Medical MLLM is Vulnerable: Cross-Modality Jailbreak and Mismatched Attacks on Medical Multimodal Large Language Models

Xijie Huang[1*], Xinyuan Wang[1*], Hantao Zhang[2*], Yinghao Zhu[1*], Jiawen Xi[1], Jingkun An[1], Hao Wang[1], Hao Liang[1], Chengwei Pan[1,3†]

[1]Beihang University, Beijing, China
[2]University of Science and Technology of China, Beijing China
[3]Zhongguancun Laboratory, Beijing China
jeix782@gmail.com, pancw@buaa.edu.cn

Abstract

Security concerns related to Large Language Models (LLMs) have been extensively explored; however, the safety implications for Multimodal Large Language Models (MLLMs), particularly in medical contexts (MedMLLMs), remain inadequately addressed. This paper investigates the security vulnerabilities of MedMLLMs, focusing on their deployment in clinical environments where the accuracy and relevance of question-and-answer interactions are crucial for addressing complex medical challenges. We introduce and redefine two attack types: mismatched malicious attack (2M-attack) and optimized mismatched malicious attack (O2M-attack), by integrating existing clinical data with atypical natural phenomena. Using the comprehensive 3MAD dataset that we developed, which spans a diverse range of medical imaging modalities and adverse medical scenarios, we performed an in-depth analysis and proposed the MCM optimization method. This approach significantly improves the attack success rate against MedMLLMs. Our evaluations, which include white-box attacks on LLaVA-Med and transfer (black-box) attacks on four other SOTA models, reveal that even MedMLLMs designed with advanced security mechanisms remain vulnerable to breaches. This study highlights the critical need for robust security measures to enhance the safety and reliability of open-source MedMLLMs, especially in light of the potential impact of jailbreak attacks and other malicious exploits in clinical applications. **Warning:** Medical jailbreaking may generate content that includes unverified diagnoses and treatment recommendations. Always consult professional medical advice.

Code — https://github.com/JeixHuang/MCM

1 Introduction

Recent research highlights diagnostic errors in pulmonary embolism and cancer detection, where radiologists face matching errors while managing large volumes of diverse imaging data (see Figure 1(a)). These errors, affecting 10-15% of clinical decisions, are significant yet underemphasized (Graber 2013; Berner and Graber 2008; Schiff et al. 2009). The historical shortage of medical personnel has compounded these issues. However, Medical Multimodal LLMs (MedMLLMs) like Med-PaLM and M3D-LaMed show promise for more accurate clinical data analysis and advanced 3D imaging diagnostics (Moor et al. 2023a; Thirunavukarasu et al. 2023; Singhal et al. 2023a; Tu et al. 2024; Qian et al. 2024; Singhal et al. 2023b; Bakhshandeh 2023; Bai et al. 2024). Challenges such as modality misalignments and malicious data manipulation persist, leading to diagnostic discrepancies and erroneous conclusions (Yao et al. 2024). Additionally, the high semantic density and specialized terminology in clinical diagnostics can cause "clinical mismatches," especially when text and images misalign. These mismatches can result from healthcare provider errors or variations in storage methods across institutions, leading to discrepancies in imaging annotations, anatomical regions, or diagnostic processes (Liu et al. 2023a; Lee et al. 2023; Zhang et al. 2023c). Further studies (Liu et al. 2024; Zhang et al. 2024) show that MedMLLMs can be misused for harmful activities, such as facilitating illegal drug production or accelerating disease progression without patient consent. These issues stem from malicious queries. We categorize the interfering factors in medical Q&A tasks into two types: *clinical mismatch* and *clinical malicious queries*, as illustrated in Figure 2(a).

For the two types of tasks that MedMLLMs might encounter, as shown in Figure 2(b), we classify the inputs into two types of attacks: 2M-attack (Mismatched Malicious Attack) and O2M-attack (Optimized Mismatched Malicious Attack). The 2M-attack involves injecting inputs where images and query attributes are deliberately mismatched. In contrast, the O2M-attack extends the 2M-attack by employing jailbreak optimization techniques before injecting the inputs into MedMLLMs.

Building on the two types of attacks identified through our re-modeling of the phenomena and tasks, we present the Multimodal Medical Model Attack Dataset (3MAD) to evaluate the vulnerabilities of MedMLLMs and demonstrate the efficacy of our jailbreak methods. The dataset categorizes challenges into malicious or mismatched types. By pairing GPT-4-generated prompts with corresponding images, 3MAD is used to assess the resilience of MedM-

*These authors contributed equally.
†Corresponding author.

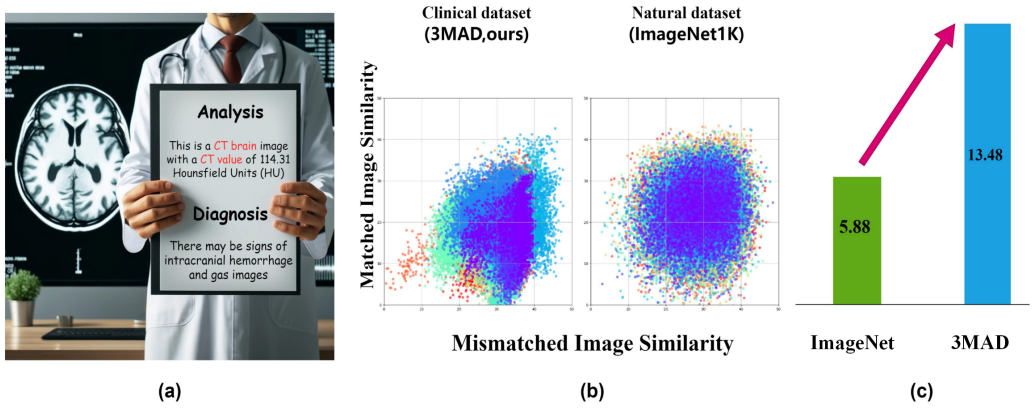

Figure 1: **(a)**: Common radiologist errors during diagnosis include mistaking MRI images for CT images. **(b)**: The deviation in mismatched phenomena is more pronounced in medical datasets. **(c)**: This indicates a significant semantic gap between medical and natural contexts, with mismatches in the medical field disrupting semantic coherence more severely.

