2025b; Epic, 2024), but have also exposed important security and safety gaps, driven in part by fast deployment cycles and persistent misconceptions about model capabilities (Gu et al., 2025). Adversarial evaluations consistently reveal a high susceptibility of LLMs to minor input perturbations and pattern variations (Gallifant et al., 2024; Griot et al., 2025c).

The extension of LLMs to multimodal models capable of processing images alongside text introduces both new opportunities and new risks. Vulnerabilities identified in text-based systems increasingly surface in image-based settings (Gu et al., 2025), including sensitivity to image artifacts such as motion, rotation, and noise (Cheng et al., 2025). In clinical contexts, multimodal systems offer a compelling pathway to integrating heterogeneous EHR data, particularly when imaging must be interpreted without accompanying written reports. Recent models already demonstrate performance surpassing that of medical trainees on radiology benchmarks (Datta et al., 2025a;b), and are increasingly integrated into radiology workflows as assistive tools whose outputs are reviewed by clinicians (Lawrence et al., 2025), often supported by interpretability techniques such as gradient-based visualizations (Huff et al., 2021).

Despite these advances, fundamental questions surrounding the reliability, trustworthiness, and safety of medical LLMs remain unresolved. Prior work has documented multiple classes of failure, including demographic biases (Zack et al., 2024), metacognitive limitations and systematic overconfidence (Griot et al., 2025a), and flawed or fabricated reasoning processes that nonetheless yield correct answers (Jin et al., 2024). Security threats are equally pressing: models may be compromised during training via data poisoning (Alber et al., 2025), or during inference through prompt injection attacks (Greshake et al., 2023; Lee et al., 2025). While text-based attacks can often be detected through input logging and access controls, they are comparatively difficult to execute within tightly regulated EHR infrastructures.

Medical vision–language models (VLMs) fundamentally

[1]Institute of NeuroSciences, Université catholique de Louvain, Brussels, Belgium. Correspondence to: Maxime Griot <maxime.griot@uclouvain.be>.

reshape this adversarial landscape. Unlike textual inputs, medical images are continuous, high-dimensional signals in which carefully crafted perturbations can remain imperceptible to human observers while substantially altering model behavior (Goodfellow et al., 2015). This introduces a distinct attack surface in which bounded, clinically plausible image manipulations—resembling benign preprocessing or encoding artifacts—can embed malicious signals capable of influencing downstream predictions, with direct implications for patient care (Zhao et al., 2023; Tu et al., 2023). The risk is further amplified by the growing deployment of health assistant models outside traditional clinical security and regulatory frameworks (Anthropic, 2026; OpenAI, 2026).

These concerns are particularly salient given the increasing availability of open and foundation models across medical modalities, including pathology (Xu et al., 2024), electrocardiography (Li et al., 2025), ophthalmology (Kim et al., 2025), and general medical imaging (Ma et al., 2025), alongside broader trends toward open model ecosystems (Riedemann et al., 2024). Recent work has shown that multimodal systems can be attacked without direct access to model parameters by exploiting cross-model transferability from open surrogate models (Jia et al., 2025). However, such approaches do not directly evaluate the robustness of closed-weight medical VLMs under realistic inference-time assumptions.

In this work, we provide a security evaluation of medical vision–language models under a black-box threat model with query access to next-token scores. We study adversarial attacks that operate directly on medical images without access to model weights, gradients, or surrogate models. We use restricted multiple-choice decoding as a methodological probe: by constraining the model to emit a single token from a small candidate set, we obtain a controlled surrogate classification signal that reflects the model's image-conditioned preference structure. We then apply query-based zero-order optimization with a low-frequency prior to perturb the image in a way that reduces alignment between the visual input and the model's original interpretation, operationalized as a flip away from the correct choice on MedXpertQA (Zuo et al., 2025).

Together, these findings indicate that restricting access to model parameters is insufficient to ensure robustness in clinical deployments and motivate defenses that account for exposed inference signals and the unique adversarial surface introduced by multimodal medical inputs.

2. Methods

Data. We evaluate on the multimodal test split of MedXpertQA. We restrict to questions with exactly one associated image (pathology, radiology, ECG, photos, drawings, etc.), yielding 1,581 examples. In addition, we filter images containing more than 1,500,000 pixels to accommodate a restricted VRAM budget. The total number of examples after filtering is 1,200.

Threat model. We consider a black-box (gradient-free) attacker who can query the model on chosen images under a fixed prompt and observe next-token scores over a restricted answer set[1]. The goal is to induce a change in the model's image-conditioned preference over a restricted answer set; in our probe setting, this manifests as flipping a *single-token* multiple-choice decision.

API interface and image encoding. To simulate realistic deployment conditions, we serve models via an OpenAI-compatible API using vLLM (Kwon et al., 2023), mimicking how production vision-language models are typically accessed. Queries are made through standard HTTP requests with images encoded as base64 JPEG (quality=85). This compression is applied after perturbation generation—while adversarial optimization operates on uncompressed tensors, the final attack images are evaluated through the compressed API pathway. This design choice reflects that: (1) many deployed systems apply compression to reduce bandwidth and storage costs, and (2) compression can serve as a partial defense against adversarial attacks. Our setting thus represents a conservative estimate of attack success under practical constraints.

Transfer to free-form interfaces. While restricted decoding is our main robustness probe, we also test whether the resulting perturbations transfer to a more deployment-like, logprob-free interface by re-querying original/adversarial images with (i) a *letter-only* prompt (greedy; `max_tokens` $= 1$) and (ii) a self-consistency variant based on majority vote over sampled generations.

2.1. Restricted decoding as a surrogate classification probe

For each question, the model must output one choice from $\mathcal{A} = \{A, B, C, D, \dots\}$. Let $z(\mathbf{x}) \in \mathbb{R}^V$ denote next-token logits given image $\mathbf{x}$ and the fixed prompt. Each answer $a \in \mathcal{A}$ maps to a single tokenizer id $\tau(a)$ (checked with/without a leading space). Prediction under restricted decoding is

$$\hat{a}(\mathbf{x}) = \arg\max_{a \in \mathcal{A}} z_{\tau(a)}(\mathbf{x}).$$

Let $a^\star$ be the ground-truth answer and $t^\star = \tau(a^\star)$. We perform *untargeted* attacks that reduce preference for $a^\star$.

[1]Many inference APIs expose log-probabilities (e.g., for calibration or structured decoding). Our implementation uses top-k logprobs over the candidate set; this preserves relative preferences over candidates, which is what the attack exploits.

We report attack success rate (ASR) conditional on initially correct predictions.

We emphasize that restricted decoding is not a deployment assumption but an experimental probe. Constraining the output to a single token in a fixed candidate set turns the image-conditioned generation process into a well-defined surrogate classification problem with an explicit decision boundary between candidate tokens. This provides a controlled measurement of how small perturbations to the image move the model's internal multimodal state across a decision surface, enabling a precise robustness evaluation in a black-box setting.

Score function. Let $\mathcal{T}$ denote the set of candidate answer tokens (e.g., option letters), and let $\ell_t(\mathbf{x})$ be the model log-probability of emitting token $t \in \mathcal{T}$ for image $\mathbf{x}$ under the fixed prompt. These log-probabilities are obtained directly from the API response, without access to full logits. For a reference choice $t^\star$ (the ground-truth option in the untargeted setting), we define

$$s(\mathbf{x}) = \Big(\ell_{t^\star}(\mathbf{x}) - \max_{t \in \mathcal{T} \setminus \{t^\star\}} \ell_t(\mathbf{x})\Big) + \lambda\, \ell_{t^\star}(\mathbf{x}), \quad (1)$$

with $\lambda = 0.5$ throughout. The first term corresponds to the log-probability margin between the reference option and its strongest competitor, while the second term penalizes absolute confidence in $t^\star$. Minimizing $s(\mathbf{x})$ via a zeroth-order NES estimator and projected ℓ_∞ pixel updates decreases both the relative and absolute preference for $t^\star$, thereby inducing a switch to alternative options under bounded perturbations.

2.2. Zeroth-order optimization with a low-frequency prior

We optimize $s(\mathbf{x})$ using query-based zeroth-order gradient estimates and signed updates as introduced by Liu et al. (2019) with ZO-SignSGD. At iteration t, we draw m random directions in a reduced-resolution Gaussian space, upsample them to image resolution (bilinear), and use antithetic NES queries:

$$\hat{g} = \frac{1}{2\sigma m} \sum_{i=1}^{m} \big(s(\mathbf{x} + \sigma u_i) - s(\mathbf{x} - \sigma u_i)\big) u_i,$$

where σ is the smoothing scale and u_i are the low-frequency directions. This reduces variance and improves query efficiency compared to pixel-wise search.

