香港中文大學計算機科學與工程學系
Department of Computer Science and Engineering
The Chinese University of Hong Kong

The Chinese University of Hong Kong

Department of Computer Science and Engineering

Final Year Project

ESTR 4999 Graduation Thesis II

LYU2308 Evaluation of Multimodal Models:
Assessing Performance and Finding Improvements

Authors:
Haoran WU
Yushan WU

Supervised by:
Michael Rung Tsong Lyu

April 2024

Contents

1 Introduction

In the contemporary landscape of healthcare in the United States, medical errors are identified as the third leading cause of mortality, responsible for over 250,000 deaths each year[43]. Diagnostic inaccuracies stand out within this grim statistic, posing significant challenges to patient safety[22, 45]. The integration of computer systems with medical imaging has been a pivotal advancement, fostering the development of automated tools aimed at enhancing the accuracy of clinical diagnoses[39]. The advent and progression of Artificial Intelligence (AI) in medical imaging have marked a significant leap forward, refining these tools to meet industrial standards. The spectrum of techniques employed ranges from basic image processing to sophisticated neural networks, each contributing to the field's evolution[61, 49, 29, 3].

The immense potential within this domain has not gone unnoticed by major technology firms, propelling the AI healthcare market to a valuation of USD 16.3 billion in 2022, with a pronounced emphasis on diagnostic imaging products[14]. Pioneers such as IBM Watson Health and Google DeepMind have deployed AI-driven tools in leading hospitals, showcasing their efficacy in critical tasks like breast cancer screening, often surpassing human experts[44, 50].

However, the journey is not without its pitfalls. For instance, discrepancies between IBM Watson for Oncology's assessments and clinicians' decisions in gastric cancer cases underscore the limitations of current systems[32]. Given the stakes involved in medical diagnostics, ensuring the reliability of AI tools is of utmost importance. This has sparked a demand for robust testing frameworks, similar to those employed in soft-

ware and other AI applications, like autonomous vehicles[73]. Traditional testing approaches for computer vision software face challenges when adapted to the medical imaging diagnosis context, due to its inherent complexities and the critical nature of its applications[81, 47, 65, 26, 69].

The dearth of effective testing frameworks for medical imaging diagnosis software underscores the intricacy of this challenge. The creation of testing oracles necessitates a profound understanding of medical and clinical knowledge. Additionally, the prevalent image generation models, primarily trained on natural image datasets, are inadequate for producing realistic medical images essential for accurate testing.

The expansion of AI into medical diagnostics now includes multimodal models capable of interpreting diverse data types, offering a holistic analysis beyond single-modality capabilities. However, testing these advanced models demands a sophisticated approach that considers their complexity and the nuances of interpreting multimodal data.

This report introduces MedTest, an innovative metamorphic testing framework devised for the analysis of medical image diagnosis software. This framework encompasses both state-of-the-art academic models and extensive multimodal models. Through a pilot study involving over 2,500 images from three hospitals, we have identified nine metamorphic relations across four artifact categories: lightness, motion, object, and non-object. These relations have been incorporated into MedTest to create test cases reflecting real-world clinical scenarios, thereby ensuring the framework's relevance and effectiveness in evaluating medical image diagnosis applications.

Our application of MedTest to commercial and state-of-the-art diagnostic tools has revealed substantial variances in performance when confronting

original versus artifact-introduced images, highlighting areas for potential enhancement. Our ongoing research will delve into the challenges and prospects of multimodal models, with the aim of significantly contributing to AI's role in medical diagnostics.

The primary contributions of this paper are manifold:

- Introduction of MedTest, a novel and comprehensive testing framework tailored for validating medical image diagnosis software, marking a significant progression in the testing of medical imaging software.
- Execution of a pilot study with 2,553 real-world medical images, identifying 9 metamorphic relations crucial for MedTest's operationalization.
- Application of 9 distinct perturbation types across 5 datasets, totaling over 2,052 segmentation images and 3,022 classification images, resulting in the generation of 42,644 artifact-embedded images.
- Comprehensive evaluation of MedTest, demonstrating its utility across various commercial and academic state-of-the-art models. Our findings indicate that MedTest can reliably identify errors in these systems and substantially improve the robustness of leading algorithms, thereby advancing medical imaging technology.

2 Background

2.1 Medical Image Analysis

For decades, various medical imaging modalities such as computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), mammography, ultrasound, and X-ray have been instrumental in early disease detection, diagnosis, and management[6]. The task of interpreting these images has traditionally been reserved for hu-

man experts, like radiologists and physicians. Nonetheless, the subjective nature of interpretation and the risk of human fatigue have prompted the healthcare sector to lean towards computer-assisted analysis. Although computational analysis in medical imaging has lagged behind the technological advancements in imaging modalities, the recent surge in machine learning applications within this domain marks a significant leap forward.

The crux of machine learning in medical image analysis lies in the identification or development of features that accurately represent the underlying patterns in the data. Traditionally, these features were manually defined by experts, leveraging their specialized knowledge. This approach, however, presented barriers to non-experts, hindering the wider adoption of machine learning in medical research. Recent trends have shifted towards employing sparse representations, either through predefined dictionaries or those derived from training data. Rooted in the principle of parsimony, common across several scientific fields, this approach advocates for simpler explanations of phenomena. Techniques like sparsity-inducing penalization and dictionary learning have proven effective in feature representation and selection in medical image analysis, albeit with a reliance on relatively simplistic architectures for pattern identification[56, 71, 58, 12].

Contrastingly, deep learning transcends these constraints by embedding the feature engineering process within the algorithmic learning phase. This paradigm shift allows for autonomous identification of informative features with minimal human intervention, streamlining the feature extraction process[54]. By removing the dependency on manual feature design, deep learning enables broader engagement with advanced analytical techniques, fostering innovation and discovery in medical diagnostics

and treatment planning.

2.2 Metamorphic Testing

Metamorphic testing is a robust testing strategy that effectively tackles the oracle problem and has seen extensive application across different software engineering domains[10]. It is grounded in the identification and utilization of Metamorphic Relations (MRs), which define expected relationships between sequences of input-output pairs during software tests. By transforming an initial test case into a related one through a specific rule and comparing the outcomes, this method allows for the evaluation of software performance even when the exact output is unknown.

For example, in testing a software implementation of the $\sin x$ function, one might use the mathematical identity "$\sin(\pi - x) = \sin x$" as a metamorphic relation. This principle allows for the validation of the software's accuracy in computing sine values without needing the exact expected result, simply by comparing the sine of an angle and its supplementary angle[55].

In the realm of AI, metamorphic testing has gained traction for its ability to uncover errors in AI systems by creating and assessing innovative MRs. It has been applied in diverse areas, including bioinformatics, where Chen et al.[11] demonstrated its applicability, and in machine learning algorithms like k-Nearest Neighbors and Naive Bayes, with Xie et al.[75] establishing specific MRs for performance assessment. Further, Dwarakanath et al.[16] developed MRs for evaluating image classifiers based on SVM and ResNet architectures. Metamorphic testing has also been applied in autonomous driving software testing by Zhang et al.[81], who utilized GANs to generate varied driving scenarios for system eval-

uation. These developments underscore the expanding utility of metamorphic testing in validating the robustness and reliability of complex AI-driven software systems.

3 MedTest

In this section, we commence with an insightful pilot study, which delves into an analysis of authentic medical images that have been sourced directly from hospital environments (as detailed in Section 3.1). This preliminary exploration sets the stage for the subsequent introduction of nine metamorphic relations (MRs). These relations, derived and inspired by the findings of the pilot study, represent a significant step in understanding and evaluating medical image analysis processes.

We have meticulously categorized these nine MRs into four distinct groups, each based on the type of perturbation they involve. The first category focuses on lightness perturbations, where we examine how variations in image brightness and contrast can impact medical image analysis (discussed in Section 3.2). The second category, motion perturbations, explores the effects of simulated motion artifacts such as blurring, which can occur during image capture in dynamic clinical settings (covered in Section 3.3).

The third category revolves around object perturbations (Section 3.4), where the emphasis is on alterations related to the objects within the medical images. This includes changes in the size, shape, or position of clinically relevant features within the image. The final category, non-object perturbations (Section 3.5), addresses modifications that do not directly involve the primary objects of interest in the images. This could include alterations to background elements or other aspects of the image

that, while not directly related to the primary diagnostic features, may still influence the overall analysis process.

Each of these categories plays a pivotal role in understanding how different types of perturbations can affect the accuracy and reliability of medical image analysis, thereby contributing to the enhancement of diagnostic processes and tools in the healthcare sector. This structured approach allows for a comprehensive exploration of the complexities involved in medical image analysis and paves the way for developing more robust and reliable diagnostic methodologies.

3.1 Pilot Study

In our research, we set out with the ambitious goal of developing a set of MRs tailored to the field of medical imaging. These MRs are designed on the premise that a 'seed' test case (an original medical image) and its 'perturbed' counterpart (the same image but with added artifacts) should yield consistent classification labels or similar segmentation masks when analyzed by medical image analysis software. To ensure that these test cases are both effective and relevant, we have established a set of criteria for the perturbations incorporated in our MRs, which include:

- *Clinical-semantic-preserving*: This criterion ensures that the perturbed test cases should maintain the integrity of the analysis results, matching those of the original seed image.
- *Realistic*: The perturbations should closely mirror the types of artifacts encountered in actual clinical settings.
- *Unambiguous*: Clarity and precision in definition are key, ensuring that the perturbations are well-defined and easily interpretable.

To establish a foundation for designing these perturbations, we embarked

on a pilot study focusing on the types of artifacts typically encountered in medical images used in real-world clinical scenarios. This involved an extensive review of 103 endoscopic videos sourced from three hospitals. From these videos, we extracted $2,553$ individual images. We then engaged ten highly qualified annotators, each holding at least a postgraduate degree in medicine, to meticulously label these images. These annotators underwent thorough training, including guidelines, test tasks, and sessions specifically focused on endoscopic images and the identification of artifacts. During the annotation process, each image was evaluated to determine the presence of any artifact. The consensus among the annotators was used to establish the final human label, resulting in a dataset of $1,199$ endoscopic images identified as containing artifacts.

Upon detailed examination of these artifact-laden images, we identified and summarized 9 distinct methods of perturbation, commonly encountered in clinical settings. These methods are categorized from different perspectives: 1) those related to endoscopic imaging cameras, including lightness and motion perturbations; and 2) those pertaining to the visual content within the endoscopic images, such as object and non-object perturbations. Building on these insights, we formulated nine corresponding MRs, each based on a specific perturbation method. As shown in Table 1, we introduce 4 different perturbation groups, i.e. lightness, motion, objects, and non-objects, where each group includes at least one perturbation type. Fig. 1 demonstrates the visual perturbed images of different perturbation types. According to these MRs, the diagnostic label assigned by the medical analysis software to a perturbed endoscopic image (i.e., the generated test case) should align with the label given to the original, unperturbed seed image. Through this approach, we aim to rigorously test and validate the robustness and reliability of medical

Table 1: Categorization of Perturbation Types in Medical Images: A Pilot Study.

Perturbation Group	Type	Description
Lighting	Saturation	Over-saturation caused by excessive lighting
	Contrast	Resulting from underexposure or obstructions in the field of view
	White Balance	Color distortions due to presence of white objects
	Specularity	Reflections resembling a mirror-like surface
Motion	Blur	Blurring from hand movements or rapid camera motion
Objects	Instrument	Presence of surgical instruments in the image frame
	Feces	Incomplete colon cleansing in patients
	Blood	Visible bleeding from wounds
Non-objects	Text	Embedded clinical information related to patients

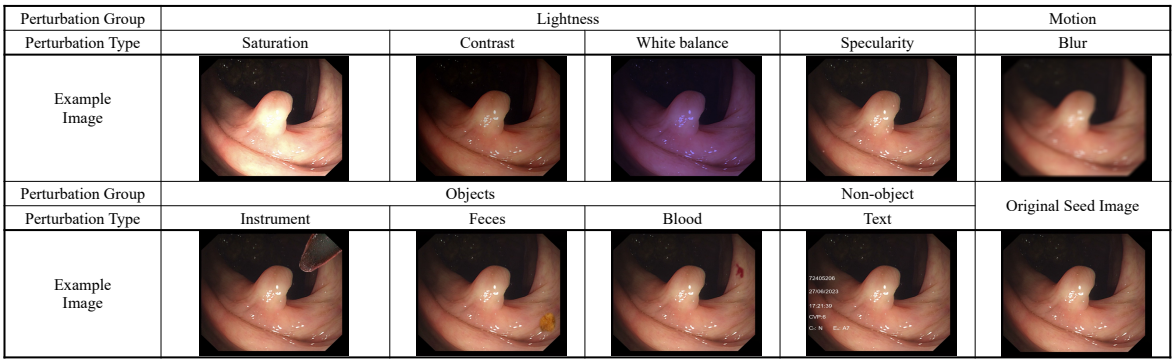


Figure 1: The visualization of the different perturbations groups.

image analysis software, ensuring its effectiveness even in the presence of common clinical artifacts.

3.2 MRs with Lightness Perturbations

These MRs leverage the lightness perturbations that imitate the various illumination conditions during the endoscopic camera imaging.

3.2.1 MR1-1 Saturation

To address saturation issues in endoscopic imaging, a key concern is the proximity of the light source to colon tissue. Overexposure can occur when the light source is too close, leading to saturation artifacts. Our method for simulating this effect involves applying variable levels of sat-

uration to endoscopic images to mimic different degrees of overexposure. This is achieved by adjusting the saturation of an image using a random factor selected from a predefined range $[s_1, s_2]$.