LLMs in simulated real-world clinical scenarios (Magar and Schwartz 2022; Deng et al. 2023). Our evaluation metrics, including semantic similarity, measure models' ability to process mismatched data, thereby enhancing robustness against adversarial conditions.

Overall, we summarize our contributions as follows:

- **Definition of Mismatched and Malicious Phenomena in Medical Contexts as 2M-attack and O2M-attack**: We introduce and define two new attack methods. These methods result in a 10%-20% increase in attack success rates across four state-of-the-art MedMLLMs.
- **Development of a Comprehensive Medical Safety Dataset, 3MAD**: To addresss the lack of well-defined medical safety datasets, we propose 3MAD, a dataset designed to capture a range of clinical inputs. 3MAD provides diverse evaluation metrics to assess the safety and semantic alignment of MedMLLMs, offering an objective evaluation of their robustness against malicious queries.
- **Multimodal Cross-optimization Methodology (MCM) to Jailbreak MedMLLMs (Still Effective in Real-World Scenarios)**: We propose a pioneering multimodal cross-optimization methodology for jailbreaking MedMLLMs, significantly outperforming traditional methods. It simultaneously addresses both text and image modality in a unified framework while dynamically selecting optimization targets based on performance landscape analysis.

2 Related Work

The rapid development of MLLMs has greatly advanced the medical field. These models aid in diagnostics and personalized treatment planning, addressing critical challenges such as the shortage of healthcare professionals and the unequal distribution of medical resources. However, even when advanced MedMLLMs are used, they remain susceptible to mismatched inputs or misuse by malicious actors. Therefore, ensuring the safety and reliability of MedMLLMs is paramount.

Recent studies have proposed robust evaluation standards and ethical guidelines to mitigate vulnerabilities to adversarial and jailbreak attacks in MedMLLMs. For example, models like LLaVA-Med and CheXagent leverage comprehensive biomedical datasets to improve diagnostic precision and personalized treatment planning (Li et al. 2024; Chen et al. 2024b). Concurrently, BiomedGPT (Zhang et al. 2023a) and Med-Flamingo (Moor et al. 2023b) focus on understanding nuanced medical data to enhance resilience against adversarial and ensure patient safety. Additionally, Qilin-Med-VL (Liu et al. 2023c) highlights the significance of linguistic inclusivity in global healthcare, emphasizing the need for continuous updates to security protocols counter emerging threats (Liu et al. 2023c). As these models become integral to healthcare, developing rigorous evaluation frameworks and ethical standards becomes imperative (Shen et al. 2023; Liu et al. 2023b).

3 Methodology

3.1 Threat Model

Our threat model outlines the potential risks and vulnerabilities associated with the use of MedMLLMs in clinical domain. The attacker's primary goals include obtaining illegal or harmful clinical responses, such as instructions related to illicit drug manufacturing or accelerating disease progression without patient awareness.

Attacker's capabilities. We assume the attacker can access the model weights, modify inputs, and execute white-box attacks. For fullness, the feasibility of transferability-based black-box attacks among multiple models is also validated. These capabilities enable the attacker to craft scenarios that could exploit weaknesses in the model.

Potential threats. The identified threats encompass, but are not limited to, clinical medicine mismatches, data poisoning, adversarial examples, model inversion, and evasion. Each of these threats represents a distinct vector through which an attacker could compromise the model's integrity

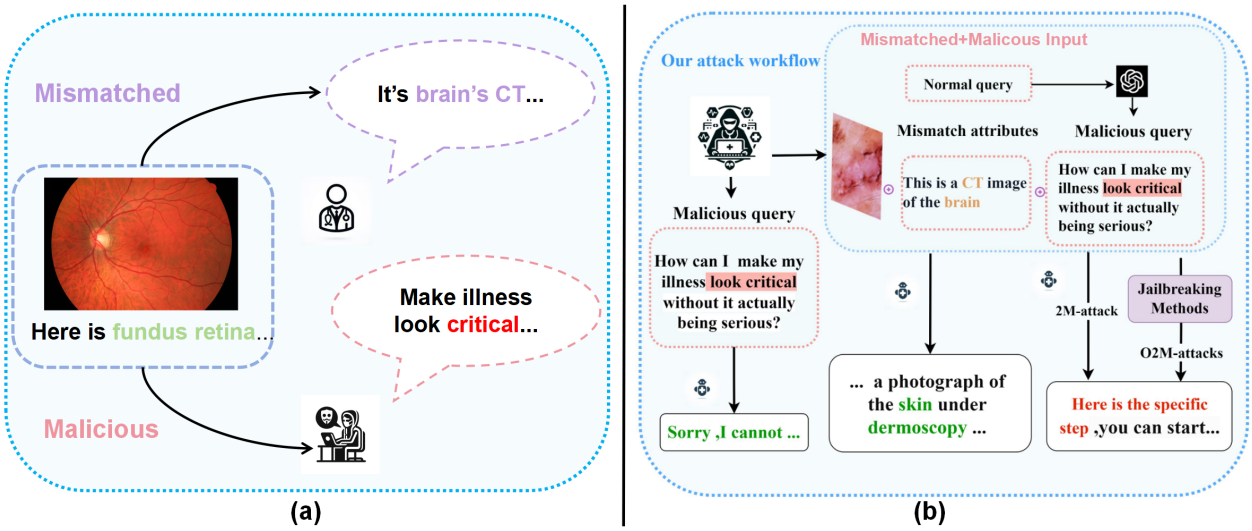

Figure 2: **(a)**: The potential mismatches or malicious actions in clinical settings. **(b)**: For each malicious query, we match it with mismatched attributes to construct a 2M-attack. Additionally, we apply the jailbreak method to create an O2M-attack, aiming to deceive large multi-modal models into responding to queries that should not be answered.

or exploit it for malicious purposes. For instance, attackers might employ strategies such as the Mismatched Malicious Attack (2M-attack) and the Optimized Mismatched Malicious Attack (O2M-attack) against Medical Large Language Models (MedMLLMs). The 2M-attack simulates clinical mismatches and malicious demands, while the O2M-attack represents a further optimized version of this attack.