2.3. Update rule and constraints

All optimization is carried out directly in raw pixel space. Let $\mathbf{x}_0$ denote the original image. We enforce an ℓ_∞ constraint

$$\|\mathbf{x} - \mathbf{x}_0\|_\infty \le \epsilon,$$

and after each update project back onto this ℓ_∞ ball and clamp pixel values to the valid input range. The iterative update rule is

$$\mathbf{x} \leftarrow \mathbf{x} - \alpha_t \operatorname{sign}(\tilde{g}_t),$$

where $\tilde{g}_t$ is a normalized or preconditioned variant of the zeroth-order gradient estimate $\hat{g}_t$ (using either Adam-style first and second moments or simple mean normalization) (Ilyas et al., 2018). The step size α_t follows a cosine schedule over T optimization steps. For grayscale-constrained attacks, the update direction is restricted to a luminance channel and replicated across the RGB channels.

Each iteration requires $2m$ model queries due to antithetic sampling, resulting in a worst-case budget of $2mT$ queries per example, with early stopping applied when the restricted-decoding prediction changes. In all experiments, we set $m = 128$ and $T = 30$.

2.4. Human perceptibility evaluation

To assess detectability by a medical expert, we conducted a paired comparison with one physician on 150 randomized original/adversarial pairs at $\epsilon = 0.02$, stratified by model (`MedGemma 4B`/`27B`) and grayscale vs. color. The participant indicated which image appeared digitally altered or reported being unable to distinguish them; we report detection accuracy and modality-specific outcomes.

3. Results

We evaluate the effectiveness of zero-order adversarial attacks on medical vision-language models under black-box access constraints. Our experiments focus on `MedGemma 4B`, `MedGemma 27B`, and `Qwen3-VL-Instruct 4B` models across 1,200 single-image multimodal questions from MedXpertQA, testing two perturbation budgets ($\varepsilon \in \{0.01, 0.02\}$) and two image settings (grayscale and color) (Bai et al., 2025; Sellergren et al., 2025).

3.1. Attack success rates and perturbation magnitudes

Table 1 summarizes aggregate attack success rates (ASR) and perturbation statistics across models, perturbation budgets, and input representations. For `MedGemma 4B` at $\varepsilon = 0.01$, the ZO-SignSGD attack achieves an ASR of 59.8% on grayscale images and 68.3% on color images. At $\varepsilon = 0.02$, ASR remains comparable at 56.7% for grayscale and 71.6% for color. Across both budgets, attacks on color images consistently outperform their grayscale counterparts, reflecting the reduced degrees of freedom imposed by the grayscale projection constraint, which restricts perturbations to a single luminance channel replicated across RGB channels.

A similar trend is observed for `MedGemma 27B`. At $\varepsilon =$

Table 1. Attack success rates (ASR, %) comparing zero-order optimized attacks with uninformed noise baselines and no image provided in the query. Optimized attacks substantially outperform random, high-frequency noise, and the absence of input image, confirming that success stems from adversarial optimization rather than model brittleness.

Model	ε	Method	Type	ASR	L_∞ (px)
MedGemma 4B	-	No image	-	34.8%	-
	0.01	**ZO-SignSGD + Adam**	**Grayscale**	**59.8%**	0.0066±0.0073
		ZO-SignSGD + Adam	**Color**	**68.3%**	0.0059±0.0073
		Random	Uniform	12.5%	0.01±0.0000
		High-Frequency	Checkerboard	18.1%	0.01±0.0000
	0.02	**ZO-SignSGD + Adam**	**Grayscale**	**56.7%**	0.0126±0.0134
		ZO-SignSGD + Adam	**Color**	**71.6%**	0.0109±0.0173
		Random	Uniform	12.3%	0.02±0.0000
		High-Frequency	Checkerboard	16.3%	0.02±0.0000
MedGemma 27B	-	No image	-	9.4%	-
	0.01	**ZO-SignSGD + Adam**	**Grayscale**	**22.3%**	0.0078±0.0047
		ZO-SignSGD + Adam	**Color**	**39.5%**	0.0073±0.0060
		Random	Uniform	4.0%	0.01±0.0000
		High-Frequency	Checkerboard	6.3%	0.01±0.0000
	0.02	**ZO-SignSGD + Adam**	**Grayscale**	**25.0%**	0.0145±0.0101
		ZO-SignSGD + Adam	**Color**	**40.5%**	0.0134±0.0147
		Random	Uniform	4.3%	0.02±0.0000
		High-Frequency	Checkerboard	4.3%	0.02±0.0000
Qwen3-VL-Instruct 4B	-	No image	-	30.7%	-
	0.01	**ZO-SignSGD + Adam**	**Grayscale**	**49.4%**	0.0065±0.0073
		ZO-SignSGD + Adam	**Color**	**58.1%**	0.0061±0.0073
		Random	Uniform	10.6%	0.01±0.0000
		High-Frequency	Checkerboard	10.2%	0.01±0.0000
	0.02	**ZO-SignSGD + Adam**	**Grayscale**	**48.2%**	0.0118±0.0173
		ZO-SignSGD + Adam	**Color**	**58.9%**	0.0109±0.0173
		Random	Uniform	11.7%	0.02±0.0000
		High-Frequency	Checkerboard	12.1%	0.02±0.0000

0.01, the attack achieves an ASR of 22.3% on grayscale images and 39.5% on color images, increasing to 25.0% and 40.5%, respectively, at $\varepsilon = 0.02$ (Table 1). Across perturbation budgets and input representations, `MedGemma 27B` consistently exhibits lower attack success rates than its `4B` counterpart, suggesting increased resistance to black-box, zero-order attacks, potentially due to reduced reliance on visual inputs compared to text in VLMs (Zhao et al., 2025).

For `Qwen3-VL-Instruct 4B`, the attack achieves 49.4% ASR on grayscale and 58.1% on color images at $\varepsilon = 0.01$, with comparable rates of 48.2% and 58.9% at $\varepsilon = 0.02$. This model shows intermediate vulnerability between the two `MedGemma` variants, with success rates closer to the smaller `MedGemma 4B` model.

Across all settings, optimized NES attacks substantially outperform uninformed noise baselines, including both uniform random perturbations and high-frequency checkerboard noise. For example, `MedGemma 4B` at $\varepsilon = 0.02$ achieves 71.6% ASR with optimization versus only 12.3% for random noise and 16.3% for checkerboard patterns. `MedGemma 27B` shows similar gaps: 40.5% ASR versus 4.3% for both baselines. These results confirm that the observed attack success is driven by adversarial optimization rather than incidental sensitivity to arbitrary perturbations.

3.2. Transfer to free-form interfaces

Restricted decoding provides a controlled robustness probe, but real deployments often use free-form generation without exposing log-probabilities. To test whether perturbations found using our restricted-decoding objective transfer to a more deployment-like interface, we re-query the same original/adversarial images using a *letter-only* free-form prompt (greedy decoding), and additionally measure a self-consistency variant where we sample $N = 20$ completions (temperature 0.7, top-p 0.9) and take a majority vote.

Table 2 shows that restricted-decoding flips partially transfer to this letter-only interface and lead to meaningful accuracy degradation. For both `MedGemma 4B` and `MedGemma 27B` (color, $\varepsilon = 0.02$), letter-only accuracy drops by 8–11 percentage points, and agreement between restricted de-

Table 2. **Probe-to-behavior transfer under free-form querying** (color, $\varepsilon = 0.02$). We evaluate original vs. adversarial images using (i) restricted decoding (logprob-based probe) and (ii) a logprob-free letter-only interface (greedy, one token), plus an optional self-consistency majority vote over $N = 20$ samples. Accuracy is computed w.r.t. ground-truth labels under the letter-only interface.

Model	n	Accuracy (orig)	Accuracy (adv)	Agreement drop
`MedGemma 4B`	271	43.5%	32.8%	45.9% → 33.6%
`MedGemma 27B`	299	91.6%	83.3%	94.5% → 74.7%

coding and letter-only outputs decreases substantially post-attack. In this limited two-model evaluation, the results suggest that free-form decoding can mask some restricted-decoding instability without fully eliminating its behavioral effects. However, because this transfer analysis covers only two models and a letter-only interface, we treat it as preliminary evidence rather than a general conclusion about open-ended medical generation.