The process utilizes the torchvision library functions that control brightness, contrast, and saturation. We define a fluctuation range and randomly select a saturation factor within this range, with a bias towards values greater than 1 to replicate the overexposure effect. This factor is then used to modulate the saturation level, where a value of 1 indicates no change, values less than 1 decrease saturation, contrast, and brightness, and values greater than 1 increase them, thereby simulating the impact of light source proximity on the colon tissue.

3.2.2 MR1-2 Contrast

In the context of endoscopic examinations, the distance between the colon tissue and the light source, or obstructions, can result in underexposure. To simulate this scenario, our method focuses on altering the contrast of endoscopic images. Beginning with a seed endoscopic image, we establish a contrast range denoted as $[c_1, c_2]$. A value is then randomly selected from this range, which is used to adjust the image's contrast level.

This technique parallels the approach used for saturation adjustments, but with an inclination towards lower levels of contrast, brightness, and saturation, corresponding to the underexposed nature of the images. By carefully modulating the contrast in this manner, we aim to authentically replicate the conditions of underexposure commonly encountered during endoscopic procedures.

3.2.3 MR1-3 White Balance

In endoscopic imaging, we often observe color biases, predominantly manifesting as green or purple hues. The likely reason for these color biases could be attributed to the white balance settings of the endoscopic camera or the lighting conditions within the endoscopic environment, which may not always accurately represent the true colors of the tissue.

To simulate these white balance discrepancies in endoscopic images, we selectively modify the RGB channels. For images with a green bias, we reduce the red and blue channels by approximately half of their original values, maintaining their proportional relationship. Similarly, for images exhibiting a purple color bias, we decrease the values of both the red and green channels proportionately. This method allows us to realistically replicate the color distortions that might occur due to white balance issues in endoscopic imaging.

3.2.4 MR1-4 Specularity

The observed phenomena indicate that the manifestation of spots, attributable to specular reflection, predominantly occurs in a compact region as opposed to being dispersed throughout the entire image. Our initial approach involved identifying clusters as potential sites for spot generation. Subsequently, we introduced circle, ellipse, and distorted circle as the potential shape for generating white spots. Spots are generated at random locations near the cluster centers, with randomly chosen radius bounded by λ times image height in order to control the size of spots comparing with the image size. After trials and error, we found that the elliptical spots can achieve a best effect, creating most realistic specularity on the endoscopy images. The elliptical spots are decided by

the following formula:

$$\frac{(x-\bar{x})^2}{(a+\epsilon)^2}+\frac{(y-\bar{y})^2}{(b+\epsilon)^2}=1 \tag{1}$$

where $\bar{x}$ and $\bar{y}$ denote the center of the ellipse, and a and b are the two axes of the ellipse respectively. The additional ϵ acts as a term to avoid zero denominator due to the randomized selection of parameters. The following process involves the application of Gaussian blur to facilitate their seamless integration into the image. Additionally, we integrated these spots with a gray mask, derived from our algorithm, to modulate their intensity, particularly ensuring they do not exhibit excessive brightness in the darker regions of the image.

3.3 MRs with Motion Perturbations

3.3.1 MR2-1 Blur

We have noted that possible camera movement and tissue movement when the image is captured can often cause motion blur in images. To replicate this phenomenon, we employed Gaussian Blur, a technique that involves convolving each pixel of the image with a Gaussian function. The blurring degradation is defined as following:

$$x' = x \cdot G_B(r_B, \sigma_B) + n \tag{2}$$

where G_B is a Gaussian filter with a radius r_B and a spatial constant σ_B, and n is the random Gaussian noise added to the image. [57] In our implementation, we first generate a random number in the range of (5, 15] as the σ value. Based on the chosen σ, we randomly chose odd integers within the range of $\frac{\sigma}{2}$ to σ for the width and height of the kernel of the Gaussian filter, respectively. This process effectively blends each

pixel with the information from its neighboring pixels, creating a blurring effect reminiscent of a weighted average of the surrounding area. By using Gaussian Blur, we can simulate the kind of blur typically introduced by camera motion, enhancing the realism of our simulated images.

3.4 MRs with Object Perturbations

3.4.1 MR3-1 Instrument

In this study, we utilized the Kvasir-Instrument dataset [28], which comprises 590 images featuring medical instruments and their corresponding segmentation masks. Our initial step involved extracting these instruments from the original images and documenting their positions. Subsequently, we employed our algorithm to identify an optimal target area for each instrument, ensuring it met the following criteria:

- Avoidance of overlap with the Polyp.
- Preservation of a position and orientation akin to those in the original image.
- Maintenance of an appropriate size, neither excessively large nor small.

Finally, we repositioned the extracted instruments into these target areas. To enhance realism, we applied Gaussian blur and integrated our blending algorithm, ensuring a natural appearance of the instruments in their new context.

3.4.2 MR3-2 Feces

In this section, we employed fecal matter images extracted from the Kvasir dataset [30] using Meta's Segment Anything algorithm [34]. Similar to the aforementioned method used for instruments, we replicated

this approach, albeit without constraints on the positioning and orientation of the feces. Crucially, we calculated a brightness ratio by comparing the fecal matter with the target image, enabling us to adjust the feces' brightness for a more coherent integration. Furthermore, to prevent excessive brightness in particularly dark areas of the target image, we again utilized the gray mask previously mentioned in the context of specular reflections, providing an additional layer of realism to the adjusted fecal images.

3.4.3 MR3-3 Blood

In this phase, we focused on the blood images and their associated masks from the EAD2020 dataset [53, 2, 1]. Our methodology mirrored the approach previously described for pasting feces, with an emphasis on modifying various lighting parameters. This adjustment was crucial to enhance the natural appearance of the blood when integrated into the target images, ensuring a realistic representation in the context of the dataset.

3.5 MRs with Non-Object Perturbations

3.5.1 MR4-1 Text

Our analysis revealed a consistent pattern in the text displayed on endoscopic images, as illustrated in figure 2. Although the specific position and content of the text varied across images, it predominantly comprised temporal data and device parameters. To replicate this characteristic, we employed the ImageDraw method, generating text that adhered to the observed pattern through random generation, thereby maintaining consistency with the original text format in the images.

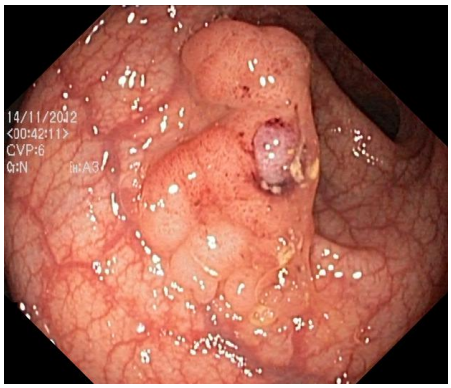


Figure 2: Pattern of the text in Kvasir dataset

4 Experimental Settings

4.1 Datasets

In our endeavor to thoroughly validate MedTest, we have utilized a diverse array of datasets as seed data, drawing upon the extensive work of previous researchers who have meticulously collected, labeled, and made available various types of data for research applications. Our evaluation spanned a variety of tasks, encompassing segmentation, Visual Question answering (VQA), and classification. For each of the task, we constructed a separate dataset for the tailored experiments. In the following sub-sections, we will discuss the tasks and the corresponding datasets separately.

4.1.1 Segmentation

For the purpose of evaluating medical diagnostic systems in the segmentation task, we have specifically chosen the most widely used datasets in the field of polyp segmentation, all of which are publicly accessible. These include CVC-300 [68], CVC-ClinicDB [5], CVC-ColonDB [60], and Kvasir [30]. There are in total 2052 images combined.

CVC-300, a subset of the larger EndoScene dataset, is a relatively compact dataset comprising 60 images, each with dimensions of 578×500 pixels. In addition to CVC-300, the EndoScene dataset also encompasses images from the CVC-ClinicDB dataset. To maintain clarity and precision in our analysis, we have treated these two datasets as distinct entities in separate experiments, meticulously recording and analyzing their respective results [38].

CVC-ClinicDB, also known as CVC-612, is a more extensive collection, featuring 612 publicly available polyp images sourced from 25 different colonoscopy videos. The images in this dataset are of the size 384×288 pixels, offering a distinct set of characteristics for analysis.

The CVC-ColonDB dataset is composed of 15 different endoscopy sequences, totaling 380 polyp images. Each image in this dataset shares the same resolution as the CVC-300 dataset, specifically 578×500 pixels.

Lastly, the Kvasir dataset, a more recent addition to the field, stands out due to its large scale, diverse endoscopy scenes, and varied polyp shapes. This diversity renders the segmentation task particularly challenging. The images in Kvasir vary considerably in size, ranging from 332×487 to 1920×1072 pixels. This variability not only presents a significant challenge for medical diagnosis software but also adds complexity to our method of generating simulated artifacts [30] [38]. The selection of these datasets for our validation process reflects our commitment to ensuring that MedTest is rigorously tested against a wide spectrum of real-world scenarios, thereby ensuring its robustness and applicability in diverse clinical settings.

4.1.2 Visual Question Answering

To assess the capabilities of multimodal large language models (MLLMs), we utilized the ImageCLEF MEDVQA Dataset[27]. This dataset, derived from the Hyper Kvasir dataset[30], includes 182 images that were also part of our segmentation dataset to maintain consistency in testing. For each image, participants answered 18 specific questions, as detailed in Table 2.

Question Number	**Question**
1	Are there any abnormalities in the image?
2	Are there any anatomical landmarks in the image?
3	Are there any instruments in the image?
4	Have all polyps been removed?
5	How many findings are present?
6	How many instruments are in the image?
7	How many polyps are in the image?
8	Is there a green/black box artefact?
9	Is there text?
10	Is this finding easy to detect?
11	What color is the abnormality?
12	What color is the anatomical landmark?
13	What is the size of the polyp?
14	What type of polyp is present?
15	What type of procedure is the image taken from?
16	Where in the image is the abnormality?
17	Where in the image is the anatomical landmark?
18	Where in the image is the instrument?

Table 2: Summary of Questions in VQA Experiment

However, questions 2, 10, 12, and 17 were either irrelevant to our objectives or too vague to provide meaningful insights, leading to their exclusion from our analysis. Moreover, in scenarios where we manually added text or instruments to an image, we adjusted the ground truth responses for certain questions accordingly: for the question **Is there text?**, the answer was changed to **Yes**; for **Are there any instruments in the image?**, it was also adjusted to **Yes**; and for **How many instruments are in the image?**, the answer was amended to $n+1$, where n represents

the initial count. Due to the random placement of instrument artifacts by MedTest, the responses to the question **Where in the image is the instrument?** were deemed unreliable.

4.1.3 Classification

Wireless Capsule Endoscopy (WCE) is a non-invasive medical imaging device that gastroenterologists use to investigate gastrointestinal tract disorders, which also constitutes an important scope of medical images that collect information from organs or other structures inside the human body. As for the classification task, we selected the commonly-used datasets in previous implementation of classification models on wireless capsule endoscopy images based on our literature review on previous studies. CAD-CAP [15] and KID [35] are two datasets that constitute the training and testing data of all the chosen classification methods for evaluation.

In order to keep consistency, we follow the design in the implementation of AGDN model [76] to split the data and construct the training and testing subsets.

All the involved methods for evaluation focus on the classification of three different types of images, namely normal images, vascular lesion frames and inflammatory frames, where the latter two are considered to reflect the sub-healthy status of human body. The fused dataset consists of 3022 images (1812 CAD-CAP + 1210 KID), including 1300 normal images (600 CAD-CAP + 700 KID), 888 vascular lesions (605 CAD-CAP + 283 KID), and 834 inflammatory frames (607 CAD-CAP + 227 KID). The original resolution of CAD-CAP and KID are 576×576 and 360×360 with black margins over 32 and 20 pixels separately. Therefore,

we resized all the images to 512×512 for consistency.

In the experiment design of AGDN, the whole fused dataset were randomly split for four times for effectiveness and robustness. Based on this, we chose one of the split as our dataset, which includes 600 images (200 images for each class) as the testing set and the remaining 2422 images as the training set. [76]

4.1.4 Pre-processing

To ensure a uniform approach in our analysis, we initially undertook the task of standardizing the dataset. This involved pre-processing both the images and their corresponding segmentation masks to a consistent size of 512×512 pixels, a dimension commonly accepted and utilized by various medical diagnosis algorithms. This standardization is crucial for maintaining consistency across different datasets and ensuring that the input to the medical diagnosis algorithms is uniform, thus allowing for more accurate comparisons and evaluations.

A notable characteristic of most endoscopy images is the presence of a black frame around the edges, which typically lacks a consistent pattern. This irregularity renders traditional image processing techniques, such as thresholding and region growing, ineffective for their extraction. To address this challenge, we developed a specialized model specifically designed to extract these black frames from the images. This extraction is vital, as it enables us to mask out any potential synthesized artifacts that may appear on these black edges, thereby ensuring the integrity and realism of the images used in our tests.

In addition to frame extraction, our pre-processing model plays a crucial role in assessing the brightness levels across different areas within the

images. By employing a sigmoid function, we generate gray masks that reflect these brightness variations. These gray masks are then strategically used in the process of artifact addition. They allow for precise adjustments in color and brightness of the synthesized components, effectively preventing the creation of images with overtly unnatural or artificial effects. This meticulous approach to pre-processing and artifact integration is fundamental to our objective of producing realistic test cases that accurately mimic real-world clinical scenarios. It ensures that our testing environment closely replicates the conditions under which medical diagnosis software is typically employed, thus providing a robust and reliable framework for evaluating the performance and efficacy of these algorithms.