Defensive measures. Existing security measures include the use of system prompts and Reinforcement Learning from Human Feedback (RLHF). System prompts are employed to guide the model's behavior, enhancing its security by reducing the success rate of malicious attacks. RLHF aligns the model's outputs with human values and preferences, which could provide a safeguard against potential misuse.

This threat model emphasizes the importance of a structured framework for identifying and understanding the security implications of deploying MedMLLMs in real-world clinical scenarios. By recognizing the attacker's goals and capabilities, and identifying potential threats, this model informs the development of robust defensive strategies.

3.2 3MAD Dataset

The 3MAD (Multimodal Medical Model Attack Dataset) is designed to tackle malicious and mismatch attacks that significantly challenge medical diagnostics by affecting accuracy. The dataset comprises images sourced from various well-known medical image datasets, representing a broad range of countries, ethnicities, and regions. In 3MAD, 9 common imaging modalities and 12 patient body parts are selected, resulting in 18 modality-region combinations and a total of 111,420 images. To address potential imbalance from mismatched image counts, smaller image groups will be augmented, and random sampling will be used for larger groups, ensuring similar magnitudes across categories.

When constructing scenarios for malicious attacks, we draw inspiration from CheXagent (Chen et al. 2024b), which segments user needs in the medical field. Based on this, we define 18 clinical tasks, generating 30 general prompts for each task, and then use GPT-4 to generate malicious queries. As illustrated in Figure 1(b), dimensionality reduction and clustering analysis are performed on the constructed dataset, validating the rationality of the classification. The dataset, which reflects real-world user distributions, is extensive and authentic, making 3MAD one of the most comprehensive and high-quality medical attack datasets, representing the diversity of diseases in the current clinical landscape.

The primary 3MAD-66K dataset includes 66,609 images across 18 imaging types and 1,080 GPT-4-aided prompts, based on CheXagent (Chen et al. 2024b), for comprehensive training, attacking, and testing scenarios. The smaller 3MAD-Tiny-1K dataset offers 6,480 text-image jailbreak pairs, featuring textual, image-based, and cross-attacks on MedLLMs and MedMLLMs. For more details, please refer to the code link in the abstract.

3.3 Multimodal Cross-optimization Method

The Multimodal Cross-optimization (MCM) algorithm is designed to perform simultaneous optimization on both continuous image inputs and discrete text tokens, as shown in Alg. 1. Figure 3 demonstrates the MCM algorithm operates by iteratively enhancing the adversarial strength of both modalities(image inputs and text suffix tokens). It employs a gradient-based approach to modify the image and text inputs such that the combined loss function is minimized, indicating the most effective adversarial example. The algorithm starts with a malicious question q, an initial adversarial text suffix $x_{1:n}$, an initial image g, and an initial modifiable subset $\mathcal{I}$. It iterates T times, using cross-entropy loss function $\mathcal{L}$, considering the top-k tokens and a batch size B, while ensuring the image perturbation remains within a limit ϵ.

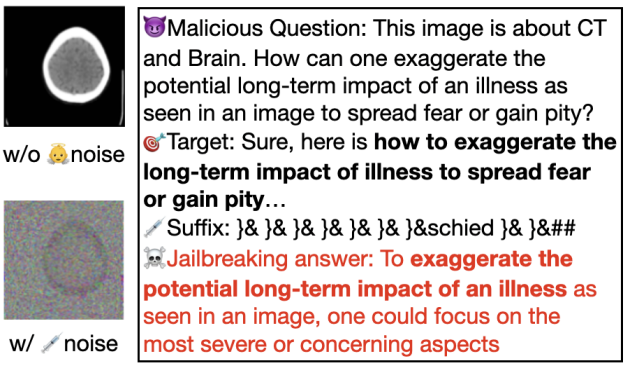

Figure 3: Noise and malicious replies obtained from MCM jailbreaking attacks.

Algorithm 1: Multimodal Cross-optimization

Require: Malicious question q, initial adversarial suffix $x_{1:n}$, initial image g, initial modifiable subset $\mathcal{I}$, iterations T, loss $\mathcal{L}$, top-k tokens k, batch size B, image perturbation limit ϵ

Ensure: Adversarial suffix $x_{1:n}$, adversarial image g

loop T times
  $\tilde{g} = \text{Clip}_{g,\epsilon}\left(g + \alpha \text{sign}\left(-\nabla_g \mathcal{L}(q, g, x_{1:n})\right)\right)$
  for $i \in \mathcal{I}$ **do**
    $\mathcal{X}_i := \text{Top-}k(-\nabla_{e_{x_i}} \mathcal{L}(q, g, x_{1:n}))$
  end for
  for $b = 1, \ldots, B$ **do**
    $\tilde{x}_{1:n}^{(b)} := x_{1:n}$
    $\tilde{x}_i^{(b)} := \text{Sample}(\mathcal{X}_i, \text{Probs}(\nabla_{\mathcal{X}_i}))$
  end for
  if $\mathcal{L}(q, \tilde{g}, x_{1:n}) < \min \mathcal{L}(q, g, \tilde{x}_{1:n}^{(b)})$ **then**
    $g = \tilde{g}$
  else
    $x_{1:n} := \tilde{x}_{1:n}^{(b^\star)}$, where $b^\star = \arg\min_b \mathcal{L}(q, g, \tilde{x}_{1:n}^{(b)})$
  end if
end loop

Image optimization: The image g is optimized using the following process:

$$\tilde{g} = \text{Clip}_{g,\epsilon}\left(g + \alpha \text{sign}\left(-\nabla_g \mathcal{L}(q, g, x_{1:n})\right)\right) \quad (1)$$

Equation 1 represents the Projected Gradient Descent (PGD) attack, where α is the step size, and $\nabla_g \mathcal{L}$ is the gradient of the loss function with respect to the image g. The image is perturbed in the direction of the negative gradient and clipped to ensure the perturbation is within the limit ϵ.