3.3. Optimization convergence and time-to-success

Figures 1 and 2 show the convergence behavior across optimization steps for all models under both perturbation budgets. The curves plot both per-step success rate (proportion of examples that flip at exactly that step) and cumulative ASR (proportion that have flipped by that step). Across models, most successful attacks occur within the first 15–20 steps, with diminishing returns thereafter. The gap between grayscale and color ASR is evident throughout optimization, consistent with the constraint-imposed limitation on effective perturbation dimensionality.

We analyze time-to-success by measuring the first optimization step at which the restricted-decoding prediction flips away from the correct answer. For `MedGemma 4B` at $\varepsilon = 0.01$, successful attacks on color images occur after a mean of 6.30 steps (median 4, $\sigma = 5.67$; $n = 185$ successes), versus 8.62 steps on grayscale (median 5, $\sigma = 7.95$; $n = 58$). At $\varepsilon = 0.02$, time-to-success remains in the same range with a mean of 6.71 steps (median 3.5, $\sigma = 7.63$; $n = 194$) for color and 7.13 steps (median 5, $\sigma = 6.52$; $n = 55$) for grayscale.

For `Qwen3-VL-Instruct 4B`, successful attacks at $\varepsilon = 0.01$ occur after 7.26 steps on color (median 4, $\sigma = 7.14$; $n = 154$) and 8.90 steps on grayscale (median 5, $\sigma = 7.96$; $n = 41$). At $\varepsilon = 0.02$, the corresponding means are 7.35 steps on color (median 3, $\sigma = 8.39$; $n = 156$) and 9.60 steps on grayscale (median 5.5, $\sigma = 10.23$; $n = 40$).

For `MedGemma 27B`, successful attacks require more iterations on average. At $\varepsilon = 0.01$, the mean first-success step is 9.84 on color (median 7, $\sigma = 8.22$; $n = 118$) and 10.80 on grayscale (median 9, $\sigma = 7.83$; $n = 25$). At $\varepsilon = 0.02$, these increase slightly to 10.07 on color (median 6, $\sigma = 9.29$; $n = 121$) and 10.89 on grayscale (median 8.5, $\sigma = 8.80$; $n = 28$).

Overall, these statistics show that successful attacks require on the order of 1,000 queries on average, which remains feasible even under strict rate limits.

3.4. Stratification by medical domain and task type

Attack effectiveness varies across medical specialties, imaging modalities, and question types, and these trends are largely consistent across model scale. Tables 4 to 6 in the appendix provide detailed stratification by body system, medical task, and question type for `MedGemma 4B`, `Qwen3-VL-Instruct 4B`, and `MedGemma 27B`, respectively.

Body system variation. Across all models, ASR varies substantially by body system, with color images consistently more vulnerable than grayscale. For `MedGemma 4B` at $\varepsilon = 0.02$ on color images, ASR ranges from 57.1% (Integumentary; 16/28) to 85.2% (Cardiovascular; 46/54), with particularly high vulnerability in Nervous (79.5%; 31/39), Lymphatic (77.8%; 7/9), and Endocrine (77.8%; 7/9) systems.

For `MedGemma 27B` at $\varepsilon = 0.02$, the same pattern emerges with lower overall rates: Lymphatic shows highest vulnerability at 71.4% (5/7), followed by Cardiovascular at 50.9% (28/55) and Nervous at 48.7% (19/39). The lowest ASR is observed in Integumentary (30.2%; 13/43) and Digestive (30.4%; 7/23) systems.

`Qwen3-VL-Instruct 4B` at $\varepsilon = 0.02$ shows highest vulnerability in Lymphatic (75.0%; 6/8) and Integumentary (72.7%; 24/33), with lower rates in Digestive (43.8%; 7/16) and Nervous (54.5%; 18/33) systems.

Medical task dependency. Task-level stratification reveals consistent patterns across models. For `MedGemma 4B` at $\varepsilon = 0.02$ on color images, ASR reaches 74.5% (41/55) for *Basic Science*, 73.5% (111/151) for *Diagnosis*, and 64.6% (42/65) for *Treatment*. `MedGemma 27B` shows lower but similarly ordered rates: 52.2% (24/46) for *Basic Science*, 42.4% (73/172) for *Diagnosis*, and 29.6% (24/81) for *Treatment*. `Qwen3-VL-Instruct 4B` exhibits the highest vulnerability on *Basic Science* tasks at 78.0% (39/50), followed by *Diagnosis* at 54.9% (79/144) and *Treatment* at 53.5% (38/71).

Reasoning vs. understanding. Question type analysis shows mixed patterns. For `MedGemma 4B` at $\varepsilon = 0.02$ on color images, *Reasoning* questions achieve 72.1% ASR (158/219) while *Understanding* questions reach 69.2% (36/52). `MedGemma 27B` shows 39.9% ASR (97/243) for *Reasoning* and 42.9% (24/56) for *Understanding*.

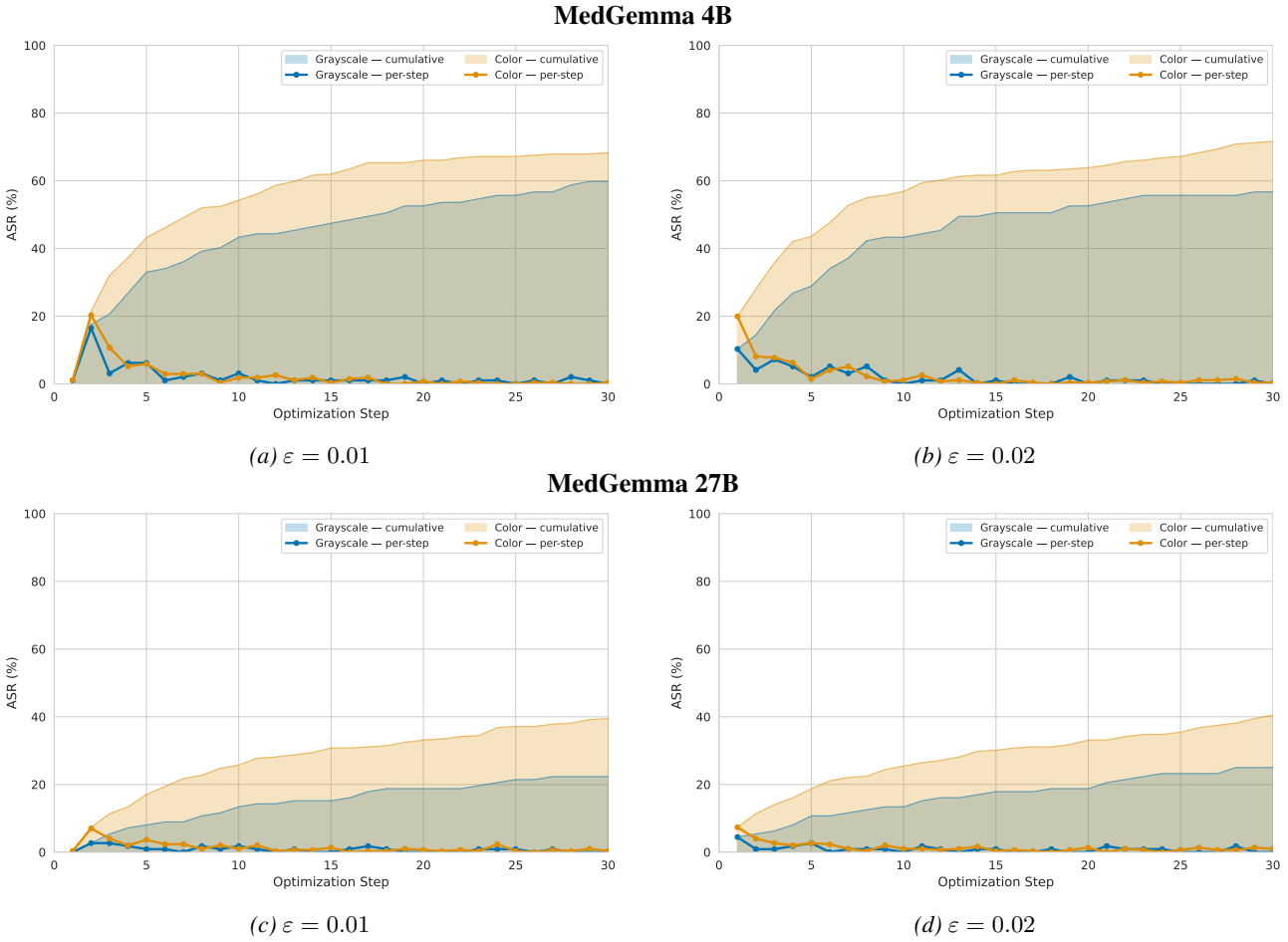


Figure 1. Attack success for medical VLMs showing per-step and cumulative attack success rates (ASR) across optimization steps. Each row corresponds to a model, and columns compare perturbation budgets $\varepsilon = 0.01$ and $\varepsilon = 0.02$. Per-step rates indicate the proportion of examples successfully attacked at each iteration, while cumulative rates show the total proportion attacked up to that step.