4.2 Software and Models Under Test

We use MedTest to test commercial medical image diagnosis software products and SOTA academic models. Commercial software products include ChatGPT[46] and Gemini[63], on which we want to test the performance on the VQA task on polyp-related diagnosis given endoscopic images as the input. SOTA academic models can be divided into three parts centering segmentation, classification and VQA task. For the segmentation task, our tested objects consist of PraNet [17], SANet [74], TGANet [66] and SSFormer [70], all targeting the polyp segmentation task on endoscopy images. Regarding the classification task, we found relevant implementations mainly on categorizing WCE images on normal, vascular lesion and inflammatory type. Therefore, we carefully investigated previous studies and select DSI-Net [82] and AGDN [76] as the models to be tested on. As for the VQA task, we tested on the two well-known MLLMs, GPT-4V and Gemini, which has the ability to respond

to both the text and image data.

4.2.1 Segmentation

PraNet

PraNet model is a well-established model on polyp segmentation task and paved the path for the later ones. The special design on PraNet is that it uniquely combines high-level feature aggregation via a Parallel Partial Decoder (PPD) and detailed segmentation through Reverse Attention (RA) modules. This approach can enable the model to effectively handle the variability in polyps' size, color, and texture, as well as the often-blurred boundaries in colonoscopy images. PraNet has demonstrated excessive performance over existing methods in terms of segmentation accuracy, generalizability, and real-time efficiency. Because it is one of the initial influential model, we decided to investigate deep into it and evaluate its robustness.

SANet

Besides, we explored the Shallow Attention Network (SANet), also designed to address key challenges in polyp segmentation but with more unique designs and implementation to handle detailed problems in previous studies. SANet innovatively tackles issues like inconsistent color distributions in samples, degradation of small polyps due to repeated downsampling, and imbalance between foreground and background pixels. Based on this idea, the model employs a color exchange operation to reduce overfitting by decoupling image content from color, enhancing focus on shape and structure. It also introduces a shallow attention module to filter background noise in shallow features, which helps preserve small polyps more effectively. Additionally, the probability correction strategy

during inference improves model performance, especially for small polyps. SANet's extensive testing across five benchmarks shows its outstanding capability in polyp segmentation task, suitable for us to evaluate.

TGANet

We also investigated the TGANet model, which focuses on enhanced polyp segmentation in colonoscopy images using the auxiliary text input as additional information. The model aimed at the challenges posed by the variability in polyp size and number, which can impact the effectiveness of segmentation models. Targeting this, TGANet innovatively employs text-guided attention mechanisms, leveraging attributes like polyp size and count through additional text input to improve segmentation accuracy. The text input containing the polyp information serves as an auxiliary classification task and further enhance model's learned representations of important features within the image. After that, a feature enhancement module and multi-scale feature aggregation within the network are present to allow for more precise adaptation to varying polyp characteristics. With these implementation, especially the module to incorporate text description information, the model is expected to have better performance because of the excessive learning of image feature representations. As is describe in their inference part, text input is unnecessary, so we leverage this novel design to test our evaluation framework MedTest.

SSFormer

We also explored the SSFormer model, which also targeted the challenges imposed by the complex and diverse structure of polyps image and the varying shapes of poly. These problems, together with the indistinctive bound between polyp and other categories, make the whole segmentation

task difficult and the learning on existing dataset prone to over-fitting. This model stands out by incorporating a pyramid Transformer encoder, significantly enhancing the model's generalization capabilities. The Progressive Locality Decoder (PLD) in it emphasizes local features while integrating them into global features. This can effectively address the common issue of attention dispersion in Transformer models. Such delicate design improves the detail processing ability of the neural network and allows the establishment of its SOTA performance in polyp segmentation tasks. Because this model demonstrates exceptional learning and generalization abilities on unseen datasets, we want to test whether its performance is robust enough on our MedTest.

4.2.2 Visual Question Answering

GPT-4V

Introduced by OpenAI, GPT-4V represents a significant advancement in the field of large language models (LLMs) with integrated vision capabilities. This model demonstrates a profound understanding of visual content, enabling it to perform VQA tasks across a diverse spectrum. Although OpenAI has cautioned that GPT-4V's reliability in medical diagnosis is limited, the model's potential application in healthcare and related fields is noteworthy. Specific details regarding the model's architecture and training methodologies remain proprietary and have not been disclosed to the public.

Gemini

Developed by Google, the Gemini model has been reported to surpass the capabilities of GPT-4V in several key areas, including but not limited to performance metrics and task-specific benchmarks. Notably, Gemini

distinguishes itself through its exceptional multimodal capabilities, with a particular emphasis on visual processing and interpretation. This enhancement in visual modality allows Gemini to engage with and analyze image-based content with remarkable efficiency and accuracy.

4.2.3 Classification

AGDN

The Attention Guided Deformation Network (AGDN) for Wireless Capsule Endoscopy (WCE) image classification is a pioneering approach designed to address the inherent challenges of identifying small and often indistinct lesions within the gastrointestinal tract. By innovatively employing a two-branch architecture that utilizes attention maps to guide the deformation of input images, AGDN achieves precise amplification of lesion regions, which significantly enhances lesion visibility and classification accuracy. This network introduces the Third-order Long-range Feature Aggregation (TLFA) modules and the Deformation based Attention Consistency (DAC) loss, which together capture long-range dependencies, aggregate contextual features, and refine attention maps for improved diagnostic performance. Through comparing experiments, the study demonstrated to outperform existing models on public WCE datasets. AGDN represents a substantial advancement in automated endoscopic analysis, offering new avenues for accurate and efficient gastrointestinal disease diagnosis. Its innovative mechanism of focusing on lesion-specific features while minimizing background interference positions AGDN as a key model within our evaluation framework, underscoring its potential to transform clinical practices through enhanced diagnostic precision.

DSI-Net

Moreover, we explored the Deep Synergistic Interaction Network (DSI-Net), which introduces a novel approach to the joint classification and segmentation of endoscopic images, specifically targeting the challenges in gastrointestinal tract disease diagnosis. Uniquely integrating a classification branch, a coarse segmentation branch, and a fine segmentation branch, DSI-Net leverages deep synergistic interactions between these tasks to significantly enhance performance. Central to its innovation are the Lesion Location Mining module and the Category-Guided Feature Generation (CFG) module. The Lesion Location Mining module refines attention on lesion regions by accurately identifying neglected lesion areas and eliminating misclassified background areas, which aids in precise classification. On the other hand, the CFG module utilizes category prototypes from the classification branch to generate category-aware features, thereby improving segmentation accuracy. Furthermore, DSI-Net incorporates a task interaction loss to ensure consistency between classification and segmentation tasks, which helps to enhance the mutual guidance and improving the overall diagnostic capabilities. Demonstrating superior performance over state-of-the-art methods on public datasets, DSI-Net marks a significant advancement in computer-aided diagnosis systems for gastrointestinal diseases. Its approach to leveraging complementary information across different tasks not only improves diagnostic accuracy but also highlights the potential for synergistic interactions in medical image analysis. Considering its novel contributions and exceptional performance, DSI-Net was a prime candidate for our evaluation framework, aiming to assess its robustness and applicability in real-world diagnostic scenarios.

5 Evaluation: Research Questions

To rigorously assess the efficacy of MedTest, our methodology has been applied to SOTA algorithms in various medical image diagnosis tasks focusing on endoscopy and colonscopy image inputs. Relevant tasks mainly consist of segmentation and classification, where the segmentation task involves segmenting polyps in images and the classification task is dedicated to categorizing medical images into normal and disease-detected ones. Additionally, we extended this evaluation to include several commercial software products derived from MLLMs on the VQA task, including GPT-4V and Gemini.

Based on our design of experiments, this section is dedicated to exploring and providing insights into four critical Research Questions (RQs), which are as follows:

- RQ1: Does MedTest generate test cases that are diagnostically consistent with the original seed images and maintain a realistic appearance?
- RQ2: Is MedTest effective in identifying incorrect outputs produced by medical image diagnosis software and algorithms?
- RQ3: Can the test cases generated by MedTest be utilized to enhance the performance of medical image diagnosis software?
- RQ4: What are the various factors that influence MedTest's performance and how do they do so?

The prerequisite of identifying the effectiveness of our method MedTest is to ensure that the result is convincing and reliable. Therefore, we need to guarantee that our test cases after perturbations are clinically equivalent to the original seed images as well as reflective to realistic

situations and artifacts present in actual clinical scenarios. Therefore, we designed RQ1 mainly aiming to validate whether the perturbations introduced in the test cases preserve the clinical diagnosis and realism, as assessed by human annotation.

The key point in the evaluation is to determine whether our method MedTest possess the capability to attack and incur hidden errors in our selected medical image diagnosis systems. Hence, in RQ2, we discussed in detail the approaches we conducted to reveal the robustness and consistency when models encounter corner cases constructed by our perturbations, together with the description of the chosen tasks and models.

Furthermore, the discovery of errors naturally leads to their rectification. Intuitively, we hope to see our method MedTest can be further utilized in refining the performance of our evaluated software. Therefore, RQ3 is dedicated to discussing how the generated perturbed dataset can be leveraged to improve the performance of how the medical image diagnosis tools react when they are faced with potentially error-triggering corner cases.

Besides previous concerns, RQ4 is utilized to illustrate the potential influence of different factors present in the overall experiment process on the performance of MedTest, given the fact that our method is relatively a pioneering approach in evaluating the diagnostic systems in the interdisciplinary field of biomedicine. This sub-sections aims to draw a closer look into how various elements in both the perturbation generation and experiment setting can affect the efficacy of MedTest in showcasing its ability in the medical diagnosis model evaluation.

To sum up, we hope to investigate deeper into the proposed research questions above and provide a comprehensive insight into our method

and whole evaluation process through answering to the questions. In the effort, we can further ensure that we not only validate the effectiveness of MedTest but also to contribute significantly to the advancement of medical image diagnosis technology.

6 RQ1: Are the test cases generated by MedTest diagnosis-identical to seed images and realistic?

In this study, MedTest is designed with the specific objective of creating test cases that not only yield identical diagnostic results compared to their corresponding seed images but also closely resemble the types of artifacts encountered by medical professionals in real-world clinical settings. To assess the effectiveness of MedTest in achieving these goals, we designed an evaluation based on human annotations.

First of all, we conducted thorough literature review on related medical studies and summarized the observed patterns inside the medical images. Based on the results, we generated a sample set containing all the perturbation method. These images were then evaluated by professional expert in medicine through surveys. Upon our survey result, we can conclude that images generated from MedTest can be regarded as realistic and clinically equivalent to real-world scenarios. It underscores the high degree of fidelity and realism that MedTest achieves in simulating clinical artifacts, as well as its effectiveness in maintaining diagnostic consistency. This built up the foundation of our further experiments on applying our method MedTest to the actual evaluation of medical diagnosis software and models, ensuring its readiness for practical application in clinical environments.

Answer to RQ1: The test cases generated by MedTest are diagnosis-identical to seed images and realistic.

7 RQ2: Can MedTest find erroneous outputs returned by medical image diagnosis software?

MedTest aims to automatically generate test cases to find potential errors in current medical image diagnosis software. Hence, in this section, we evaluate the number of errors that MedTest can find in the outputs of commercial software and academic models.

We first input all the original seed images and obtain the original output for each software product or model under test. Then we conduct perturbations in MedTest's MRs described in section 3 on the seed image to generate test cases. Finally, we use the generated test cases to validate the software products and academic models.

In particular, we check whether the generated test cases have identical diagnosis results as the corresponding seed images. If not, the diagnosis-identical perturbation affects the diagnosis of the software products or academic models, indicating erroneous outputs.

To evaluate how well MedTest does on generating test cases that trigger errors, we calculate Error Finding Rate (EFR), which is defined as follows:

$$\text{EFR} = \frac{\text{Number of misclassified test cases}}{\text{Number of generated test cases}} * 100\%. \tag{3}$$

7.1 Segmentation

Segmentation task on medical images is a computer vision task that involves dividing an medical image into multiple segments, where different segments represent separate objects or structures of interest. Medical image segmentation aims to provide an accurate representation of diverse information present in the image, which can be leveraged for further diagnosis and quantitative analysis. Therefore, the segmentation task contributes an important division in medical diagnosis and ensuring the robustness of SOTA methods in this area is of great significance.

7.1.1 Evaluation Criteria

In our testing on polyp segmentation task, we apply two similarity coefficients, Dice score and IoU score, which are proved to be simple and useful summary measures of spatial overlap and can measure the accuracy in image segmentation [84]. A test case is considered misclassified when its scores, both Dice and IoU, are 50% less than the scores tested on the seed image. The Dice score is given by

$$\begin{aligned} \text{Dice}(\hat{Y}, Y) &= \frac{2 \times |\hat{Y} \cap Y|}{|\hat{Y}| + |Y|} \\ &= \frac{2 \times TP}{(TP + FP) + (TP + FN)}. \end{aligned} \tag{4}$$

And the IoU score is given by

$$\begin{aligned} \text{IoU}(\hat{Y}, Y) &= \frac{|\hat{Y} \cap Y|}{|\hat{Y} \cup Y|} \\ &= \frac{TP}{TP + FP + FN}. \end{aligned} \tag{5}$$

In both equations 4 and 5, the $\hat{Y}$ stands for the predicted segmentation mask output by the models, while Y is the ground truth segmentation mask. Here, the TP, FP, and FN are all calculated pixel-wise on the masks.