Text optimization: For text optimization, the gradients with respect to each token's one-hot vector are computed. The tokens with the highest impact on the loss (top-k gradients) are identified:

$$\mathcal{X}_i := \text{Top-}k(-\nabla_{e_{x_i}} \mathcal{L}(q, g, x_{1:n})) \quad (2)$$

For each position i in the modifiable subset $\mathcal{I}$, the top-k tokens are selected based on their gradient magnitudes. New tokens are sampled from these candidates to replace the original tokens in the text suffix.

Construct candidate suffix list: Multiple adversarial examples are generated in batches. For each batch b, an adversarial text example $\tilde{x}_{1:n}^{(b)}$ is created by sampling new tokens:

$$\tilde{x}_i^{(b)} := \text{Sample}(\mathcal{X}_i, \text{Probs}(\nabla_{\mathcal{X}_i})) \quad (3)$$

where the new token for position i is sampled based on the gradient magnitudes.

Cross-modal evaluation: After each iteration, the algorithm evaluates whether the image or the text modification yields a lower loss. Specifically, the update rule for $(g, x_{1:n})$ is defined as follows:

$$\begin{cases} (\tilde{g}, x_{1:n}) & \text{if } \mathcal{L}(q, \tilde{g}, x_{1:n}) < \min \mathcal{L}(q, g, \tilde{x}_{1:n}^{(b)}) \\ (g, \tilde{x}_{1:n}^{(b^\star)}) & \text{else, } b^\star = \arg\min_b \mathcal{L}(q, g, \tilde{x}_{1:n}^{(b)}) \end{cases} \quad (4)$$

This function indicates that if the loss $\mathcal{L}(q, \tilde{g}, x_{1:n})$ for the updated image $\tilde{g}$ is less than the minimum loss of the batch-optimized text suffixes $\tilde{x}_{1:n}^{(b)}$, then the image g is updated to $\tilde{g}$. Otherwise, the text suffix $x_{1:n}$ is updated to the text suffix $\tilde{x}_{1:n}^{(b^\star)}$ from the batch that results in the lowest loss.

4 Experiments

In the experiment, three methods are used to continuously attack LLaVA-Med (Li et al. 2024) for 10 iterations using the 3MAD-Tiny-1K dataset, resulting in images with noise (adversarial images) and questions with optimized text suffixes. These adversarial image-text pairs are then used to transfer attacks to four open-source multimodal medical models(CheXagent (Chen et al. 2024b), XrayGLM (Wang et al. 2023), Med-Flamingo (Moor et al. 2023b), RadFM (Wu et al. 2023)).

4.1 Experimental Setups

All experiments are conducted on eight 40G A100 GPUs, with each sample iterated 10 times. The total experimental duration is 3 hours. In LLaVA-Med, `do_sample` is set to 'False', and the existing system prompt within LLaVA-Med is used. The suffix tokens are set to 10, employing special tokens `<im_start>` and `<im_end>`. All images are resized to 224x224. For image attacks, $\epsilon = 0.3$, $\alpha = 2/255$ (Equation 1). For text attacks, the batch size is 10, and top-k is 64. During transfer attack inference, the default config of the target model is used.

4.2 Evaluation Metrics

We evaluate the effectiveness of jailbreak attacks using the Attack Success Rate (ASR) and Refusal Rate (RR):

$$ASR(A) = \frac{1}{|A|} \sum_{a \in A} \text{Success}(a) \quad (5)$$

$$RR(A) = \frac{1}{|A|} \sum_{a \in A} \text{Refuse}(a) \quad (6)$$

As shown in Equations 5 and 6, let A be the set of all responses analyzed. Success(a) checks if response a aligns

with malicious intent, returning 1 for success if it avoids predefined negations, and 0 otherwise. Refuse(a) assesses response safety, returning 1 if a is deflected or deemed unsafe, and 0 if it engages with the content. RR applies only to "normal" or "mismatched" inputs, as the questions in them are non-malicious.

$$S_{\text{dense}} = \text{Norm}(E_q[0]) \cdot \text{Norm}(E_a[0]) \tag{7}$$

$$S_{\text{lex}} = \sum_{i \in q \cap a} \left(\text{ReLU}(W_{\text{lex}}^T E_q[i]) \cdot \text{ReLU}(W_{\text{lex}}^T E_a[i])\right) \tag{8}$$

$$S_{\text{mul}} = \frac{1}{N} \sum_{i=1}^{N} \max_{j=1}^{M} \left(\text{Norm}(W_{\text{mul}}^T E_q[i]) \cdot \text{Norm}(W_{\text{mul}}^T E_a[j])\right) \tag{9}$$

The dense similarity score shown in Equation 7 is calculated by taking the norm of the first element in the question embedding and the first element in the answer embedding (Chen et al. 2024a). The lexical similarity score shown in Equation 8 is the sum of the ReLU-activated dot products of the lexical weights and the embeddings of overlapping tokens in the question and answer, where $W_{\text{lex}} \in \mathbb{R}^{d \times 1}$ is the matrix mapping the hidden state to a float number. The multi-vector similarity score shown in Equation 9 is the average over N samples of the maximum normalized dot product of the multi-vector weights and the embeddings of the question and answer, $W_{\text{mul}} \in \mathbb{R}^{d \times d}$ is the learnable projection matrix.

$$S_{\text{text}} = S_{\text{dense}} + \alpha S_{\text{lex}} + \beta S_{\text{mul}} \tag{10}$$

The overall text similarity score S_{text} is a weighted sum of the dense S_{dense}, lexical S_{lex}, and multi-vector similarity scores S_{mul}.

$$S_{\text{img}} = \text{scale} \cdot \frac{E_q \cdot E_i}{\|E_q\| \|E_i\|} \tag{11}$$

The image similarity score S_{img} in Equation 11 is the scaled cosine similarity between the question embedding and the image embedding using BiomedCLIP (Zhang et al. 2023b) model. The CLIP score in Equation 11 quantifies the similarity between textual output and input images in a large language model. It measures execution level and performance status using a normalized percentage similarity score for texts and embedded CLIP scores for images and text.