Table 3. Human perceptibility study results

Category	Overall Accuracy	Cannot Tell Rate
Overall	57.0%	72.7%
`MedGemma 27B`	58.5%	66.0%
`MedGemma 4B`	54.6%	79.6%
`Qwen3-VL-Instruct 4B`	58.2%	71.4%
Color	53.6%	80.0%
Grayscale	66.2%	52.5%

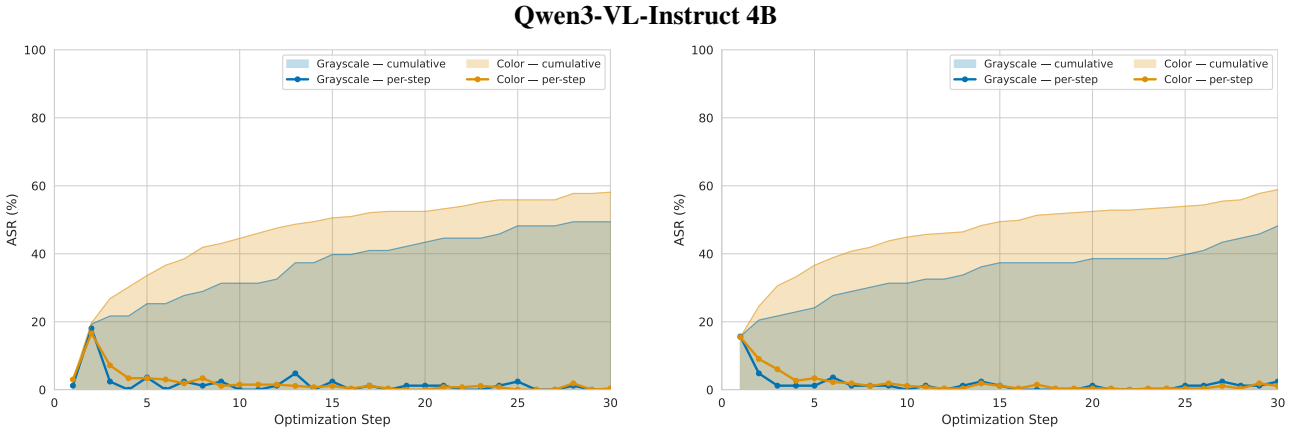
`Qwen3-VL-Instruct 4B` demonstrates a larger gap, with *Understanding* questions substantially more vulnerable at 75.0% (39/52) compared to *Reasoning* at 54.9% (117/213).

3.5. Qualitative analysis of adversarial perturbations

Figure 3 presents representative examples of adversarial images, amplified perturbations, and original images across different medical imaging modalities. The low-frequency nature of the NES-optimized perturbations is visible in the amplified difference maps, which show smooth, spatially correlated noise patterns rather than high-frequency salt-and-pepper artifacts.

Table 3 summarizes a small paired perceptibility study with one physician over 150 randomized original/adversarial pairs at $\epsilon = 0.02$. Overall, detection accuracy is only slightly above chance (57.0%), and the expert reported being unable to distinguish the pair in most trials (72.7% "cannot tell" rate), suggesting that these low-frequency perturbations can be visually subtle in this small expert assessment. Because the evaluation involved a single physician, these results should not be interpreted as evidence that the perturbations are clinically believable across readers, specialties, or deployment settings.

Perceptibility differs sharply between color and grayscale settings. For color images, detection accuracy drops to 53.6% with an 80.0% "cannot tell" rate, consistent with successful perturbations blending into the natural texture



(a) $\varepsilon = 0.01$ *(b)* $\varepsilon = 0.02$

Figure 2. Attack success for non-medical VLMs showing per-step and cumulative attack success rates (ASR) across optimization steps. Each row corresponds to a model, and columns compare perturbation budgets $\varepsilon = 0.01$ and $\varepsilon = 0.02$. Per-step rates indicate the proportion of examples successfully attacked at each iteration, while cumulative rates show the total proportion attacked up to that step.

and staining variability of pathology slides and clinical photographs. Grayscale images are more readily distinguished (66.2% accuracy) and yield a lower "cannot tell" rate (52.5%), suggesting that constraining updates to a shared luminance channel can produce more globally coherent intensity changes that are easier to notice, even though it reduces attack effectiveness (Section 3.1).

Across models, perceptibility is broadly similar (accuracy 54.6–58.5%; "cannot tell" 66.0–79.6%), suggesting that the visual characteristics of the perturbations are primarily determined by the optimization prior and the ℓ_∞ budget rather than the specific target model.

3.6. Comparison with baseline conditions

To validate that attack success stems from adversarial optimization rather than model brittleness, we compare NES-optimized attacks against two uninformed baselines: uniform random noise and high-frequency checkerboard patterns, both at the same perturbation budget ε. Table 1 shows that optimized attacks substantially outperform both baselines. For `MedGemma 4B` at $\varepsilon = 0.02$ on color images, NES achieves 71.6% ASR versus 12.3% for random noise and 16.3% for checkerboard. `MedGemma 27B` shows similar gaps: 40.5% ASR versus 4.3% for both baselines. For `Qwen3-VL-Instruct 4B`, optimized attacks reach 58.9% ASR compared to 11.7% for random noise and 12.1% for checkerboard patterns. Grayscale attacks follow the same pattern, with optimized perturbations achieving substantially higher success rates across all models.

These results show that success is not explained by generic noise sensitivity: both uninformed baselines saturate the ℓ_∞ budget yet yield much lower ASR. Instead, NES finds structured, low-frequency perturbations that move examples along vulnerable directions of the decision boundary. More broadly, because these models can often perform well even under weakened or missing visual inputs (i.e., strong textual bias) (Zhao et al., 2025), our perturbations do more than remove visual signal—they inject an additional, adversarial signal that overrides these priors.

4. Discussion

This study evaluates whether a medical vision-language model can be induced to change its image-conditioned preference under restricted, single-token multiple-choice decoding using only black-box query access to next-token logits and zero-order optimization. Across 1,200 single-image MedXpertQA examples (after filtering for image size), we observe non-trivial attack success rates across models under tight ℓ_∞ constraints (Table 1). In particular, attacks are consistently more effective in the color setting than under the grayscale (luminance-projected) constraint.

A consistent pattern is that attacks are more effective on color than on grayscale inputs at the same ε. This is expected under our constraint, since the optional grayscale projection reduces the effective degrees of freedom of the perturbation, and thus limits the ability of low-frequency directions to reshape the vision encoder representation.