Based on these two similarity coefficients, we define a segmentation output case as "missclassified" or "error" when difference between the performance of the seed image and that of the synthesize image to be larger than a proportion of the performance of the seed image. We set the proportion to be a threshold t. Our definition is therefore by the following:

$$\frac{OriginalScore - ArtifactScore}{OriginalScore} > t \tag{6}$$

where $OriginalScore$ represents the Dice/IoU score calculated from the seed image and $ArtifactScore$ represents the Dice/IoU score calculated from the synthesized image with specific artifact. In the table showed in later sections, we recorded the statistics for choosing both 50% and 25%.

Regarding the EFR, we found that it can be analyzed in different dimension, regarding different artifacts, datasets and models. The EFR varies in different experiment settings and we will use the following sections to illustrate the influence of above factors, with our proposed explanations for the situations. The EFR and performance with regard to different models will be discussed in the following section, while the varying situation of EFR concerning different datasets and artifacts will be further illustrated in later section 10.

7.1.2 Model Performance

Though the models selected for evaluation are targeting the same task, i.e. polyp segmentation, and there are all accepted by top conferences

or have high citations, vary in their emphasis on different components in the network designs. Therefore, the difference between their robustness should be discussed.

PraNet	CVC-300		CVC-ClinicDB		CVC-ColonDB		Kvasir	
t=0.5	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
Blood	0.0	0.0	0.0	0.0	1.1	1.3	0.1	0.3
Feces	0.0	0.0	0.0	0.0	3.7	5.0	0.0	0.2
Instrument	**3.3**	5.0	1.1	2.6	7.9	9.8	0.1	0.3
Spot	0.0	0.0	0.2	0.2	1.8	1.8	0.1	0.1
Saturation	**3.3**	**6.7**	0.7	0.7	4.2	4.5	2.3	3.7
Contrast	0.0	0.0	0.3	0.3	4.0	4.2	0.3	0.6
White Balance	**3.3**	3.3	**7.4**	**10.9**	**15.0**	**16.9**	2.9	4.9
Blur	1.7	1.7	5.2	8.0	7.9	11.3	**7.4**	**11.7**
Text	0.0	0.0	0.5	0.5	2.4	2.6	0.0	0.1

Table 3: EFR(%) of PraNet Model on Various Datasets with $t = 0.5$

SANet	CVC-300		CVC-ClinicDB		CVC-ColonDB		Kvasir	
t=0.5	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
Blood	0.0	0.0	0.0	0.0	1.6	1.8	0.0	0.0
Feces	0.0	1.7	0.0	0.2	5.0	5.5	0.1	0.1
Instrument	0.0	0.0	0.0	0.0	3.4	3.7	0.0	0.0
Spot	0.0	0.0	0.2	0.2	2.1	2.6	0.0	0.0
Saturation	**3.3**	**3.3**	0.2	0.5	2.4	2.4	0.9	1.3
Contrast	0.0	0.0	0.0	0.0	2.4	2.4	0.0	0.0
White Balance	0.0	0.0	**1.5**	**2.3**	**8.2**	**9.0**	**1.0**	**3.4**
Blur	0.0	0.0	0.2	0.2	2.6	3.4	0.3	0.6
Text	0.0	0.0	0.2	0.2	4.0	4.0	0.0	0.0

Table 4: EFR(%) of SANet Model on Various Datasets with $t = 0.5$

PraNet

PraNet, renowned in the domain of Polyp segmentation, exhibits commendable performance on the original dataset. Its Error Finding Rate (EFR) on the Dice score, with a threshold of 0.25, stands at 4.38%, indicative of its relative robustness. However, the inclusion of CVC-ClinicDB and Kvasir datasets in PraNet's training set may predispose the results to bias. A more critical examination using the CVC-ColonDB

TGANet	CVC-300		CVC-ClinicDB		CVC-ColonDB		Kvasir	
t=0.5	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
Blood	10.0	10.0	9.2	11.3	12.9	15.3	6.5	7.8
Feces	1.7	8.3	1.8	2.5	5.8	7.4	1.1	1.4
Instrument	11.7	16.7	2.6	4.2	8.2	10.8	1.5	1.8
Spot	3.3	3.3	0.2	0.2	3.2	3.4	0.1	0.3
Saturation	8.3	10.0	9.3	15.2	10.8	15.0	22.4	30.4
Contrast	0.0	0.0	6.7	8.7	15.3	17.6	5.5	7.3
White Balance	**23.3**	**26.7**	**30.2**	**37.7**	**21.3**	**40.8**	**25.4**	**32.6**
Blur	15.0	21.7	4.7	6.5	6.6	7.1	5.8	7.5
Text	5.0	6.7	1.1	1.8	4.2	5.0	0.6	1.0

Table 5: EFR(%) of TGANet Model on Various Datasets with $t = 0.5$

SSFormer	CVC-300		CVC-ClinicDB		CVC-ColonDB		Kvasir	
t=0.5	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
Blood	0.0	1.7	0.2	0.2	3.2	3.9	0.0	0.0
Feces	0.0	0.0	0.0	0.0	4.7	5.8	0.0	0.0
Instrument	0.0	1.7	0.5	1.1	5.8	5.8	0.0	0.0
Spot	0.0	0.0	0.2	0.2	1.6	1.8	0.0	0.0
Saturation	**6.7**	**6.7**	0.8	1.8	1.6	2.1	0.2	0.4
Contrast	0.0	0.0	0.2	0.2	3.4	3.4	0.1	0.2
White Balance	0.0	1.7	**2.5**	**3.9**	**9.5**	**10.5**	**0.9**	**1.5**
Blur	0.0	0.0	0.2	0.2	2.4	2.6	0.1	0.2
Text	0.0	0.0	0.2	0.2	0.8	1.6	0.0	0.0

Table 6: EFR(%) of SSFormer Model on Various Datasets with $t = 0.5$

dataset reveals a heightened EFR of 8.56%. Focusing on the CVC-ColonDB analysis, PraNet demonstrates proficiency in handling Specularity and Blood artifacts, but shows vulnerability to White Balance and Blur. This suggests a higher resilience to object-based distortions as opposed to those induced by lighting variations. See Table 3 7 for the specific results.

SANet

SANet, introduced a year subsequent to PraNet, is evaluated for its enhanced robustness. The EFR of SANet on the Dice score, with the threshold set at 0.25, is recorded at 1.7%. To mitigate the potential bias

PraNet	CVC-300		CVC-ClinicDB		CVC-ColonDB		Kvasir	
t=0.25	**Dice**	**IoU**	**Dice**	**IoU**	**Dice**	**IoU**	**Dice**	**IoU**
Blood	3.3	6.7	0.5	1.0	4.0	5.0	0.5	0.6
Feces	0.0	1.7	0.8	2.0	7.4	9.2	0.5	1.5
Instrument	6.7	11.7	4.1	5.6	12.1	14.0	0.4	1.1
Spot	1.7	1.7	0.5	0.5	3.2	4.2	0.1	0.5
Saturation	**8.3**	**13.3**	1.6	3.4	6.6	8.4	5.8	9.6
Contrast	1.7	5.0	0.3	0.8	4.7	6.1	1.3	2.2
White Balance	**8.3**	**13.3**	**12.7**	**18.0**	**19.8**	**22.7**	7.5	12.3
Blur	**8.3**	8.3	9.6	13.6	14.2	17.2	**14.2**	**18.8**
Text	0.0	0.0	0.7	0.8	5.0	5.8	0.2	0.3

Table 7: EFR(%) of PraNet Model on Various Datasets with $t = 0.25$

SANet	CVC-300		CVC-ClinicDB		CVC-ColonDB		Kvasir	
t=0.25	**Dice**	**IoU**	**Dice**	**IoU**	**Dice**	**IoU**	**Dice**	**IoU**
Blood	0.0	0.0	0.0	0.0	2.9	3.4	0.1	0.1
Feces	1.7	1.7	0.3	0.5	6.9	7.4	0.1	0.2
Instrument	1.7	1.7	0.2	0.7	5.5	5.8	0.0	0.0
Spot	0.0	0.0	0.2	0.3	4.2	4.5	0.0	0.0
Saturation	**5.0**	**6.7**	1.0	1.8	3.4	5.5	1.7	3.1
Contrast	0.0	0.0	0.2	0.2	3.4	4.0	0.0	0.2
White Balance	0.0	0.0	**3.4**	**6.2**	**10.8**	**14.0**	**5.4**	**9.2**
Blur	3.3	5.0	0.3	0.5	6.3	8.7	1.1	2.0
Text	0.0	0.0	0.5	1.0	5.5	5.8	0.0	0.1

Table 8: EFR(%) of SANet Model on Various Datasets with $t = 0.25$

from images in the training set, we scrutinized its performance on the CVC-ColonDB, where the EFR is noted to be 5.43%. SANet exhibits a markedly reduced EFR across most artifact categories in the CVC-ColonDB. It is predominantly impacted by White Balance and Blur, while demonstrating greater resistance to artifacts related to Blood, Saturation, and Contrast. See Table 4 8 for the specific results.

TGANet

TGANet has a special design of incorporating the text embedding to provide additional information to enhance feature representations. We found that its performance is unsatisfying on both the original seed im-

TGANet	CVC-300		CVC-ClinicDB		CVC-ColonDB		Kvasir	
t=0.25	**Dice**	**IoU**	**Dice**	**IoU**	**Dice**	**IoU**	**Dice**	**IoU**
Blood	16.7	20.0	15.8	22.1	23.9	29.2	12.9	15.7
Feces	13.3	25.0	4.4	7.0	13.9	18.2	2.7	3.7
Instrument	30.0	**46.7**	9.2	14.9	18.9	24.2	4.4	6.9
Spot	3.3	3.3	1.5	2.1	5.5	6.6	0.8	1.0
Saturation	16.7	18.3	21.2	28.9	21.8	24.7	**46.1**	**53.7**
Contrast	0.0	1.7	12.9	17.3	26.8	29.2	14.3	18.0
White Balance	**31.7**	38.3	**47.5**	**59.5**	**35.3**	**40.8**	43.0	49.8
Blur	28.3	31.7	4.7	6.5	9.7	11.8	15.3	18.3
Text	8.3	8.3	3.9	5.1	10.8	13.4	3.9	5.6

Table 9: EFR(%) of TGANet Model on Various Datasets with $t = 0.25$

SSFormer	CVC-300		CVC-ClinicDB		CVC-ColonDB		Kvasir	
t=0.25	**Dice**	**IoU**	**Dice**	**IoU**	**Dice**	**IoU**	**Dice**	**IoU**
Blood	3.3	3.3	0.2	0.2	5.0	5.3	0.1	0.1
Feces	0.0	0.0	0.3	0.5	7.6	8.2	0.0	0.0
Instrument	3.3	6.7	1.8	2.5	7.1	7.6	0.0	0.0
Spot	0.0	0.0	0.3	0.3	2.4	2.4	0.0	0.0
Saturation	**6.7**	**10.0**	1.0	1.3	2.6	4.5	0.4	0.8
Contrast	1.7	3.3	0.2	0.2	3.9	4.7	0.3	0.5
White Balance	3.3	5.0	**4.7**	**7.5**	**11.8**	**13.9**	**2.0**	**4.0**
Blur	0.0	1.7	0.2	0.2	3.4	3.4	0.3	0.4
Text	0.0	0.0	0.3	0.3	2.1	2.6	0.0	0.1

Table 10: EFR(%) of SSFormer Model on Various Datasets with $t = 0.25$

ages and the synthesized ones, and our synthesized image inputs have triggered even more errors. When threshold t is set to 0.25, the network produced an EFR up to 15.70% on the overall 4 datasets. Among all artifacts, EFR was much higher on White Balance, Instrument, and Blur. Contrast and Blood artifacts also sometimes caused severe problems on specific datasets. Our conjecture on the explanation is that the model were trained using insufficient data and the presence of auxiliary text inputs may result in the overfitting problem when learning feature representations. See Table 5 9 for the specific results.

SSFormer

SSFormer is the most robust models when tested using the synthesized images. When we relax the threshold to $t = 0.25$, the EFR is only 1.47% for the whole 4 datasets when calculating using Dice Score. As the newly released model, SSFormer showed its robustness and strong capability in addressing the task even when faced with bad image conditions. Many of the artifacts only trigger a small number of errors in specific datasets. As is the common case in other experiments, artifacts synthesized on seed images in CVC-ColonDB, as a dataset seldom used in training, can confuse the model the most and generate more errors consequently. Light-reltaed perturbations, including White Balance, Contrast, and Saturation, are able to find corner cases most often, which exactly aligns with our previous conjectures. Also, SSFormer's performance on images synthesized with Blood also decreased, which may suggest that blood in medical images has the potential to fool the model into misclassification even on a relatively robust model. See Table 6 10 for the specific results.

Using the statistics present in tables 7 8 9 10, we calculated the EFR for all four academic models we have tested, including PraNet, SANet, TGANet, SSFormer, and their EFR are 4.38%, 1.70%, 1.47%, and 15.70%, respectively. We can clearly see that SANet and SSFomer are relatively more robust with much lower EFR on our synthesized images, While PraNet and TGANet performed worse and more errors are triggered using synthesized images generated from our framework MedTest.