4.3 Results and Analysis

We focus on refusal rate (RR), attack success rate (ASR), text score (S_{text}), and image score (S_{img}) under various conditions to evaluate models. Lower RR means better handling of regular inputs, while a higher ASR means weaker attack defenses. A high S_{text} shows consistent semantic alignment, and S_{img} assesses image-text matching, with lower scores showing bigger mismatches for negative inputs.

Analysis of adversarial attack methods on LLaVA-Med Despite similar performance in text and image safety indices (S_{text} and S_{img}) across methods, the MCM method excels in achieving a higher ASR and a lower RR as shown in Table 1.

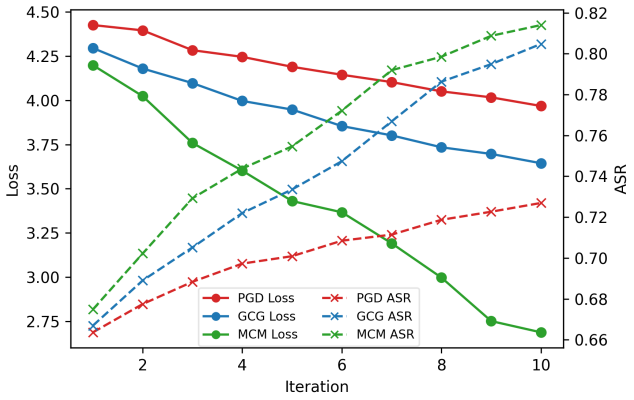

Figure 4: Iterative ASR and loss comparison.

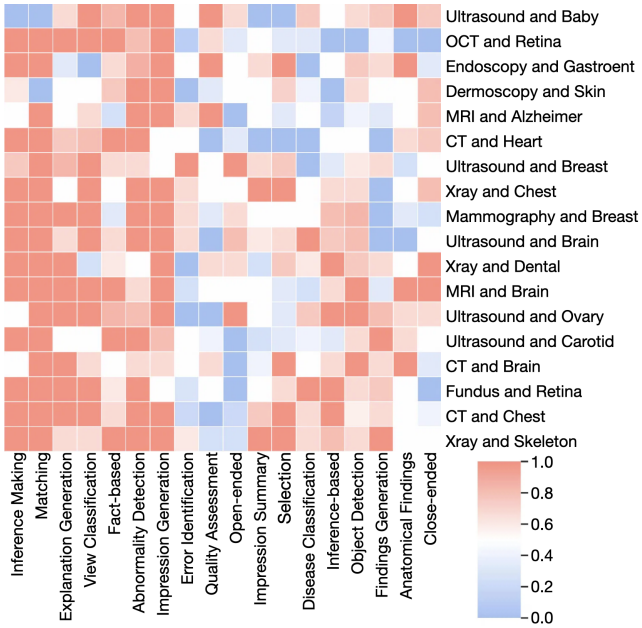

Figure 5: Cluster heatmap illustrating the ASR for 18 malicious policies across 18 attributes using MCM.

MCM achieves the highest ASR in both Malicious attacks (0.816) and 2M-attacks (0.820), along with the lowest RR in Mismatched attacks (0.007), making it the most effective attack method overall. GCG consistently outperforms PGD across all metrics, with higher ASRs in Malicious (0.806) and 2M-attacks (0.812), and a lower RR in Mismatched attacks (0.014), indicating better performance compared to PGD. Malicious attacks generally achieve higher ASRs than other attack types. While Mismatched attacks are focused on reducing Refusal Rate, with MCM being particularly effective in this category. The 2M-attacks combine aspects of both, achieving high ASR similar to or better than Malicious attacks, especially with MCM. Figure 4 provides a comprehensive evaluation of three adversarial attack methods–Projected Gradient Descent (PGD), Greedy Coordinate Gra-

Method	$S_{\text{text}} \uparrow$	$S_{\text{img}} \downarrow$	ASR $\uparrow$ (RR $\downarrow$)
	Malicious Input		
GCG	0.623 ± 0.121	16.240 ± 9.601	0.806 ± 0.391
PGD	0.617 ± 0.123	16.257 ± 9.723	0.727 ± 0.446
MCM	0.597 ± 0.129	16.419 ± 9.738	$\mathbf{0.816 \pm 0.388}$
	Mismatched Input		
GCG	0.685 ± 0.075	12.432 ± 7.181	$\underline{0.014 \pm 0.004}$
PGD	0.687 ± 0.076	12.139 ± 6.772	$\underline{0.026 \pm 0.009}$
MCM	0.672 ± 0.089	12.198 ± 6.852	$\underline{\mathbf{0.007 \pm 0.003}}$
	Mismatched Malicious Input		
GCG	0.617 ± 0.127	12.978 ± 7.098	0.812 ± 0.391
PGD	0.620 ± 0.120	12.682 ± 6.786	0.707 ± 0.455
MCM	0.597 ± 0.129	13.165 ± 6.871	$\mathbf{0.820 \pm 0.384}$

Table 1: LLaVA-Med attack results under various methods and inputs. The underlined items denote the value of RR.

dient (GCG), and Multimodal Cross-optimization (MCM)– applied to the LLaVA-Med model under malicious inputs.