The main practical implication supported by these experiments is that gradient access is not required to induce answer flips in this restricted-decoding setting, provided that an attacker can repeatedly query the model and observe next-token logits (or equivalently log-probabilities over a fixed candidate set). Importantly, this is a relatively accessible threat model: the attacker needs only API access and the ability to submit many slightly modified images, without

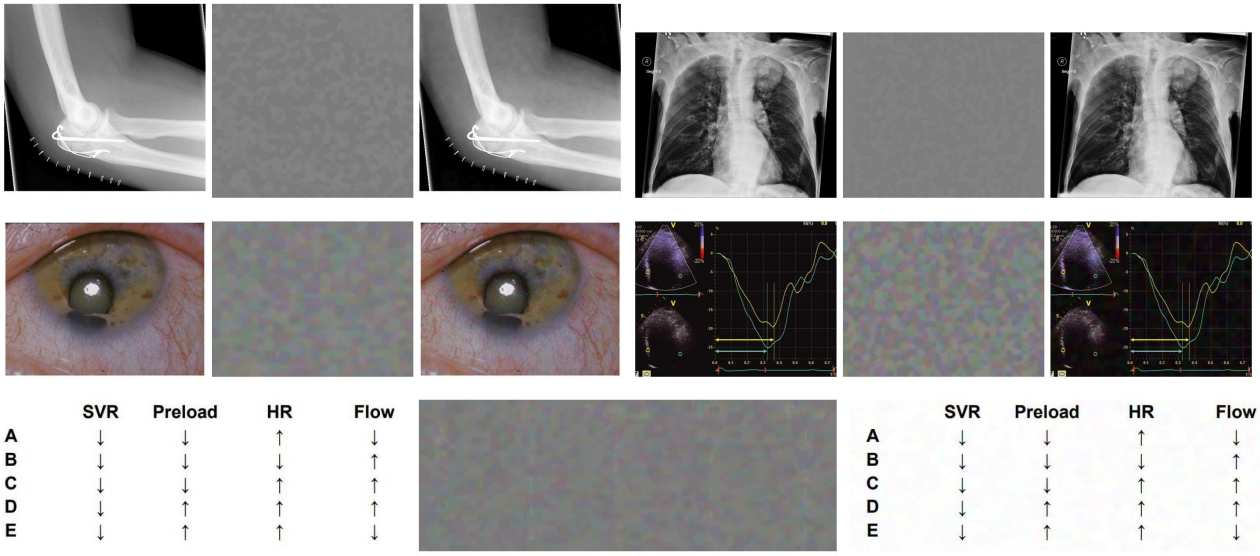


Figure 3. **Representative adversarial examples across medical imaging modalities.** For each sample, we display three images from left to right, the original image, the adversarial noise amplified by $4\times$ for visualization, and the resulting adversarial image. All generated with $\varepsilon = 0.02$, 30 steps, 128 samples/step. The first two rows are using grayscale noise and the last two rows are using color noise.

any privileged system integration.

In our configuration, each optimization step uses $2m$ queries, so the worst-case query count is $2mT = 7{,}680$ forward passes per example (before early stopping), which is substantially higher than typical white-box attacks. However, the cost–benefit profile remains favorable for a motivated adversary: (i) we observe that most successful attacks flip within 15–20 steps (Section 3.3), and (ii) simple heuristic perturbations (random noise and checkerboard patterns at the same ε budget) yield much lower success rates (Table 1), implying that optimization is needed but does not require specialized resources beyond repeated querying.

This also highlights a realistic detection and mitigation surface: repeated near-duplicate image queries from a single client are amenable to rate limiting and anomaly detection. Nevertheless, such controls do not eliminate the risk entirely (e.g., distributed querying across accounts/IPs, long-horizon attacks spread over time, or on-prem deployments without strict ingress controls), and so should be viewed as friction rather than a guarantee.

More broadly, exposing token-level scores provides a particularly strong optimization signal for black-box attacks, as also suggested by prior evaluations of black-box VLM robustness (Zhao et al., 2023). A simple and actionable defense implication is therefore to minimize exposed inference signals: masking log-probabilities, or returning only sampled text where possible reduces the direct gradient-free optimization channel. These changes may not be sufficient by themselves considering LLMs are capable of expressing their uncertainty (Xiong et al., 2024). Additional system-level mitigations—such as strengthening the image encoding pipeline or introducing stochastic image preprocessing/encoders—may further increase query complexity, though they can also affect benign performance and require careful evaluation in medical settings where image artefacts already impact model reliability (Cheng et al., 2025).

Several limitations follow from the probe setting used here. First, we evaluate restricted decoding where the model emits a single token from a small candidate set. We use this as a controlled measurement operator that converts multimodal generation into a surrogate classification signal with an explicit decision boundary. While this isolates the effect of image perturbations on the model's immediate image-conditioned preferences, it does not fully capture longer-horizon dynamics in open-ended generation (e.g., multi-token rationales, tool use, or downstream clinical pipelines). Extending query-based attacks and robustness evaluation to such settings remains an important direction for future work. Second, the human perceptibility evaluation is limited to one physician and 150 paired comparisons. It provides an initial check of visual subtlety, but it cannot establish clinical believability or diagnostic innocuousness. Larger reader studies across specialties and imaging modalities would be needed before making stronger claims about clinical plausibility.

Impact Statement

This work examines the security properties of medical vision-language models under a specific black-box threat model and is intended to improve understanding of potential failure modes in high-stakes clinical decision support systems. By demonstrating that image-based adversarial perturbations can influence model outputs without access to model weights or gradients, this study highlights risks that may arise if such systems are deployed without appropriate safeguards.

The results do not suggest that current clinical practice is immediately compromised, as real-world deployments typically involve human oversight, additional preprocessing, and access controls that may limit the feasibility of the attacks studied here. However, the findings underscore the importance of carefully defining threat models, limiting exposed inference signals, and incorporating monitoring or human-in-the-loop mechanisms when deploying multimodal AI systems in clinical contexts.

The attack should also be interpreted in light of its interface assumption: it is most directly applicable to systems that expose token-level scores or log-probabilities. Systems that suppress these signals, rate-limit repeated near-duplicate queries, or introduce robust preprocessing may substantially reduce practical feasibility.

More broadly, this work aims to inform the development of more robust and trustworthy medical AI systems by encouraging systematic security evaluation prior to deployment. The insights presented here are intended to support safer model design and deployment practices rather than enable malicious use.

References

Alber, D. A., Yang, Z., Alyakin, A., Yang, E., Rai, S., Valliani, A. A., Zhang, J., Rosenbaum, G. R., Amend-Thomas, A. K., Kurland, D. B., Kremer, C. M., Eremiev, A., Negash, B., Wiggan, D. D., Nakatsuka, M. A., Sangwon, K. L., Neifert, S. N., Khan, H. A., Save, A. V., Palla, A., Grin, E. A., Hedman, M., Nasir-Moin, M., Liu, X. C., Jiang, L. Y., Mankowski, M. A., Segev, D. L., Aphinyanaphongs, Y., Riina, H. A., Golfinos, J. G., Orringer, D. A., Kondziolka, D., and Oermann, E. K. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine*, 31 (2):618–626, February 2025. ISSN 1546-170X. doi: 10.1038/s41591-024-03445-1. URL https://doi.org/10.1038/s41591-024-03445-1.

Anthropic. Advancing Claude in healthcare and the life sciences, January 2026. URL https://www.anthropic.com/news/healthcare-life-sciences.

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., and Zhu, K. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Cheng, Z., Ong, A. Y., Wagner, S. K., Merle, D. A., Ju, L., Zhang, H., Chen, R., Pang, L., Li, B., He, T., Ran, A., Jiang, H., Yang, D. G., Zou, K., Goh, J. H. L., Srinivasan, S., Altmann, A., Alexander, D. C., Cheung, C. Y., Tham, Y. C., Keane, P. A., and Zhou, Y. Understanding the robustness of vision-language models to medical image artefacts. *npj Digital Medicine*, 8(1): 727, November 2025. ISSN 2398-6352. doi: 10.1038/s41746-025-02108-w. URL https://www.nature.com/articles/s41746-025-02108-w. Publisher: Nature Publishing Group.

Datta, S., Buchireddygari, D., Kaza, L. V. C., Bhalke, M., Singh, K., Pandey, A., Vasipalli, S. S., Karnwal, U., Bhatti, H. B. S., Maroo, B. R., Hebbar, S., Joseph, R., Kaur, G., Singh, D., V, A., Prasad, D. D. S., Mahajan, N., Arisha, A., Vanagundi, R., Nandy, R., Vuthoo, K., Rajvanshi, S., Kondaveeti, N., Gunjal, S., Jain, R., Jain, R., and Agrawal, A. Radiology's Last Exam (RadLE): Benchmarking Frontier Multimodal AI Against Human Experts and a Taxonomy of Visual Reasoning Errors in Radiology, September 2025a. URL http://arxiv.org/abs/2509.25559. arXiv:2509.25559 [cs].

Datta, S., Buchireddygari, D., Kaza, L. V. C., Karnwal, U., Bhatti, H. B. S., and Singh, K. Gemini 3.0 Pro Surpasses Radiology Trainees on RadLE v1, November 2025b. URL https://drdattaaiims.github.io/Gemini-3.0-Radiology-2025.html.

Epic. Cool Stuff Now: Epic and Generative AI | Epic, 2024. URL https://www.epic.com/epic/post/cool-stuff-now-epic-and-generative-ai/.