Detailed visualization of the artifacts and the corresponding output are illustrated in Table11 and Table12, where PraNet and SANet are used as example models.

Table 11: Comparison of PraNet Model Outputs with Different Artifacts

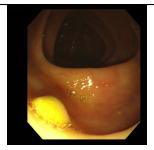
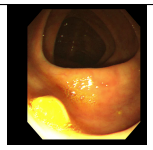
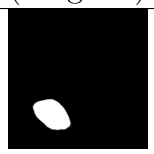
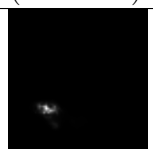
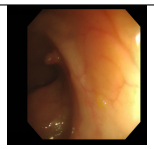
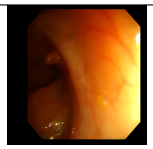
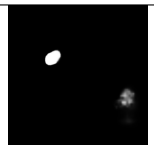
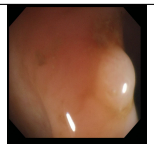
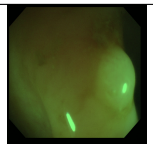
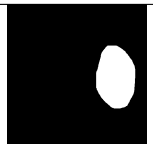
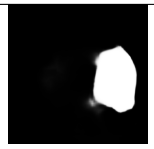
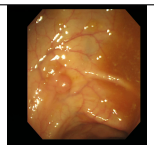
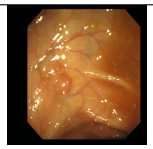
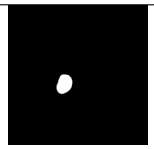
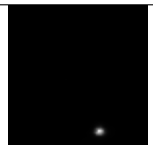
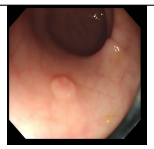
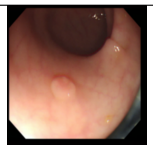
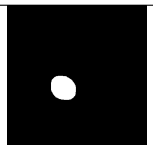
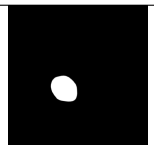
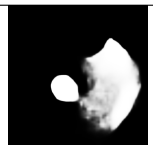
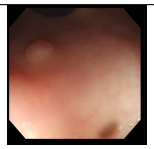
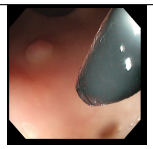
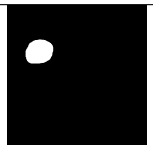
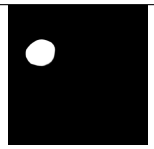
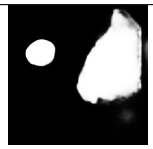
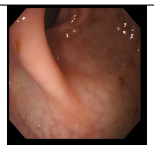
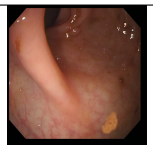
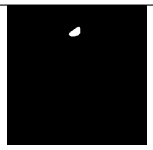
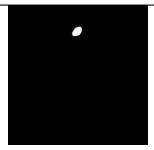
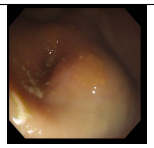
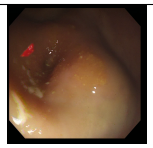
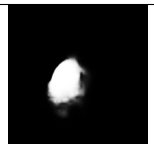
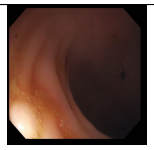
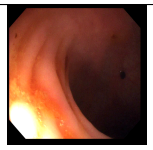
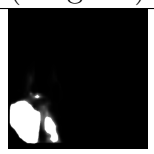
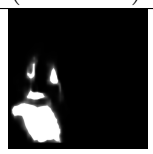
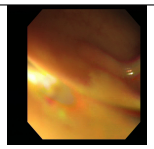
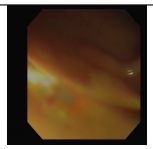
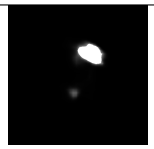
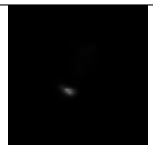
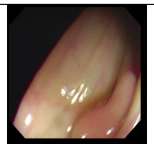
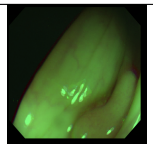
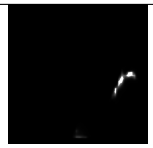
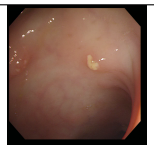
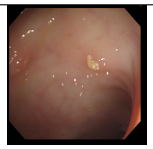
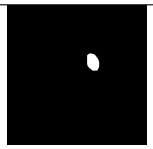
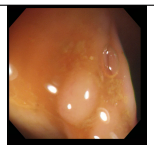
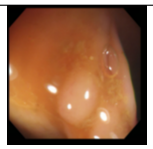
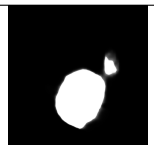
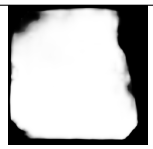
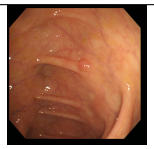
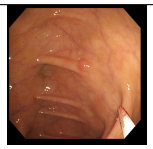
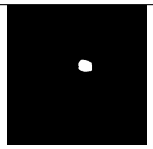
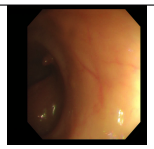
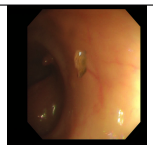
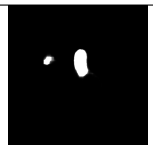
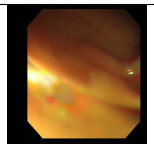
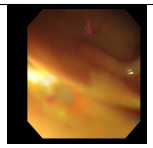
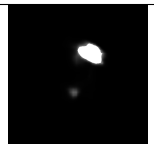
Artifact	Original Image	Image with Artifact	Ground Truth	Output (Original)	Output (Artifact)
Saturation					
Contrast					
White-Balance					
Specularity					
Blur					
Instrument					
Feces					
Blood					
Text					

Table 12: Comparison of SANet Model Outputs with Different Artifacts

Artifact	Original Image	Image with Artifact	Ground Truth	Output (Original)	Output (Artifact)
Saturation					
Contrast					
White-Balance					
Specularity					
Blur					
Instrument					
Feces					
Blood					
Text					

7.2 Visual Question Answering

VQA in the medical domain is an emerging interdisciplinary field that combines techniques from computer vision and natural language processing to interpret and answer questions about medical images. This technology enables automated systems to analyze visual medical data such as X-rays, MRI scans, or endoscopic images, and provide precise textual responses to clinically relevant questions posed by users. Such capabilities are crucial for enhancing diagnostic processes, supporting medical education, and facilitating more efficient clinical decision-making. VQA systems in healthcare help bridge the gap between complex visual data interpretation and actionable medical insights, leveraging large-scale medical datasets and advanced multimodal models to improve accuracy and reliability in medical diagnostics and research.

7.2.1 Evaluation Criteria

Prompt Engineering

Effective prompt engineering is crucial for evaluating LLMs, particularly in VQA tasks where precise responses are essential. Through iterative refinement, we discovered that employing a few-shot learning approach yielded the most consistent and accurate results. Our finalized prompt structure is detailed below.

```
I'm working on the Visual Question Answering tasks on medical endoscopic
images. I will be sending you some endoscopic images, you need to answer all
the questions I give you following the format of sample answer sets. You must
not repeat the question or elaborate on your answer. From now on, answer all
the questions below for the images that I send you.
"Question": "What types of abnormalities are there in the image?"
"Question": "Are there any anatomical landmarks in the image?"
"Question": "Are there any instruments in the image?"
"Question": "Have all polyps been removed?"
"Question": "How many findings are present?"
```

```
"Question": "How many instruments are in the image?"
"Question": "How many polyps are in the image?"
"Question": "Is there a green/black box artefact?"
"Question": "Is there text?"
"Question": "Is this finding easy to detect?"
"Question": "What color is the abnormality?"
"Question": "What color is the anatomical landmark?"
"Question": "What is the size of the polyp?"
"Question": "What type of polyp is present?"
"Question": "What type of procedure is the image taken from?"
"Question": "Where in the image is the abnormality?"
"Question": "Where in the image is the anatomical landmark?"
"Question": "Where in the image is the instrument?"
Sample answers set1: "Polyp" "No" "Biopsy forceps" "No" "2" "1"
"1" "No" "Yes" "Yes" "Red", "Pink" "Not relevant"
">20mm" "Paris ip" "Colonoscopy" "Center-right", "Upper-left" "Not relevant" "
Lower-right"
Sample answers set2: "Polyp" "No" "No" "No" "1" "0"
"1" "Yes" "Yes" "No" "Pink", "White" "Not relevant"
"11-20mm" "Paris iia" "Colonoscopy" "Center", "Lower-right" "Not relevant" "
Not relevant"
In case the question is not applicable to the image or you don't know, please
answer "Not relevant". Please follow the format of the sample answers strictly
.
```

Answer Validation

To assess the accuracy of LLM responses, we developed a software tool that compares each LLM-generated answer to the corresponding ground truth. An answer is deemed correct if it encompasses the ground truth content. Given the few-shot learning context, the format of the response is presumed to be accurate. However, if the number of correct responses for an image falls below a predetermined threshold n, a manual review is initiated to ensure accuracy. Correct answers are tallied as such, while incorrect responses are scrutinized further. It is important to note that responses classified as non-reasonable due to the model's limitations in text comprehension are marked as entirely incorrect.

Metric

For assessing performance in the VQA task, we chose to use the accuracy metric over the EFR. Conceptualizing VQA as a binary classification problem, accuracy is computed using the following formula:

$$\text{Accuracy} = \frac{\text{Number of correct answers}}{\text{Number of generated test cases}} * 100\%. \tag{7}$$

While the EFR could simply be derived as $1 - Accuracy$, it offers less intuitive insight. Thus, we rely on accuracy to clearly represent our results in the VQA evaluations.

7.2.2 Model Performance

The performance of the LLMs, specifically GPT-4V and Gemini, in the VQA task exhibits notable variation.

GPT-4V	Original	Blood	Feces	Instrument	Spot	Saturation	Contrast	WhiteBalance	Blur	Text
Are there any abnormalities in the image?	0.888	0.904	0.871	0.888	0.889	0.878	0.855	0.827	0.777	0.859
Are there any anatomical landmarks in the image?	0.341	0.351	0.314	0.39	0.364	0.355	0.327	0.335	0.364	0.32
Are there any instruments in the image?	0.947	0.947	0.945	0.312	0.946	0.943	0.943	0.941	0.936	0.947
Have all polyps been removed?	0.284	0.31	0.286	0.461	0.231	0.333	0.263	0.204	0.197	0.253
How many findings are present?	0.835	0.859	0.837	0.833	0.827	0.82	0.81	0.782	0.761	0.816
How many instrumnets are in the image?	0.963	0.963	0.957	0.714	0.96	0.957	0.959	0.961	0.946	0.957
How many polyps are in the image?	0.79	0.784	0.747	0.792	0.771	0.784	0.771	0.739	0.716	0.765
Is there a green/black box artefact?	0.624	0.594	0.618	0.653	0.64	0.68	0.659	0.669	0.732	0.48
Is there text?	0.98	0.984	0.982	0.976	0.982	0.973	0.984	0.984	0.863	0.935
Is this finding easy to detect?	0.594	0.631	0.598	0.598	0.598	0.563	0.612	0.527	0.541	0.637
What color is the abnormality?	0.465	0.406	0.416	0.451	0.423	0.371	0.412	0.312	0.402	0.463
What color is the anatomical landmark?	0.369	0.369	0.339	0.422	0.382	0.382	0.363	0.416	0.465	0.351
What is the size of the polyp?	0.21	0.173	0.198	0.194	0.233	0.224	0.192	0.186	0.225	0.229
What type of polyp is present?	0.176	0.159	0.135	0.169	0.179	0.159	0.169	0.171	0.171	0.153
What type of procedure is the image taken from?	0.976	0.98	0.982	0.982	0.984	0.984	0.978	0.91	0.903	0.976
Where in the image is the abnormality?	0.653	0.604	0.616	0.676	0.67	0.647	0.624	0.629	0.65	0.676
Where in the image is the anatomical landmark?	0.365	0.365	0.337	0.416	0.384	0.38	0.365	0.414	0.461	0.349
Where in the image is the instrument?	0.924	0.924	0.926	0.669	0.93	0.92	0.922	0.927	0.913	0.929
Average	0.632	0.628	0.617	0.589	0.633	0.631	0.623	0.607	0.612	0.616

Table 13: The VQA results of GPT-4V

GPT-4V

The GPT-4V model demonstrated robust performance on several questions such as **Is there text?**, **What type of procedure is the image taken from?**, and **How many instruments are in the image?** with respect to the original images. However, its performance was less satisfactory on questions like **What type of polyp is present?**, **What is the size of the polyp?**, and **Have all polyps been removed?**.