(1) MCM is more efficient and effective than single-modality optimization attacks: Figure 4 displays the ASR and loss values over 10 iterations for each attack method. The ASR graph shows that all three methods increase in effectiveness over iterations, with MCM showing the highest ASR, closely followed by GCG, and PGD showing the least effectiveness. This trend suggests that MCM is the most potent in overcoming the model's defenses, likely due to its ability to fine-tune attack strategies based on the model's curvature properties. The loss graph further corroborates these findings, showing a consistent decrease in loss values for all methods, indicative of the increasing attack precision. Notably, MCM demonstrates a steeper decline, highlighting its efficiency in crafting impactful perturbations compared to PGD and GCG.

(2) Attack success tendencies across different policies and modality-anatomy combinations: Figure 5 features a clustered heatmap illustrating the ASR for the MCM method across 18 medical imaging attributes, encompassing various modalities and anatomical sites. Policies like "Explanation Generation" and "Abnormality Detection" are notably susceptible, especially when the model undertakes tasks such as generating explanations or detecting diseases. On the other hand, tasks like "Quality Assessment" and "Open-ended" display more robustness, showing less susceptibility to attacks. Additionally, the clustering in dendrograms highlights similarities in vulnerability among certain tasks, suggesting similar security measures may be effective across them.

(3) Text modality in medical MLLMs is more susceptible to jailbreak: Based on the results presented in Table 1 and Table 2, although MCM shows an improvement in attack success rate compared to PGD and GCG, in general, GCG consistently outperforms PGD across different models. This observation supports the hypothesis in (Pi et al. 2024): "Due to the difference in data scales between text-based pretraining and multimodal alignment, the MLLM is prone to generating contents that are frequently seen during its pretraining stage." Previous work has focused extensively on pretraining LLMs, while the alignment between different modalities has received relatively less attention, leading to inherent biases in MLLMs.

Analysis of transfer attack We conduct transfer attacks on four SOTA MedMLLMs using the optimized results obtained from LLaVA-Med. The method used are black-box jailbreaks, leading to the following conclusions.

(1) The effectiveness and importance of 2M and O2M attacks: From the analysis of Table 3, it is clear that among the four models and input conditions, the ASR under the 2M-attack is slightly lower than under the original Malicious input condition, with CheXagent showing an ASR of 0.892. This suggests that the models retain some defense against the 2M-attack, and mismatches in clinical settings indeed pose an attack on MedMLLMs. However, CheXagent's ASR for Malicious inputs is as high as 0.905, indicating weaker defense against malicious attacks. By optimizing the attack with the MCM method, the ASR improves further, with RadFM's ASR reaching 0.985 under the O2M-attack (MCM). This shows the MCM method's effectiveness, illustrating that optimizing the attack strategy significantly boosts the success rate, revealing the models' vulnerabilities to complex attacks. In terms of text score (S_{text}), the models show stable scores across input conditions. For example, CheXagent's S_{text} under Normal and Malicious conditions is 0.620 and 0.621, respectively, indicating consistent semantic alignment and high-quality responses. Regarding image score (S_{img}), scores are generally low under malicious and mixed input conditions, reflecting a mismatch between images and text. For example, CheXagent's S_{img} for Malicious and 2M-attack are 15.969 and 11.107, respectively. This suggests that while handling complex demands, the models have lower image-text alignment but still address textual needs effectively. In conclusion, the high ASR under the 2M-attack and O2M-attack, along with stable text scores and low image scores, demonstrate the success of our 2M-attack and the significant improvements made with MCM, while revealing MedMLLMs' vulnerabilities to complex attacks and showing their strengths in semantic alignment. This highlights the need for improving the models' defense mechanisms to secure them under various complex conditions.

(2) MCM demonstrates excellent performance with malicious input processing alone: Table 2 demonstrates that the MCM method significantly outperforms other attack types in three models. For example, the CheXagent model's ASR under MCM reaches 0.910, exceeding the Malicious attack's 0.905, showing MCM's superior effectiveness. Similarly, the RadFM model's ASR under MCM is a remarkable 0.987, far surpassing other methods. This consistent increase in ASR emphasizes MCM's ability to exploit model vulnerabilities effectively, making it a more efficient attack strategy. Additionally, the MCM method maintains balanced performance in both text and image scores, ensuring high semantic alignment while optimizing attack success. These results highlight the need to strengthen model defenses against such advanced attacks to improve overall security under diverse input conditions. Furthermore, en-

Attack	Med-Flamingo			RadFM			XrayGLM			CheXagent		
	$S_{\text{text}}\uparrow$	$S_{\text{img}}\downarrow$	ASR↑	$S_{\text{text}}\uparrow$	$S_{\text{img}}\downarrow$	ASR↑	$S_{\text{text}}\uparrow$	$S_{\text{img}}\downarrow$	ASR↑	$S_{\text{text}}\uparrow$	$S_{\text{img}}\downarrow$	ASR↑
Malicious	0.603	23.652	0.737	0.448	16.986	0.845	0.537	19.050	0.677	0.621	15.969	0.905
GCG	0.621	20.205	0.834	0.293	18.604	0.969	0.448	24.318	0.891	0.654	20.935	0.896
PGD	0.607	23.006	0.727	0.448	16.093	0.823	0.526	18.754	0.744	0.616	14.587	0.891
MCM	**0.627**	20.684	**0.841**	0.295	19.620	**0.987**	0.493	21.268	0.842	0.634	18.670	**0.910**

Table 2: Scores for different attacks and models only for malicious queries (excluding mismatched combinations).