Gallifant, J., Chen, S., Moreira, P., Munch, N., Gao, M., Pond, J., Celi, L. A., Aerts, H., Hartvigsen, T., and Bitterman, D. Language Models are Surprisingly Fragile to Drug Names in Biomedical Benchmarks, June 2024. URL http://arxiv.org/abs/2406.12066.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and Harnessing Adversarial Examples. March 2015. doi: 10.

48550/arXiv.1412.6572. URL http://arxiv.org/abs/1412.6572. arXiv:1412.6572 [stat].

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., and Fritz, M. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. *Proceedings of the 16th ACM workshop on artificial intelligence and security*, pp. 79–90, 2023. doi: 10.1145/3605764.3623985. URL https://doi.org/10.1145/3605764.3623985.

Griot, M., Hemptinne, C., Vanderdonckt, J., and Yuksel, D. Large Language Models lack essential metacognition for reliable medical reasoning. *Nature Communications*, 16 (1):642, January 2025a. ISSN 2041-1723. doi: 10.1038/s41467-024-55628-6. URL https://doi.org/10.1038/s41467-024-55628-6.

Griot, M., Vanderdonckt, J., and Yuksel, D. Implementation of large language models in electronic health records. *PLOS Digital Health*, 4(12):1–18, December 2025b. doi: 10.1371/journal.pdig.0001141. URL https://doi.org/10.1371/journal.pdig.0001141. Publisher: Public Library of Science.

Griot, M., Vanderdonckt, J., Yuksel, D., and Hemptinne, C. Pattern Recognition or Medical Knowledge? The Problem with Multiple-Choice Questions in Medicine. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5321–5341, Vienna, Austria, July 2025c. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.266. URL https://aclanthology.org/2025.acl-long.266/.

Gu, Y., Fu, J., Liu, X., Valanarasu, J. M. J., Codella, N. C., Tan, R., Liu, Q., Jin, Y., Zhang, S., Wang, J., Wang, R., Song, L., Qin, G., Usuyama, N., Wong, C., Cheng, H., Lee, H., Sanapathi, P., Hilado, S., Bian, J., Alvarez-Valle, J., Wei, M., Malik, K., Gao, J., Horvitz, E., Lungren, M. P., Poon, H., and Vozila, P. The illusion of readiness: Stress testing large frontier models on multimodal medical benchmarks, 2025. URL https://arxiv.org/abs/2509.18234. arXiv:2509.18234 [cs.AI].

Huff, D. T., Weisman, A. J., and Jeraj, R. Interpretation and visualization techniques for deep learning models in medical imaging. *Physics in Medicine & Biology*, 66(4):04TR01, February 2021. doi: 10.1088/1361-6560/abcd17. URL https://doi.org/10.1088/1361-6560/abcd17. Publisher: IOP Publishing.

Ilyas, A., Engstrom, L., Athalye, A., and Lin, J. Black-box adversarial attacks with limited queries and information. In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th international conference on machine learning, ICML 2018, stockholmsmässan, stockholm, sweden, july 10-15, 2018*, volume 80 of *Proceedings of machine learning research*, pp. 2142–2151. PMLR, 2018. URL http://proceedings.mlr.press/v80/ilyas18a.html. tex.bibsource: dblp computer science bibliography, https://dblp.org tex.timestamp: Wed, 03 Apr 2019 18:17:30 +0200.

Jia, X., Gao, S., Qin, S., Pang, T., Du, C., Huang, Y., Li, X., Li, Y., Li, B., and Liu, Y. Adversarial attacks against closed-source MLLMs via feature optimal alignment. In *The thirty-ninth annual conference on neural information processing systems*, 2025. URL https://openreview.net/forum?id=ktC3cDu320.

Jin, Q., Chen, F., Zhou, Y., Xu, Z., Cheung, J. M., Chen, R., Summers, R. M., Rousseau, J. F., Ni, P., Landsman, M. J., Baxter, S. L., Al'Aref, S. J., Li, Y., Chen, A., Brejt, J. A., Chiang, M. F., Peng, Y., and Lu, Z. Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. *npj Digital Medicine*, 7 (1):1–6, July 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01185-7. URL https://www.nature.com/articles/s41746-024-01185-7.

Kim, H., Ai, X., Srinivasan, S., Gilson, A., Singer, M. B., Pushpanathan, K., Xie, Q., Park, J., Applebaum, S., Yang, G. D., Zou, M., Chen, D. Z., Zou, K., Sarrafpour, S., Liu, J., Yin, Y., Huang, J., Nguyen, Q. N., Long, E., Wan, P., Liu, D., Hintz, R., Zheng, W. J., Wang, S. Y., Ohno-Machado, L., Xu, H., Adelman, R. A., Priore, L. V. D., Tham, Y.-C., and Chen, Q. LEME: Open Large Language Models for Ophthalmology with Advanced Reasoning and Clinical Validation, November 2025. URL http://arxiv.org/abs/2410.03740. arXiv:2410.03740 [cs].

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J., Zhang, H., and Stoica, I. Efficient Memory Management for Large Language Model Serving with PagedAttention. *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, October 2023. doi: 10.1145/3600006.3613165. URL https://dl.acm.org/doi/10.1145/3600006.3613165.

Lawrence, R., Dodsworth, E., Massou, E., Sherlaw-Johnson, C., Ramsay, A. I., Walton, H., O'Regan, T., Gleeson, F., Crellin, N., Herbert, K., Ng, P. L., Elphinstone, H., Mehta, R., Lloyd, J., Halliday, A., Morris, S., and Fulop, N. J. Artificial intelligence for diagnostics in radiology practice: a rapid systematic scoping review. *eClinicalMedicine*,

83, May 2025. ISSN 2589-5370. doi: 10.1016/j.eclinm.2025.103228. URL https://doi.org/10.1016/j.eclinm.2025.103228. Publisher: Elsevier.

Lee, R. W., Jun, T. J., Lee, J.-M., Cho, S. I., Park, H. J., and Suh, J. Vulnerability of Large Language Models to Prompt Injection When Providing Medical Advice. *JAMA Network Open*, 8(12):e2549963, December 2025. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2025.49963. URL https://doi.org/10.1001/jamanetworkopen.2025.49963.

Li, J., Aguirre, A. D., Junior, V. M., Jin, J., Liu, C., Zhong, L., Sun, C., Clifford, G., Brandon Westover, M., and Hong, S. An Electrocardiogram Foundation Model Built on over 10 Million Recordings. *NEJM AI*, 2(7):AIoa2401033, June 2025. doi: 10.1056/AIoa2401033. URL https://ai.nejm.org/doi/full/10.1056/AIoa2401033. Publisher: Massachusetts Medical Society.

Liu, S., Chen, P.-Y., Chen, X., and Hong, M. signSGD via Zeroth-Order Oracle. In *International conference on learning representations*, 2019. URL https://openreview.net/forum?id=BJe-DsC5Fm.

Ma, D., Pang, J., Gotway, M. B., and Liang, J. A fully open AI foundation model applied to chest radiography. *Nature*, 643(8071):488–498, July 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09079-8. URL https://www.nature.com/articles/s41586-025-09079-8. Publisher: Nature Publishing Group.

OpenAI. Introducing OpenAI for Healthcare, January 2026. URL https://openai.com/index/openai-for-healthcare/.

Riedemann, L., Labonne, M., and Gilbert, S. The path forward for large language models in medicine is open. *npj Digital Medicine*, 7(1):1–5, November 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01344-w. URL https://www.nature.com/articles/s41746-024-01344-w.

Saab, K., Tu, T., Weng, W.-H., Tanno, R., Stutz, D., Wulczyn, E., Zhang, F., Strother, T., Park, C., Vedadi, E., Chaves, J. Z., Hu, S.-Y., Schaekermann, M., Kamath, A., Cheng, Y., Barrett, D. G. T., Cheung, C., Mustafa, B., Palepu, A., McDuff, D., Hou, L., Golany, T., Liu, L., Alayrac, J.-b., Houlsby, N., Tomasev, N., Freyberg, J., Lau, C., Kemp, J., Lai, J., Azizi, S., Kanada, K., Man, S., Kulkarni, K., Sun, R., Shakeri, S., He, L., Caine, B., Webson, A., Latysheva, N., Johnson, M., Mansfield, P., Lu, J., Rivlin, E., Anderson, J., Green, B., Wong, R., Krause, J., Shlens, J., Dominowska, E., Eslami, S. M. A., Chou, K., Cui, C., Vinyals, O., Kavukcuoglu, K., Manyika, J., Dean, J., Hassabis, D., Matias, Y., Webster, D., Barral, J., Corrado, G., Semturs, C., Mahdavi, S. S., Gottweis, J., Karthikesalingam, A., and Natarajan, V. Capabilities of Gemini Models in Medicine, May 2024. URL http://arxiv.org/abs/2404.18416. arXiv:2404.18416 [cs].

Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., Chen, J., Mahvar, F., Yatziv, L., Chen, T., Sterling, B., Baby, S. A., Baby, S. M., Lai, J., Schmidgall, S., Yang, L., Chen, K., Bjornsson, P., Reddy, S., Brush, R., Philbrick, K., Asiedu, M., Mezerreg, I., Hu, H., Yang, H., Tiwari, R., Jansen, S., Singh, P., Liu, Y., Azizi, S., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Riviere, M., Rouillard, L., Mesnard, T., Cideron, G., Grill, J.-b., Ramos, S., Yvinec, E., Casbon, M., Buchatskaya, E., Alayrac, J.-B., Lepikhin, D., Feinberg, V., Borgeaud, S., Andreev, A., Hardin, C., Dadashi, R., Hussenot, L., Joulin, A., Bachem, O., Matias, Y., Chou, K., Hassidim, A., Goel, K., Farabet, C., Barral, J., Warkentin, T., Shlens, J., Fleet, D., Cotruta, V., Sanseviero, O., Martins, G., Kirk, P., Rao, A., Shetty, S., Steiner, D. F., Kirmizibayrak, C., Pilgrim, R., Golden, D., and Yang, L. MedGemma Technical Report, July 2025. URL http://arxiv.org/abs/2507.05201.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Scharli, N., Chowdhery, A., Mansfield, P., Arcas, B. A. y., Webster, D., Corrado, G. S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A., and Natarajan, V. Large Language Models Encode Clinical Knowledge, December 2022. URL http://arxiv.org/abs/2212.13138. arXiv:2212.13138 [cs].

Tu, H., Cui, C., Wang, Z., Zhou, Y., Zhao, B., Han, J., Zhou, W., Yao, H., and Xie, C. How Many Unicorns Are In This Image? A Safety Evaluation Benchmark For Vision LLMs. *arXiv preprint arXiv:2311.16101*, 2023.

Xiong, M., Hu, Z., Lu, X., LI, Y., Fu, J., He, J., and Hooi, B. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *The twelfth international conference on learning representations*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., Xu, Y., Wei, M., Wang, W., Ma, S., Wei, F., Yang, J., Li, C., Gao, J., Rosemon, J., Bower, T., Lee, S., Weerasinghe, R., Wright, B. J., Robicsek,

A., Piening, B., Bifulco, C., Wang, S., and Poon, H. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07441-w. URL https://www.nature.com/articles/s41586-024-07441-w. Publisher: Nature Publishing Group.

Zack, T., Lehman, E., Suzgun, M., Rodriguez, J. A., Celi, L. A., Gichoya, J., Jurafsky, D., Szolovits, P., Bates, D. W., Abdulnour, R.-E. E., Butte, A. J., and Alsentzer, E. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22, January 2024. ISSN 2589-7500. doi: 10.1016/S2589-7500(23)00225-X. URL https://www.thelancet.com/journals/landig/article/PIIS2589-7500(23)00225-X/fulltext.

Zhao, H., Si, S., Chen, L., Zhang, Y., Sun, M., Chang, B., and Zhang, M. Looking Beyond Text: Reducing Language Bias in Large Vision-Language Models via Multimodal Dual-Attention and Soft-Image Guidance. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19666–19690, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.995. URL https://aclanthology.org/2025.emnlp-main.995/.

Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.-M., and Lin, M. On evaluating adversarial robustness of large vision-language models. In *Proceedings of the 37th international conference on neural information processing systems*, Nips '23, New Orleans, LA, USA, 2023. Curran Associates Inc. Number of pages: 28 tex.address: Red Hook, NY, USA tex.articleno: 2355.

Zuo, Y., Qu, S., Li, Y., Chen, Z.-R., Zhu, X., Hua, E., Zhang, K., Ding, N., and Zhou, B. MedXpertQA: Benchmarking expert-level medical reasoning and understanding. *Forty-second international conference on machine learning*, 2025. URL https://openreview.net/forum?id=IyVcxU0RKI.

A. Technical Details

A.1. Query-based zeroth-order optimization algorithm

A.2. Inference prompts and parameters

All model queries are issued to an OpenAI-compatible `/chat/completions` endpoint and use restricted single-token decoding. The implementation exposes inference settings through environment variables, with the following defaults.

Prompting format. Each request contains two messages:

- **System prompt:**

  ```
  You are a medical VQA assistant. Answer with exactly one letter from the
  ↪  provided choices and nothing else.
  ```

- **User message:** a two-part multimodal content block consisting of (i) text and (ii) one image. The text is constructed as

  ```
  {question_text}

  Choices: {A, B, C, ...}
  Answer:
  ```

  where the choice letters are taken from the keys of `question["options"]`.

Decoding and log-probability extraction. We request log-probabilities for the next token and select the option with maximum log-probability among the candidate letters:

- `max_tokens = 1`
- `temperature = 0.0`
- `logprobs = true`
- `top_logprobs = 20`

Image encoding and transport. Images are sent inline as a data URL (base64) under an `image_url` content item. By default images are encoded as JPEG with `quality = 85`.

Algorithm 1 Query-based zeroth-order adversarial optimization (per example)

Input: question text q; original image $x_0 \in [0,1]^{3\times H\times W}$; answer set $\mathcal{A}$; ground-truth option $a^\star \in \mathcal{A}$; budget ϵ; steps T; NES samples m; base smoothing σ_0; grayscale flag BW.
Output: adversarial image x.
$x \leftarrow x_0$; $M \leftarrow 0$; $V \leftarrow 0$; $(\beta_1, \beta_2) \leftarrow (0.9, 0.999)$.
for $t = 1$ **to** T **do**
  $p \leftarrow (t-1)/\max(T-1, 1)$.
  $\alpha_t \leftarrow \alpha_{\min} + (\alpha_{\max} - \alpha_{\min}) \cdot \frac{1+\cos(\pi p)}{2}$.
  $\sigma_t \leftarrow \sigma_0 \cdot (1 + 2(1-p))$.
  Choose reduced resolution (h, w) (e.g., $h = \max(24, \lfloor H/12 \rfloor)$, $w = \max(24, \lfloor W/12 \rfloor)$).
  for $i = 1$ **to** m **do**
    Sample $u_i^{(s)} \sim \mathcal{N}(0, I)$ in $\mathbb{R}^{3\times h\times w}$.
    Normalize $u_i^{(s)} \leftarrow u_i^{(s)} / \|u_i^{(s)}\|_2$.
    Upsample $u_i \leftarrow \textsc{BilinearUpsample}(u_i^{(s)}, H, W)$.
    Normalize $u_i \leftarrow \sqrt{3HW} \cdot u_i / \|u_i\|_2$.
    $x_i^+ \leftarrow \mathrm{clip}_{[0,1]}(x + \sigma_t u_i)$.
    $x_i^- \leftarrow \mathrm{clip}_{[0,1]}(x - \sigma_t u_i)$.
    Query logprobs on (q, x_i^+) to obtain $\{\ell_a(x_i^+)\}_{a\in\mathcal{A}}$.
    Query logprobs on (q, x_i^-) to obtain $\{\ell_a(x_i^-)\}_{a\in\mathcal{A}}$.
    Compute $s(x) = \big(\ell_{a^\star}(x) - \max_{a\neq a^\star} \ell_a(x)\big) + \lambda\, \ell_{a^\star}(x)$ with $\lambda = 0.5$.
  end for
  $\hat{g} \leftarrow \frac{1}{2\sigma_t m} \sum_{i=1}^{m} \big(s(x_i^+) - s(x_i^-)\big)\, u_i$.
  $M \leftarrow \beta_1 M + (1-\beta_1)\hat{g}$
  $V \leftarrow \beta_2 V + (1-\beta_2)(\hat{g} \odot \hat{g})$.
  $\tilde{g} \leftarrow \frac{M/(1-\beta_1^t)}{\sqrt{V/(1-\beta_2^t)} + 10^{-8}}$.
  if BW **then**
    $\tilde{g} \leftarrow \textsc{ReplicateRGB}(0.2989\tilde{g}_R + 0.5870\tilde{g}_G + 0.1140\tilde{g}_B)$.
  end if
  $x' \leftarrow x - \alpha_t \,\mathrm{sign}(\tilde{g})$.
  $\delta \leftarrow \mathrm{clip}_{[-\epsilon,\epsilon]}(x' - x_0)$.
  $x \leftarrow \mathrm{clip}_{[0,1]}(x_0 + \delta)$.
  if $\arg\max_{a\in\mathcal{A}} \ell_a(x) \neq a^\star$ **then**
    break
  end if
end for
return x.