GPT-4V	Original	Average Perturbation	Difference (Original - Average)
Are there any abnormalities in the image?	0.888	0.861	0.027
Are there any anatomical landmarks in the image?	0.341	0.347	-0.006
Are there any instruments in the image?	0.947	0.873	0.074
Have all polyps been removed?	0.284	0.282	0.002
How many findings are present?	0.835	0.816	0.019
How many instrumnets are in the image?	0.963	0.93	0.033
How many polyps are in the image?	0.79	0.763	0.027
Is there a green/black box artefact?	0.624	0.636	-0.012
Is there text?	0.98	0.963	0.017
Is this finding easy to detect?	0.594	0.589	0.005
What color is the abnormality?	0.465	0.406	0.059
What color is the anatomical landmark?	0.369	0.388	-0.019
What is the size of the polyp?	0.21	0.206	0.004
What type of polyp is present?	0.176	0.163	0.013
What type of procedure is the image taken from?	0.976	0.964	0.012
Where in the image is the abnormality?	0.653	0.644	0.009
Where in the image is the anatomical landmark?	0.365	0.386	-0.021
Where in the image is the instrument?	0.924	0.896	0.028
Average	0.632	0.617	0.015

Table 14: The summary of VQA results of GPT-4V

GPT-4V	Original	Blood	Feces	Instrument	Spot	Saturation	Contrast	WhiteBalance	Blur	Text
Are there any abnormalities in the image?	0.888	0.904	0.871	0.888	0.889	0.878	0.855	0.827	0.777	0.859
Are there any instruments in the image?	0.947	0.947	0.945	0.312	0.946	0.943	0.943	0.941	0.936	0.947
Have all polyps been removed?	0.284	0.31	0.286	0.461	0.231	0.333	0.263	0.204	0.197	0.253
How many findings are present?	0.835	0.859	0.837	0.833	0.827	0.82	0.81	0.782	0.761	0.816
How many instrumnets are in the image?	0.963	0.963	0.957	0.714	0.96	0.957	0.959	0.961	0.946	0.957
How many polyps are in the image?	0.79	0.784	0.747	0.792	0.771	0.784	0.771	0.739	0.716	0.765
Is there a green/black box artefact?	0.624	0.594	0.618	0.653	0.64	0.68	0.659	0.669	0.732	0.48
Is there text?	0.98	0.984	0.982	0.976	0.982	0.973	0.984	0.984	0.863	0.935
What color is the abnormality?	0.465	0.406	0.416	0.451	0.423	0.371	0.412	0.312	0.402	0.463
What is the size of the polyp?	0.21	0.173	0.198	0.194	0.233	0.224	0.192	0.186	0.225	0.229
What type of polyp is present?	0.176	0.159	0.135	0.169	0.179	0.159	0.169	0.171	0.171	0.153
What type of procedure is the image taken from?	0.976	0.98	0.982	0.982	0.984	0.984	0.978	0.91	0.903	0.976
Where in the image is the abnormality?	0.653	0.604	0.616	0.676	0.67	0.647	0.624	0.629	0.65	0.676
Where in the image is the instrument?	0.924	0.924	0.926	0.669	0.93	0.92	0.922	0.927	0.913	0.929
Average	0.694	0.685	0.680	0.626	0.690	0.691	0.682	0.660	0.657	0.674

Table 15: The VQA results of GPT-4V after deleting the ambiguous questions

Despite these variances, GPT-4V showed considerable resilience to perturbations, maintaining an average accuracy difference of only 1.5% between the original and perturbed data. See Table 13 14 for the specific results.

Several questions were deemed unsuitable for our evaluations, as noted in Section 4.1.2, including **Are there any anatomical landmarks in the image?**, **Is this finding easy to detect?**, **What color is the anatomical landmark?**, and **Where in the image is the anatomical landmark?**. These were omitted due to their ambiguity or irrelevance to the task objectives. Upon excluding these questions, there was

GPT-4V	Original	Average Perturbation	Difference (Original - Average)
Are there any abnormalities in the image?	0.888	0.861	0.027
Are there any instruments in the image?	0.947	0.873	0.074
Have all polyps been removed?	0.284	0.282	0.002
How many findings are present?	0.835	0.816	0.019
How many instrumnets are in the image?	0.963	0.93	0.033
How many polyps are in the image?	0.79	0.763	0.027
Is there a green/black box artefact?	0.624	0.636	-0.012
Is there text?	0.98	0.963	0.017
What color is the abnormality?	0.465	0.406	0.059
What is the size of the polyp?	0.21	0.206	0.004
What type of polyp is present?	0.176	0.163	0.013
What type of procedure is the image taken from?	0.976	0.964	0.012
Where in the image is the abnormality?	0.653	0.644	0.009
Where in the image is the instrument?	0.924	0.896	0.028
Average	0.694	0.672	0.022

Table 16: The summary of VQA results of GPT-4V after deleting the ambiguous questions

an observed performance increase of 6.2% on original images and 5.5% on perturbed images, leading to an overall accuracy discrepancy of 2.2% between the two data sets. See Table 15 16 for the specific results.

Gemini

In contrast, the performance of the Gemini model was less satisfactory. It faced difficulties in interpreting the prompts accurately, even after substantial prompt engineering efforts. The responses generated by Gemini proved challenging to analyze, even for human evaluators. We are actively continuing to test and refine the interaction with Gemini, aiming to enhance the quality of its responses through various methodological improvements.

7.3 Classification

Classification is also a task of great value and importance in the field of medical imaging, which targets the problem of identifying the disease condition and category based on the visual effects captured by medical devices. To keep consistency, as our study mainly delved into the field of gastrointestinal images, including endoscopy and colonoscopy images,

we observed that most of the classification methods focus on wireless capsule endoscopy image input. Therefore, we decided to use this kind of sub-type as our major seed image source.

As a painless, noninvasive imaging tool, wireless capsule endoscopy has been widely adopted for direct visualization and early screening of the gastrointestinal(GI) diseases. Automatic recognition algorithms are of high demand due to the large amount of data in clinical videos and potential limitations caused by human factors such as subjectivity. As these intelligent methods play an vital part in assisting clinical diagnosis, ensuring their robustness and reliability is also of immense necessity, because a minor mistake in medical-related decision may result in severe and irreversible consequences.

Among the GI diseases that receive extensive concern, vascular lesion and inflammatory are two common types, mainly because they are important syndromes or indicators of other GI abnormalities such as bleeding, ulcers and Crohn's diseases.[76] These two syndromes are usually not obvious in the images, taking up tiny regions with unclear boundary, which poses a huge challenge on algorithms targeting this task. Besides, very minor differences may lead to divergent decision, which is the exact situation we want to test out for models and help them to prevent. Detailed samples of these two classes (vascular lesion and inflammatory) are shown in Fig. 3.

In our following evaluation on academic models targeting the classification task, the classification is on annotating the given image into one of the three types of normal image, vascular lesion image and inflammatory image.

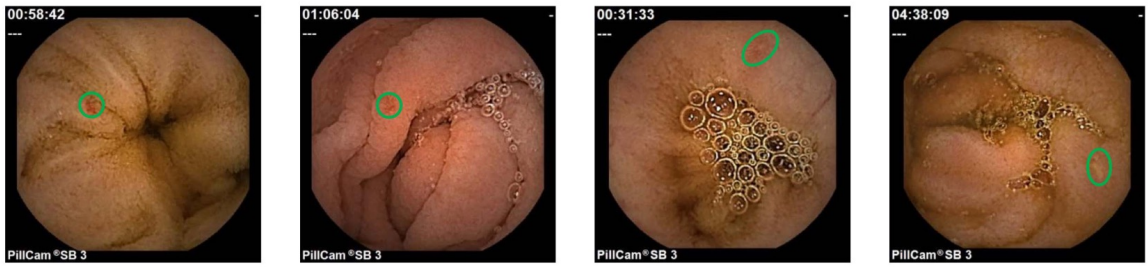


Figure 3: Samples of WCE images. First two images contain vascular lesion and the other two images show inflammatory. The lesion areas are annotated by the green circles.[76]

7.3.1 Evaluation Criteria

Given that our problem is a multi-class classification with three categories, we not only need to consider model's overall performance across the whole dataset, but also should evaluate the model's ability to distinguish each class into its corresponding true label instead of biasing towards a specific type too much. From this perspective, the evaluation criteria we decided on our classification task consists of Accuracy, Recall, Precision, F1 Score and Cohen's Kappa Statistics.

Accuracy score refers to the proportion of accurately classified samples to the total number of test cases, which adversely aligns with the definition of our previously proposed error-finding rate (EFR). In other words, the higher the accuracy is, the lower the EFR will be for each of the perturbed type of images, and their relation follows the formula derived below. (For simplicity, we denote the number of correctly classified test cases as T_n, the number of misclassified test cases as F_n and the total number of generated/seed test cases as N.

Given that,

$$\text{EFR} = \frac{F_n}{N} * 100\%, \tag{8}$$

We can derive that,

$$\begin{aligned} \text{Accuracy} &= \frac{T_n}{N} * 100\% \\ &= \frac{N - F_n}{N} * 100\% \\ &= 1 - EFR \end{aligned} \tag{9}$$

Since our classification involves multiple classes instead of simply the binary choice, we also included Precision and Recall into our evaluation criteria to examine the models' performance with regard to distinct image class. Here, we denote True Positive (TP) as the number of samples that are correctly classified into their true label, False Positive (FP) as the number of samples classified as a specific class apart from their actual class, and False Negative (FN) as the number of samples that are in reality in a specific class but classified wrongly into another. Based on the above definition, we can further illustrate the formula for Precision and Recall as follows. For a specific class C,

$$\text{Precision(Class = C)} = \frac{TP(Class = C)}{TP(Class = C) + FP(Class = C)} \tag{10}$$

$$\text{Recall(Class = C)} = \frac{TP(Class = C)}{TP(Class = C) + FN(Class = C)} \tag{11}$$

From the above equations, we can interpret Precision score as the proportion of samples that are actually of class C among all the samples classified into class C, measuring the ability of a classifier to identify only the correct instances for each class. Similarly, Recall represents the proportion of samples that are correctly classified as its true label among

all the samples with true label of class C, examining the ability of a classifier to find all correct instances per class.

Following the idea of Precision and Recall, F1 score emerges as a weighted harmonic mean of Precision and Recall normalized between 0 and 1, which is an indicator on the combined effect of both Precision and Recall. F1 score can be expressed as,

$$\begin{aligned} \text{F1 Score} &= \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} \\ &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \\ &= \frac{TP}{TP + \frac{1}{2}(FP + FN)} \end{aligned} \tag{12}$$

From the above relation, we can see that F1 score encourages similar values for Precision and Recall and maximizing F1 score has a joint effect of maximizing both the metrics.

Since F1 score is calculated distinctively for each class, we have to choose an averaging method to track the overall F1 score of all the classes. The commonly adopted methods are macro-averaging, micro-averaging and sample-weighted-averaging, where the latter two take the potential data imbalance issue into consideration.

Cohen's Kappa score is one of the most commonly-used and representative metrics for evaluating multi-class classifiers on imbalanced datasets. The traditional metrics for evaluation may be faced with the problem of biasing towards the majority class and assuming an identical distribution of the actual and predicted class. Therefore, Cohen's Kappa score can be leveraged to measure the proximity of the predicted classes to the actual classes when compared to a random classification. The output is

normalized between 0 and 1 the metrics for each classifier. Therefore, it can be directly compared across the classification task. In general, the closer to one the score is, the better the classifier is.

The formal definition of Cohen's Kappa score is defined as follows,

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{13}$$

where p_o is the relative observed agreement among raters, and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category.

In the following section, we will discuss the model performance with seed images and perturbed images as testing input using the above evaluation metrics as criteria.

7.3.2 Model Performance

The models selected are centering around the task of classifying WCE images into normal type and two GI diseases, namely vascular lesion and inflammatory type. Given that different implementations focus on addressing different challenges, their detailed structures vary, and thus, their robustness and stability when facing corner cases naturally may also vary. Therefore, we will discuss how our evaluated models performed when feeding our generated images with artifacts by our method MedTest as the input.

AGDN

AGDN utilizes a two-branch attention guided deformation network that uses the attention maps to locate and zoom in lesion regions to learn the

Artifact	Accuracy	Cohen's Kappa	F1 Score
Original	0.893	0.836	0.893
Blur	0.702	0.517	0.660
Contrast	0.747	0.602	0.735
Feces	0.797	0.682	0.790
Instrument	0.852	0.773	0.852
Saturation	0.685	0.516	0.682
Spot	0.817	0.712	0.811
Text	0.828	0.733	0.825
White Balance	0.532	0.226	0.463
Average	0.761	0.622	0.746

Table 17: Evaluation Metrics from AGDN Model.

important features with the specific areas.[76] However, this may lead to a problem of paying too much attention on details without thoroughly considering the whole image and the relation between different regions in the image. On the other hand, the change in light condition in images can potentially lead to the degradation of the deformation grids' ability in information representation. As we can see from the table 17, most of the perturbed image types incurred significant decrease in model's performance compared to the original input. Among all the artifacts, white balance and saturation artifacts triggered the most severe errors, possibly resulted from the change in light condition, specifically under a darker and more abnormal lighting. Another evident error-triggering perturbation is blurring, which may cause a fuzzy boundary for the model to recognize the key region. These effects can potentially confuse the model by making the lesion regions similar to other parts of the tissue, leading to inaccurate classification decisions.

DSI-Net

Artifact	Accuracy	Cohen's Kappa	F1 Score
Original	0.940	0.908	0.940
Blur	0.897	0.841	0.896
Contrast	0.883	0.823	0.884
Feces	0.907	0.858	0.907
Instrument	0.897	0.843	0.897
Saturation	0.755	0.635	0.757
Spot	0.932	0.895	0.931
Text	0.908	0.859	0.908
White Balance	0.728	0.574	0.711
Average	0.872	0.804	0.870

Table 18: Evaluation Metrics from DSI-Net Model.