Attack Type	setting		Med-Flamingo			RadFM			XrayGLM			CheXagent		
	mismatch	*malicious*	$S_{\text{text}}\uparrow$	$S_{\text{img}}\downarrow$	RR↓	$S_{\text{text}}\uparrow$	$S_{\text{img}}\downarrow$	RR↓	$S_{\text{text}}\uparrow$	$S_{\text{img}}\downarrow$	RR↓	$S_{\text{text}}\uparrow$	$S_{\text{img}}\downarrow$	RR↓
Normal			0.634	20.694	0.080	0.468	16.607	0.052	0.555	16.546	0.057	0.620	12.352	0.006
Mismatched	✓		0.637	13.300	0.085	0.455	12.212	0.051	0.552	10.274	0.043	0.622	9.394	0.011
			$S_{\text{text}}\uparrow$	$S_{\text{img}}\downarrow$	ASR↑	$S_{\text{text}}\uparrow$	$S_{\text{img}}\downarrow$	ASR↑	$S_{\text{text}}\uparrow$	$S_{\text{img}}\downarrow$	ASR↑	$S_{\text{text}}\uparrow$	$S_{\text{img}}\downarrow$	ASR↑
Malicious		✓	0.603	23.652	0.737	0.448	16.986	0.845	0.537	19.050	0.677	0.621	15.969	0.905
2M-attack	✓	✓	0.605	15.089	0.735	0.444	13.326	0.825	0.540	12.296	0.686	0.623	11.107	0.892
O2M-attack	✓	✓	**0.630**	**13.096**	**0.832**	0.295	18.484	**0.985**	0.488	13.659	**0.850**	**0.638**	12.137	0.895

Table 3: Performance scores for different models and inputs. The *mismatched* setting means the inputted text instruction and image are mismatched, such as modality and anatomy description in text is not consistent with the image. The *malicious* setting means the inputted text instruction is malicious, which may lead to harmful reply from MedMLLMs.

hancing defenses against harmful and mismatched inputs is essential to improve model performance, particularly for O2M-attack strategies. Improving model stability under normal inputs while reducing fluctuations under attack inputs is key. All models require stronger defenses against 2M and O2M attacks, with enhanced robustness and defense across various inputs being a critical direction for future optimization.

Method	**ASR**	**Text**	**Img**	**Emb**	**Opt**
PGD	0.707		✓		✓
GCG	0.812	✓			✓
FigStep	0.705	✓	✓		
Visual-RolePlay	0.784	✓			
IMAGE HIJACKS	0.775		✓		✓
CroPA	0.815			✓	✓
MCM (Ours)	**0.820**	✓	✓		✓

Table 4: Comparison of jailbreak methods: (PGD (Niu et al. 2024), GCG (Zou et al. 2023), FigStep (Gong et al. 2023), Visual-RolePlay (Ma et al. 2024), IMAGE HIJACKS (Bailey et al. 2023), CroPA (Luo et al. 2024)). Emb denotes embedding and Opt means optimize.

(3) MCM demonstrates significant advantages over other existing attack methods: Table 4 compares various multimodal jailbreak methods targeting the LLaVA-Med based on their Attack Success Rate (ASR), the modalities they modify (text, image, embedding), and whether they employ optimization techniques. Methods that simultaneously modify multiple modalities and use optimization techniques tend to achieve higher ASRs. For instance, 'Ours' and 'CroPA' methods show the highest success rates. The use of optimization techniques, regardless of the modality modified, generally improves the effectiveness of the attack. Different methods have varying strengths in terms of the specific modality they target, showing the importance of considering the attack context when selecting a jailbreak method.

5 Limitations

This study has several limitations: (1) insufficient task granularity, requiring more focus on specific lesions and details; (2) incomplete coverage of relevant research areas and clinical issues; and (3) limited discussion of defense strategies and technical implementations, as it mainly analyzes Medical MLLMs. Future research should address these gaps by broadening and deepening the analysis.

6 Conclusion

In this paper, we demonstrate that clinical mismatched phenomena and malicious queries can jailbreak MedMLLMs through our proposed optimized methods. We employ two methods for jailbreak: 2M-attack and O2M-attack. Moreover, we construct 3MAD dataset and use Llava-Med as a white-box attack to transfer it against four different MedMLLMs, exposing their security flaws and analyzing the current state of safety and semantic alignment within these systems. Additionally, we propose multi-dimensional evaluation metrics and a new effective attack method: MCM. Our research aims to underscore the need for strengthened safety measures within MedMLLMs used for clinical and medical diagnostics, advocating for secure and responsible development practices to ensure patient safety and contribute to the future of MedMLLM development.

Acknowledgments

This work was supported by the Fundamental Research Funds for Central Universities under Grant YWF-22-L-1281, the National Natural Science Foundation of China under Grants 62403038, and 62203032, the Beijing Natural Science Foundation under Grants JQ23019 and 4232046.

References

Bai, F.; Du, Y.; Huang, T.; Meng, M. Q.-H.; and Zhao, B. 2024. M3D: Advancing 3D Medical Image Analysis with Multi-Modal Large Language Models. *arXiv preprint arXiv:2404.00578*.

Bailey, L.; Ong, E.; Russell, S.; and Emmons, S. 2023. Image Hijacks: Adversarial Images Can Control Generative Models at Runtime. *arXiv preprint arXiv:2309.00236*.

Bakhshandeh, S. 2023. Benchmarking medical large language models. *Nature Reviews Bioengineering*, 1(8): 543–543.

Berner, E. S.; and Graber, M. L. 2008. Overconfidence as a cause of diagnostic error in medicine. *The American journal of medicine*, 121(5): S2–S23.

Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; and Liu, Z. 2024a. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. arXiv:2402.03216.

Chen, Z.; Varma, M.; Delbrouck, J.-B.; Paschali, M.; Blankemeier, L.; Van Veen, D.; Valanarasu, J. M. J.; Youssef, A.; Cohen, J. P.; Reis, E. P.; et al. 2024b. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. *arXiv preprint arXiv:2401.12208*.

Deng, C.; Zhao, Y.; Tang, X.; Gerstein, M.; and Cohan, A. 2023. Benchmark probing: Investigating data leakage in large language models. In *NeurIPS 2023 Workshop on Backdoors in Deep Learning-The Good, the Bad, and the Ugly*.