A.3. Additional Results

Table 4. Success counts and rates stratified by Body System, Medical Task, and Question Type under two perturbation budgets ($\varepsilon = 0.01$ and $\varepsilon = 0.02$), comparing grayscale and color images for `MedGemma 4B` with an Adam style optimizer. Results highlight modality-dependent performance differences and the effect of increasing ε on task success, with N/A indicating categories without applicable grayscale samples.

			Grayscale		Color	
Category	Value	ε	Success	Rate	Success	Rate
Body System	Cardiovascular	0.01	9/13	69.2%	42/54	77.8%
		0.02	9/13	69.2%	46/54	85.2%
	Digestive	0.01	4/12	33.3%	15/24	62.5%
		0.02	5/12	41.7%	16/24	66.7%
	Endocrine	0.01	3/3	100.0%	7/9	77.8%
		0.02	2/3	66.7%	7/9	77.8%
	Integumentary	0.01	N/A	N/A	13/28	46.4%
		0.02	N/A	N/A	16/28	57.1%
	Lymphatic	0.01	0/1	0.0%	6/9	66.7%
		0.02	0/1	0.0%	7/9	77.8%
	Muscular	0.01	1/1	100.0%	2/3	66.7%
		0.02	1/1	100.0%	2/3	66.7%
	Nervous	0.01	11/15	73.3%	30/39	76.9%
		0.02	12/15	80.0%	31/39	79.5%
	Other / NA	0.01	2/2	100.0%	3/5	60.0%
		0.02	2/2	100.0%	3/5	60.0%
	Reproductive	0.01	1/4	25.0%	8/12	66.7%
		0.02	1/4	25.0%	8/12	66.7%
	Respiratory	0.01	8/16	50.0%	19/28	67.9%
		0.02	7/16	43.8%	19/28	67.9%
	Skeletal	0.01	19/29	65.5%	36/53	67.9%
		0.02	16/29	55.2%	35/53	66.0%
	Urinary	0.01	0/1	0.0%	4/7	57.1%
		0.02	0/1	0.0%	4/7	57.1%
Medical Task	Basic Science	0.01	12/14	85.7%	40/55	72.7%
		0.02	10/14	71.4%	41/55	74.5%
	Diagnosis	0.01	30/55	54.5%	101/151	66.9%
		0.02	31/55	56.4%	111/151	73.5%
	Treatment	0.01	16/28	57.1%	44/65	67.7%
		0.02	14/28	50.0%	42/65	64.6%
Question Type	Reasoning	0.01	51/87	58.6%	152/219	69.4%
		0.02	50/87	57.5%	158/219	72.1%
	Understanding	0.01	7/10	70.0%	33/52	63.5%
		0.02	5/10	50.0%	36/52	69.2%

Table 5. Success counts and rates stratified by Body System, Medical Task, and Question Type under two perturbation budgets ($\varepsilon = 0.01$ and $\varepsilon = 0.02$), comparing grayscale and color images for `Qwen3-VL-Instruct 4B` with an Adam style optimizer. Results highlight modality-dependent performance differences and the effect of increasing ε on task success, with N/A indicating categories without applicable grayscale samples.

			Grayscale		Color	
Category	Value	ε	Success	Rate	Success	Rate
Body System	Cardiovascular	0.01	5/15	33.3%	34/53	64.2%
		0.02	8/15	53.3%	34/53	64.2%
	Digestive	0.01	4/7	57.1%	8/16	50.0%
		0.02	3/7	42.9%	7/16	43.8%
	Endocrine	0.01	1/2	50.0%	5/9	55.6%
		0.02	2/2	100.0%	6/9	66.7%
	Integumentary	0.01	N/A	N/A	25/33	75.8%
		0.02	N/A	N/A	24/33	72.7%
	Lymphatic	0.01	N/A	N/A	5/8	62.5%
		0.02	N/A	N/A	6/8	75.0%
	Muscular	0.01	1/1	100.0%	6/9	66.7%
		0.02	1/1	100.0%	6/9	66.7%
	Nervous	0.01	3/8	37.5%	16/33	48.5%
		0.02	3/8	37.5%	18/33	54.5%
	Other / NA	0.01	N/A	N/A	0/3	0.0%
		0.02	N/A	N/A	0/3	0.0%
	Reproductive	0.01	5/6	83.3%	14/21	66.7%
		0.02	5/6	83.3%	13/21	61.9%
	Respiratory	0.01	9/20	45.0%	13/27	48.1%
		0.02	7/20	35.0%	14/27	51.9%
	Skeletal	0.01	12/23	52.2%	26/47	55.3%
		0.02	10/23	43.5%	24/47	51.1%
	Urinary	0.01	1/1	100.0%	2/6	33.3%
		0.02	1/1	100.0%	4/6	66.7%
Medical Task	Basic Science	0.01	7/11	63.6%	35/50	70.0%
		0.02	8/11	72.7%	39/50	78.0%
	Diagnosis	0.01	22/42	52.4%	80/144	55.6%
		0.02	21/42	50.0%	79/144	54.9%
	Treatment	0.01	12/30	40.0%	39/71	54.9%
		0.02	11/30	36.7%	38/71	53.5%
Question Type	Reasoning	0.01	38/77	49.4%	118/213	55.4%
		0.02	36/77	46.8%	117/213	54.9%
	Understanding	0.01	3/6	50.0%	36/52	69.2%
		0.02	4/6	66.7%	39/52	75.0%

Table 6. Success counts and rates stratified by Body System, Medical Task, and Question Type under two perturbation budgets ($\varepsilon = 0.01$ and $\varepsilon = 0.02$), comparing grayscale and color images for `MedGemma 27B` with an Adam style optimizer. Results highlight modality-dependent performance differences and the effect of increasing ε on task success, with N/A indicating categories without applicable grayscale samples.

			Grayscale		Color	
Category	Value	ε	Success	Rate	Success	Rate
Body System	Cardiovascular	0.01	2/12	16.7%	27/55	49.1%
		0.02	2/12	16.7%	28/55	50.9%
	Digestive	0.01	3/14	21.4%	7/23	30.4%
		0.02	3/14	21.4%	7/23	30.4%
	Endocrine	0.01	1/2	50.0%	5/12	41.7%
		0.02	1/2	50.0%	5/12	41.7%
	Integumentary	0.01	N/A	N/A	14/43	32.6%
		0.02	N/A	N/A	13/43	30.2%
	Lymphatic	0.01	N/A	N/A	5/7	71.4%
		0.02	N/A	N/A	5/7	71.4%
	Muscular	0.01	2/4	50.0%	2/8	25.0%
		0.02	1/4	25.0%	2/8	25.0%
	Nervous	0.01	3/13	23.1%	17/39	43.6%
		0.02	5/13	38.5%	19/39	48.7%
	Other / NA	0.01	2/2	100.0%	3/5	60.0%
		0.02	2/2	100.0%	4/5	80.0%
	Reproductive	0.01	1/7	14.3%	6/18	33.3%
		0.02	1/7	14.3%	6/18	33.3%
	Respiratory	0.01	4/28	14.3%	13/40	32.5%
		0.02	3/28	10.7%	14/40	35.0%
	Skeletal	0.01	7/28	25.0%	17/44	38.6%
		0.02	10/28	35.7%	16/44	36.4%
	Urinary	0.01	0/2	0.0%	2/5	40.0%
		0.02	0/2	0.0%	2/5	40.0%
Medical Task	Basic Science	0.01	4/15	26.7%	22/46	47.8%
		0.02	4/15	26.7%	24/46	52.2%
	Diagnosis	0.01	16/71	22.5%	71/172	41.3%
		0.02	16/71	22.5%	73/172	42.4%
	Treatment	0.01	5/26	19.2%	25/81	30.9%
		0.02	8/26	30.8%	24/81	29.6%
Question Type	Reasoning	0.01	23/102	22.5%	92/243	37.9%
		0.02	25/102	24.5%	97/243	39.9%
	Understanding	0.01	2/10	20.0%	26/56	46.4%
		0.02	3/10	30.0%	24/56	42.9%