The key feature of DSI-Net is its deep synergistic interaction network for joint classification and segmentation tasks with WCE images, with one classification branch and two segmentation branches (coarse and fine segmentation). The three branches share the same backbone network and interact with each other to learn feature representations collaboratively.[82] Given this design, DSI-Net's performance is relatively more stable and consistent under extreme cases, as we can find in the table18. Similar to AGDN, white balance and saturation caused the most evident performance drop compared to other artifact types. The possible reason may also derive from the unclear difference between lesion areas and normal areas due to the bad lighting condition. This can further result in model's decreased capability in segmenting out the lesion parts, which is closely correlated with its classification accuracy on image types.

From the statistics in the two classification evaluation tables 1718, we can see dramatic performance drops with some of the perturbations as reflected in the Accuracy, Cohen's Kappa score and F1 score. Compared to the Accuracy score obtained on seed images, the white balance pertur-

bation led to the most down-side effect in both models, with a decrease up to 36.1% and 21.2% for AGDN model and DSI-Net model respectively. Moreover, the total average Accuracy score (including seed images) was 13.2% and 6.8% lower than that with only seed images (original testing score), showing that our method MedTest can effectively detect potential errors in the model with various perturbations. In general, we can observe that DSI-Net is a more robust and stable model than AGDN, though both of them have room for improvement.

Answer to RQ2: MedTest achieves up to a 15.70% EFR in testing SOTA segmentation models, causes a 2.2% average accuracy drop in the most advanced commercial LLMs, and leads to up to a 36.1% accuracy reduction in SOTA classification models. These results indicate that MedTest can effectively uncover corner cases and is valuable for further robustness testing of other models.

8 RQ3: Enhancing Medical Image Diagnosis Performance Using MedTest-Generated Test Cases

Our research has substantiated that MedTest is adept at creating diagnostically consistent and realistic test cases, which are proficient in identifying errors in both commercial software and SOTA academic models. This leads to an imperative query: Can the test cases generated by MedTest be leveraged to augment the performance of medical image diagnosis systems? Essentially, the objective is to enhance the robustness of diagnostic models.

A logical approach to achieve this enhancement is through the further training of models with test cases synthesized by MedTest, to assess if

such models exhibit increased resilience to a variety of perturbations. For this performance enhancement purpose, we could utilize our existing test set. We randomly select images with synthesized effects that might trigger errors, and constructing a new training dataset. This dataset will be an amalgamation of original and synthesized diagnostically consistent images.

8.1 Segmentation

In an effort to enhance model performance, we further trained our segmentation models using MedTest.

8.1.1 Dataset Construction

We developed a specialized dataset tailored for segmentation training. Drawing from established practices, we utilized the Kvasir and CVC-ClinicDB datasets, as referenced in the work by Fan et al.[17] and commonly employed for polyp segmentation tasks like those described by Wei et al.[74]. To adapt these datasets for training with MedTest, we incorporated an equal number of perturbed images. Specifically, for each of the 9 types of perturbations, we randomly selected an equal fraction, ensuring that the dataset comprised an equivalent number of original and perturbed images. This approach resulted in a balanced dataset containing a total of 3223 images, split between 1612 original and 1611 perturbed images.

For testing, we assembled a set using images from CVC-300 and CVC-ColonDB. Each image in this set was subjected to the same 9 perturbations, producing a comprehensive test set of 4400 images.

8.1.2 Hyperparameter Tuning

We followed the hyperparameter settings given by the original paper and code.

PraNet

- Learning rate: 1×10^{-4}
- Epoch: 20
- Batch size: 16
- Learning rate decay: 0.1 in 50 epochs
- Device: Intel(R) Xeon(R) Platinum 8352V CPU and RTX 4090 GPU

SANet

- Learning rate: 0.4
- Epoch: 128
- Batch size: 64
- Learning rate decay: 0.5 at the $64th, 96th$ epoch
- Device: Intel(R) Xeon(R) Platinum 8352V CPU and RTX 4090 GPU

SSFormer

- Learning rate: 1×10^{-5}
- Epoch: at most 100, with early stopping
- Batch size: 16
- Learning rate decay: None
- Device: Intel(R) Xeon(R) Platinum 8352V CPU and RTX 4090 GPU

8.1.3 Result

In line with the dataset configuration previously described, we proceeded to further train the evaluated segmentation models to explore potential enhancements in their performance. However, the TGANet model, which uniquely relies on text inputs for aiding information extraction in the segmentation process, could not be improved through the same continuous training approach due to the absence of these text inputs. Consequently, our experiments were limited to the remaining models: PraNet, SANet, and SSFormer. For these tests, we opted to use the accuracy metric instead of the EFR, as accuracy provides a more direct comparison of training outcomes. A comprehensive discussion of the results and their analysis will be provided in the subsequent section.

PraNet	**CVC-300**				**CVC-ColonDB**			
	Dice Score		**IoU Score**		**Dice Score**		**IoU Score**	
	Before	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**
Original	0.86	**0.889**	0.777	**0.817**	0.696	**0.701**	0.619	**0.629**
Saturation	0.818	**0.878**	0.735	**0.806**	0.677	**0.711**	0.598	**0.636**
White Balance	0.815	**0.879**	0.74	**0.805**	0.624	**0.709**	0.551	**0.634**
Contrast	0.861	**0.887**	0.777	**0.815**	0.692	**0.707**	0.618	**0.633**
Spot	0.849	**0.885**	0.764	**0.812**	0.694	**0.702**	0.615	**0.628**
Blur	0.706	**0.872**	0.619	**0.794**	0.582	**0.695**	0.491	**0.618**
Text	0.744	**0.819**	0.659	**0.746**	0.629	**0.663**	0.554	**0.593**
Instrument	0.812	**0.879**	0.717	**0.805**	0.653	**0.699**	0.571	**0.626**
Blood	0.843	**0.891**	0.751	**0.821**	0.678	**0.696**	0.599	**0.624**
Feces	0.838	**0.878**	0.75	**0.804**	0.674	**0.691**	0.592	**0.619**
Average	0.815	**0.876**	0.729	**0.803**	0.66	**0.697**	0.581	**0.624**

Table 19: Further training results on PraNet

PraNet

The enhanced training on PraNet significantly improves its segmentation capabilities across both testing sets, with a improvement of 6.1%, 7.4%, 3.7%, and 4.3% on the Dice score, IoU score of the CVC-300 and CVC-ColonDB respectively. All scores, encompassing both datasets and

SANet	CVC-300				CVC-ColonDB			
	Dice Score		IoU Score		Dice Score		IoU Score	
	Before	After	Before	After	Before	After	Before	After
Original	0.898	**0.904**	0.828	**0.836**	0.757	**0.763**	0.677	**0.69**
Saturation	0.879	**0.882**	0.807	**0.816**	**0.766**	0.757	0.683	0.683
White Balance	0.877	**0.889**	0.805	**0.822**	0.738	**0.76**	0.655	**0.687**
Contrast	**0.898**	0.897	0.827	**0.829**	0.754	**0.765**	0.673	**0.691**
Spot	0.899	**0.904**	0.828	**0.837**	0.754	**0.76**	0.675	**0.686**
Blur	0.851	**0.899**	0.773	**0.831**	0.735	**0.773**	0.646	**0.697**
Text	0.803	**0.864**	0.727	**0.793**	0.694	**0.749**	0.621	**0.674**
Instrument	0.898	**0.903**	0.829	**0.836**	0.747	**0.749**	0.668	**0.679**
Blood	0.899	**0.903**	0.829	**0.835**	0.753	**0.762**	0.675	**0.689**
Feces	0.901	**0.904**	0.831	**0.837**	0.743	**0.759**	0.665	**0.687**
Average	0.88	**0.895**	0.808	**0.827**	0.744	**0.76**	0.664	**0.686**

Table 20: Further training results on SANet

SSFormer	CVC-300				CVC-ColonDB			
	Dice Score		IoU Score		Dice Score		IoU Score	
	Before	After	Before	After	Before	After	Before	After
Original	0.891	0.891	0.825	**0.827**	0.774	0.774	0.698	**0.700**
Saturation	0.841	**0.876**	0.779	**0.806**	0.778	0.778	0.699	**0.702**
White Balance	**0.880**	0.874	**0.813**	0.811	0.731	**0.764**	0.656	**0.693**
Contrast	**0.883**	0.882	0.817	**0.817**	0.765	**0.774**	0.689	**0.700**
Spot	0.892	0.892	0.826	**0.828**	0.770	**0.775**	0.695	**0.701**
Blur	0.883	**0.893**	0.813	**0.825**	0.766	0.766	0.690	**0.693**
Text	0.892	**0.893**	0.825	**0.830**	0.768	**0.769**	0.691	**0.696**
Instrument	0.872	**0.886**	0.800	**0.820**	0.747	**0.769**	0.672	**0.695**
Blood	0.877	**0.891**	0.809	**0.827**	0.760	**0.770**	0.684	**0.697**
Feces	0.898	**0.899**	0.831	**0.835**	0.753	**0.759**	0.679	**0.687**
Average	0.881	**0.888**	0.814	**0.823**	0.761	**0.770**	0.684	**0.696**

Table 21: Further training results on SSFormer

metrics, exhibit improvements, confirming the augmented training's universal benefit to the model's performance. Notably, the "Blur" artifact demonstrates the most remarkable improvement, with over a 10% increase across both datasets and metrics, highlighting the training's effectiveness in addressing complex scenarios. See Table 19 for the specific results.

SANet

Post-enhancement training on SANet substantially uplifts its segmenta-

tion efficacy on both the CVC-300 and CVC-ColonDB datasets, demonstrating an improvement of 1.5%, 1.9%, 1.6%, and 2.2% in the Dice score and IoU score respectively. All metrics across both datasets show progress, underscoring the universal enhancement in SANet's performance through further training. The "Text" artifacts exhibit particularly notable improvements, with "Text" achieving up to a 6.6% increase in the IoU score on CVC-300, indicating the training's success in handling intricate scenarios. See Table 20 for the specific results.

SSFormer

SSFormer is already a very robust model, yet we can find clear evidence that its segmentation performance on the testing dataset improved for a certain degree. The table21 shows the increase in Dice and IOU score on CVC-300 and CVC-ColonDB of 0.7%, 0.9%, 0.9%, and 1.2% respectively, marking a positive feedback on the augmented data provided by our method MedTest. To examine closer, we can see that the model performance after continuous training with our customized dataset on some of the perturbations that triggered greater errors, such as white balance, instrument and feces, were drawn to a similar level as others. This proved our method MedTest's effectiveness in assisting in the improvement of model's robustness and stability.

8.2 Classification

As for classification, we applied the same method and techniques in order to seek further improvement on models' robustness and stability especially when faced with corner cases.

8.2.1 Dataset Construction

In the original experimental setting, we have already split the parts for training and testing, and our previous model evaluation were conducted with utilizing testing dataset as the seed image. Therefore, in this section, we following the implementation and training of AGDN model [76], that is, leverage the same subset of training data to construct our customized data for further training. Following the same logic as the dataset construction in the segmentation task, we first generated the whole perturbation dataset with our method MedTest. To keep consistent to the original model performance and prevent it from biasing too much towards artifact features in learning, we included all of the seed images into our customized dataset.

As we discovered that the blood perturbation type will affect the classification accuracy because it actually resembles medical landmarks used for GI disease diagnosis. Therefore, we removed the blood perturbation in the construction of further training dataset and included the remaining 8 perturbations in the construction.

For each of the perturbation, we randomly select $\frac{1}{8}$ of the total number of images. In this way, all the perturbed images will add up to the number of the seed images, leading to a ratio between seed images and images with artifacts as $1:1$. In total, we will obtain a customized dataset for further training of size 4934.

8.2.2 Hyperparameter Tuning

Generally, we followed the implementation details in the original repository, while making some necessary adjustment according to our situation.

AGDN

- Learning rate: 0.01
- Epoch: 80
- Batch size: 8
- Learning rate decay: Reduce to 5×10^{-3} at the $60th$ epoch
- Device: Intel(R) Xeon(R) Platinum 8352V CPU and RTX 4090 GPU

DSI-Net

- Learning rate: 1×10^{-4}
- Epoch: 100
- Batch size: 8
- Learning rate decay: None
- Device: Intel(R) Xeon(R) Platinum 8352V CPU and RTX 4090 GPU

8.2.3 Result

In the following part, we will illustrate how our method MedTest can be used to improve the robustness and overall performance of the models that underwent evaluation.

Artifact	Initial Accuracy	Enhanced Accuracy	Difference
Original	0.893	0.885	-0.008
Blur	0.702	0.808	**+0.106**
Contrast	0.747	0.812	**+0.065**
Feces	0.797	0.860	**+0.063**
Instrument	0.852	0.852	0.000
Saturation	0.685	0.770	**+0.085**
Spot	0.817	0.873	**+0.056**
Text	0.828	0.858	**+0.030**
White Balance	0.532	0.673	**+0.141**
Average	0.761	0.821	**+0.060**

Table 22: Comparison on Accuracy Score for AGDN Model Before and After Data Augmentation.