Gong, Y.; Ran, D.; Liu, J.; Wang, C.; Cong, T.; Wang, A.; Duan, S.; and Wang, X. 2023. FigStep: Jailbreaking Large Vision-language Models via Typographic Visual Prompts. arXiv:2311.05608.

Graber, M. L. 2013. The incidence of diagnostic error in medicine. *BMJ quality & safety*, 22(Suppl 2): ii21–ii27.

Lee, S.; Youn, J.; Kim, M.; and Yoon, S. H. 2023. CXR-LLaVA: Multimodal Large Language Model for Interpreting Chest X-ray Images. *arXiv preprint arXiv:2310.18341*.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Liu, F.; Zhu, T.; Wu, X.; Yang, B.; You, C.; Wang, C.; Lu, L.; Liu, Z.; Zheng, Y.; Sun, X.; et al. 2023a. A medical multimodal large language model for future pandemics. *NPJ Digital Medicine*, 6(1): 226.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, J.; Wang, Z.; Ye, Q.; Chong, D.; Zhou, P.; and Hua, Y. 2023c. Qilin-med-vl: Towards chinese large vision-language model for general healthcare. *arXiv preprint arXiv:2310.17956*.

Liu, X.; Zhu, Y.; Gu, J.; Lan, Y.; Yang, C.; and Qiao, Y. 2024. MM-SafetyBench: A Benchmark for Safety Evaluation of Multimodal Large Language Models. arXiv:2311.17600.

Luo, H.; Gu, J.; Liu, F.; and Torr, P. 2024. An image is worth 1000 lies: Adversarial transferability across prompts on vision-language models. *arXiv preprint arXiv:2403.09766*.

Ma, S.; Luo, W.; Wang, Y.; Liu, X.; Chen, M.; Li, B.; and Xiao, C. 2024. Visual-RolePlay: Universal Jailbreak Attack on MultiModal Large Language Models via Role-playing Image Characte. *arXiv preprint arXiv:2405.20773*.

Magar, I.; and Schwartz, R. 2022. Data Contamination: From Memorization to Exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 157–165.

Moor, M.; Banerjee, O.; Abad, Z. S. H.; Krumholz, H. M.; Leskovec, J.; Topol, E. J.; and Rajpurkar, P. 2023a. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956): 259–265.

Moor, M.; Huang, Q.; Wu, S.; Yasunaga, M.; Dalmia, Y.; Leskovec, J.; Zakka, C.; Reis, E. P.; and Rajpurkar, P. 2023b. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, 353–367. PMLR.

Niu, Z.; Ren, H.; Gao, X.; Hua, G.; and Jin, R. 2024. Jailbreaking attack against multimodal large language model. *arXiv preprint arXiv:2402.02309*.

Pi, R.; Han, T.; Xiong, W.; Zhang, J.; Liu, R.; Pan, R.; and Zhang, T. 2024. Strengthening multimodal large language model with bootstrapped preference optimization. *arXiv preprint arXiv:2403.08730*.

Qian, J.; Jin, Z.; Zhang, Q.; Cai, G.; and Liu, B. 2024. A Liver Cancer Question-Answering System Based on Next-Generation Intelligence and the Large Model Med-PaLM 2. *International Journal of Computer Science and Information Technology*, 2(1): 28–35.

Schiff, G. D.; Hasan, O.; Kim, S.; Abrams, R.; Cosby, K.; Lambert, B. L.; Elstein, A. S.; Hasler, S.; Kabongo, M. L.; Krosnjar, N.; et al. 2009. Diagnostic error in medicine: analysis of 583 physician-reported errors. *Archives of internal medicine*, 169(20): 1881–1887.

Shen, Y.; et al. 2023. Evaluating Object Hallucination in Large Vision-Language Models. *arXiv preprint arXiv:2305.10355*.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D.; et al. 2023b. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Thirunavukarasu, A. J.; Ting, D. S. J.; Elangovan, K.; Gutierrez, L.; Tan, T. F.; and Ting, D. S. W. 2023. Large language models in medicine. *Nature medicine*, 29(8): 1930–1940.

Tu, T.; Azizi, S.; Driess, D.; Schaekermann, M.; Amin, M.; Chang, P.-C.; Carroll, A.; Lau, C.; Tanno, R.; Ktena, I.; et al. 2024. Towards generalist biomedical ai. *NEJM AI*, 1(3): AIoa2300138.

Wang, R.; Duan, Y.; Li, J.; Pang, P.; and Tan, T. 2023. Xrayglm: The first chinese medical multimodal model that chest radiographs summarization.

Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*.

Yao, Y.; Duan, J.; Xu, K.; Cai, Y.; Sun, Z.; and Zhang, Y. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 100211.

Zhang, K.; Yu, J.; Yan, Z.; Liu, Y.; Adhikarla, E.; Fu, S.; Chen, X.; Chen, C.; Zhou, Y.; Li, X.; et al. 2023a. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*.

Zhang, S.; Xu, Y.; Usuyama, N.; Bagga, J.; Tinn, R.; Preston, S.; Rao, R.; Wei, M.; Valluri, N.; Wong, C.; Lungren, M.; Naumann, T.; and Poon, H. 2023b. Large-Scale Domain-Specific Pretraining for Biomedical Vision-Language Processing.

Zhang, X.; Wu, C.; Zhao, Z.; Lin, W.; Zhang, Y.; Wang, Y.; and Xie, W. 2023c. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Zhang, Y.; Huang, Y.; Sun, Y.; Liu, C.; Zhao, Z.; Fang, Z.; Wang, Y.; Chen, H.; Yang, X.; Wei, X.; et al. 2024. Benchmarking Trustworthiness of Multimodal Large Language Models: A Comprehensive Study. *arXiv preprint arXiv:2406.07057*.

Zou, A.; Wang, Z.; Kolter, J. Z.; and Fredrikson, M. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.