Artifact	Initial Accuracy	Enhanced Accuracy	Difference
Original	0.940	0.947	**+0.007**
Blur	0.897	0.918	**+0.021**
Contrast	0.883	0.917	**+0.034**
Feces	0.907	0.937	**+0.030**
Instrument	0.897	0.928	**+0.031**
Saturation	0.755	0.835	**+0.080**
Spot	0.932	0.942	**+0.010**
Text	0.908	0.908	0.000
White Balance	0.728	0.848	**+0.120**
Average	0.872	0.909	**+0.037**

Table 23: Comparison on Accuracy Score for DSI-Net Model Before and After Data Augmentation.

In Table 22 23, we can see performance improvement on almost all perturbations, generally leading to an overall increased accuracy on the testing dataset.

In both of the models, we can see an evident rise in the perturbations that did not perform well originally, such as lighting-related perturbations of saturation and white balance. Feeding the models with samples of these classes can allow them to draw specific information patterns within the artifact features, thus learning to classify these previously "unseen" image types into the actual categories.

Although the difference between the average accuracy scores before and after further training is not as evident as distinct artifacts because of taking the average, we can clearly conclude that our method MedTest can be leveraged to improve the robustness of the tested models.

Answer to RQ3: Test cases generated by MedTest can be leveraged to contruct our customized training dataset and effectively improve the robustness of academic medical image diagnosis models through further training on both segmentation and classification tasks.

9 RQ4: How would different factors affect the performance of MedTest?

This section delves into how three distinct external factors influence the efficacy of MedTest.

Image Structure and Overlay

The heterogeneity in the source images' locations and orientations presents challenges in our automated object perturbation system. The lack of comprehensive image analysis during object addition precludes optimal object selection and placement, potentially leading to incongruous object positioning in the synthesized images. Efforts to mitigate this include excluding objects from atypically laid out images (such as instruments positioned at corners in partial views) and constraining the target positions for merging objects to more closely resemble their original context, albeit with slight positional variations. These measures aim to minimize the incongruities arising from layout and positioning discrepancies.

Medical Landmark Characteristics

In general, the medical landmarks present in the images may potentially affect the effectiveness of our MedTest.

As for the segmentation task on polyps, the extensive diversity in the

dataset, particularly regarding polyp size and shape, poses challenges for the automated synthesis of object-related perturbations. The presence of large polyps can complicate object addition, necessitating refined automation protocols for object selection and placement. This adjustment must accommodate the variation in polyp characteristics, striking a balance between overall performance and the generation of some suboptimal results.

For the classification task, one of the most important influence is that the presence of bleeding in the image will radically alter the class that the image belongs to because the two types of GI diseases, namely vascular lesion and inflammatory, are closely related to the bleeding situation on tissues. Therefore, to prevent the undesired influence, we have to exclude the blood perturbation from the whole evaluation process of classification methods.

Ambient Lighting Conditions

Divergent lighting conditions in the seed images, especially those that are overly dark or bright, can lead to unnatural effects in object perturbations. To address this, we have implemented contrast, brightness, and color assessments for both the target and object-containing images. This enables the imposition of similarity constraints when selecting objects for synthesis, thereby mitigating unnatural outcomes. Lighting conditions also affect light-related perturbations such as saturation, contrast, and specularity. Excessively dark or bright seed images can render the application of saturation or contrast effects counter-intuitive and unnatural. Specularity synthesis is similarly impacted in underexposed images.

Answer to RQ4: The performance of MedTest can potentially be affected by the above proposed factors, including image structure, medical landmark characteristics and ambient lighting conditions. We have taken these factors into consideration in the design of MedTest and tried to mitigate the negative effect to the greatest extent in our implementation.

10 Discussion

10.1 Threats to Validity

This section elucidates potential threats that could affect the validity of our study.

10.1.1 Variability in Diagnosis Ground Truth

A primary concern is the potential alteration in the diagnostic accuracy of test cases generated by MedTest, especially after numerous perturbations, which could lead to false positives. To mitigate this risk, we engaged in expert annotation to affirm the diagnostic ground truth of these generated test cases. Additionally, annotators were instructed to assess whether the test cases authentically represent artifacts encountered in real-world clinical settings. The findings confirm that the artifacts generated by our methodology are diagnosis-neutral.

10.1.2 Scope of Application on Endoscope Image Analysis

Another concern is the applicability of MedTest primarily to endoscope image analysis, which may not be universally extendable to other types of medical images. The selection of endoscope imagery was a deliberate decision, considering its representativeness in a specific medical imag-

ing context. However, we posit that the MRs developed can be readily adapted to other medical imaging modalities. We provide a comprehensive framework encompassing the study of clinical artifacts, formulation and design of MRs, generation of test cases, and utilization of failure cases to enhance robustness.

10.1.3 Evaluation on a Limited Set of Medical Image Analysis Systems

Our evaluation encompassed 7 medical image analysis systems, which may not comprehensively represent MedTest's efficacy across diverse systems. To address this, our evaluation targeted both commercial software employing LLMs and SOTA academic models pertinent to our focused task. Future endeavors will involve extending our testing to a broader array of commercial and research models to further validate and enhance the generalizability of MedTest's performance.

10.2 Performance of Perturbations

10.2.1 Segmentation

Based on our observations on the experiment results in the segmentation task, we found evident discrepancy between the degree to which the models are influenced based on different artifact types.

Generally, light-related perturbations, including white balance, saturation and contrast, may affect the model performance more severely, that is, leading to a higher EFR.

Especially, we discovered that models ability to properly segment polyps reduce more dramatically when the boundary of the polyp is unclear. A common case is when synthesized images in contrast (underexposure)

category are input to the model, the model cannot distinguish the polyps with their surrounding, thus outputting unsatisfying prediction masks. White balance may also affect the model performance in a relatively similar way. Because we adjust the RGB channel to imbalance by reducing the undesired channel values to make the value of the vital channel outweigh other, this at the same time make the image darker (as the channel value reduces). We suspected that this is one reason why white balance biases can sometimes lead to most dramatic drop in model performance. Besides, when using the test cases with generated saturation, the rise in EFR may result from the scenario that polyps are present at the area with overexposure. In this case, the color in the overexposed areas will tend to white, making the model unable to distinguish the polyp boundary.

Blurring effect may sometimes lead to higher EFR, especially when testing on PraNet. We think the mechanism in confusing the model may be similar to the light perturbations, that is, making the edge of the polyp unclear so that the model will regard it as normal tissue as its surroundings or segmenting a much larger area with the unimportant tissues.

Surprisingly, the EFR on object perturbations do not affect models as seriously as we expected, though constantly triggering a small amount of error. The major reason for such object-based artifacts to influence model performance is mislead the model to misinterpret them as "polyps" that should be segmented out. Indeed, based on our observations on the model output predictions, this was usually the case when model performance decreased. However, because of the relatively obvious difference between these objects and polyps, their ability to fool the models is limited.

10.2.2 Visual Question Answering

Unexpectedly, the performance of the LLMs was most significantly affected by the presence of instrument artifacts. These models frequently failed to recognize newly introduced instruments, considerably lowering their scores on relevant questions. Interestingly, instruments are not commonly present in standard scenarios, leading us to question whether this issue stems from the models' inability to detect instruments or a default tendency to predict their absence. Moreover, while questions involving instruments generally led to reduced scores, the presence of instrument artifacts paradoxically improved model performance on other questions, such as **Have all polyps been removed?**. These observations suggest intriguing avenues for future investigation and discussion.

10.2.3 Classification

Our findings indicate that perturbations related to lighting have the most profound impact on model performance, significantly reducing accuracy scores more than other types of disturbances. This decline in performance is attributed to the fact that alterations in lighting can obscure the distinction between lesion areas and normal tissue, leading to erroneous model predictions. Similarly, blurring perturbations compromise model accuracy by smudging the visual clarity needed for accurate classification. Furthermore, the presence of fecal matter in images poses additional challenges, as it can mimic the appearance of lesions to some extent, further confusing the model and affecting its performance metrics.

11 Related Work

11.1 Enhanced Testing Approaches for AI Software

AI software has permeated numerous sectors, notably transforming technologies such as autonomous vehicles and advanced facial recognition systems. Nonetheless, a significant challenge is their susceptibility to errors, which could lead to severe mishaps or accidents, a concern underscored by various studies[83, 36]. In response, there has been an intensified focus on developing techniques to test these systems more rigorously. Researchers have devised numerous methods to generate adversarial examples or specialized test cases aimed at revealing vulnerabilities in AI systems[7, 48, 69, 80, 79, 67, 40, 47, 78, 52, 26, 25]. Alongside these, there are significant efforts to enhance AI resilience through robust training protocols and advanced debugging techniques[42, 4, 18, 72, 41, 62]. Our research contributes to this area by focusing on the robustness of medical image diagnosis software, a critical AI application that has not been extensively evaluated in prior studies.

11.2 Comprehensive Analysis of Robustness in Medical Image Analysis Software

Our thorough literature review has explored testing and attack methodologies applied to medical image analysis systems, integrating insights from fields like natural language processing (NLP) and computer vision (CV). Various metamorphic testing strategies have been developed for NLP applications, pioneering innovative approaches[8, 9, 21, 23, 24, 51, 59]. In parallel, the CV domain has experienced substantial progress in identifying software errors, drawing heavily from adversarial techniques[20, 31, 33, 37, 77]. AI-driven CV applications offer significant conveniences

but also pose risks; for instance, manipulated images can deceive facial recognition systems, and autopilot features may fail to recognize certain dangers. This has led to the creation of specialized testing frameworks, such as DeepTest, designed to assess and improve the robustness of CV algorithms[64].

Our work, however, distinguishes itself from these studies by introducing MedTest, a tailored method for evaluating medical imaging systems, which incorporates a wide array of MRs designed for realistic clinical perturbations. To our knowledge, the MRs utilized in MedTest have not been previously addressed in the literature. Furthermore, MedTest supports multiple diagnostic tasks such as segmentation, VQA, and classification, making it a versatile tool. Each MR in our framework is derived from real-world clinical scenarios, setting our approach apart from others that may rely on theoretical or unvalidated perturbations. Moreover, unlike most existing research that limits testing to theoretical models, MedTest also evaluates its efficacy on commercial medical imaging software, emphasizing its practical applicability and pioneering nature in the field.

12 Future Work

12.1 Image Synthesis with Generative Adversarial Networks

Since our MedTest in producing MRs mainly involves mathematical-representation-based image transformation and processing techniques, limitations exist on producing large-scale dataset with more variations within each artifact class, especially object perturbations. Due to the scarce data to serve as artifact candidates, our generated samples have specific artifact patterns, which restrict the images from being more nat-

ural. Besides, we cannot simulate the possible large-area presence of artifacts and potential existence on vital areas, such as on polyps, without affecting the original ground truth label. For instance, blood may appear in a contiguous and pervasive manner, but our simulation method only extract small parts from it and cannot produce the same effect as original.

Therefore, generative adversarial networks (GAN) may exhibit its potential in creating a more realistic blending different elements into medical images as we desired. After we surveyed the related work, we have asserted that GANs have exceptional power in generating natural fusion of image contents and styles according to the given images and segmentation label of different instance categories. [13, 19] Similar application in medical images, even in polyp related tasks, have been witnessed with promising performance. Because of this, we plan to explore deeper into this topic and try to generate more realistic images regarding the object perturbations for our customized dataset, so that we can further improve the overall evaluation on our target models.

12.2 Further Testing on Multimodal Large Language Models

The rapid advancements in artificial intelligence have catalyzed the development of numerous LLMs, with multimodality becoming a central focus. As these models become increasingly capable, particularly in performing medical VQA tasks, their potential for deployment in healthcare settings grows. We plan to extend our testing framework to continuously evaluate these emerging multimodal LLMs, ensuring they meet the rigorous demands of medical diagnostics.

13 Conclusion

In this report, we embarked on an in-depth analysis of AI-driven diagnostic tools in medical imaging, with a particular emphasis on endoscopic image diagnosis. The choice to focus initially on this area stems from its critical importance in healthcare. Accurate and reliable medical diagnostics are fundamental to patient care, and the increasing integration of AI tools in this domain necessitates a rigorous evaluation of their performance. Our development of MedTest, a specialized metamorphic testing framework, marks a significant step in this direction, enabling a detailed assessment of these tools under various clinically relevant scenarios.

Through our comprehensive pilot study, we identified and categorized common artifacts that pose challenges to the diagnostic accuracy of these tools. We generated the 9 different types of artifacts on 5 datasets, involving more than $5,000$ images and generated over $40,000$ images with artifacts. Our findings reveal that even SOTA algorithms exhibit varying degrees of performance degradation when faced with these realistic test cases, underscoring the need for continual improvement and rigorous testing.

This study provides valuable insights into the robustness of academic medical image diagnosis software targeting segmentation and classification tasks. Meanwhile, we also endeavored to extend the application of our method to the fast evolving area of multimodal models. Multimodal models, which integrate and interpret data from various modalities, are poised to revolutionize medical diagnostics by offering a more comprehensive analysis than single-modality models. However, the complexity of these models necessitates a nuanced approach to testing and valida-

tion. Combing both parts of focus, we hope to validate our method MedTest and provide a more comprehensive insights into the medical imaging areas.

To this end, our future work will focus on extending the methodologies and lessons learned from our current research further into the realm of multimodal models, as our study still has limitations on assessing the MLLMs, especially in their ability of VQA. Our ultimate goal is to ensure that as these advanced AI tools become integral to medical diagnostics, they do so with the highest standards of accuracy and reliability, thus enhancing patient outcomes and advancing healthcare services.

In conclusion, this report not only sheds light on the vulnerabilities of current medical image diagnosis software but also lays the groundwork for future explorations into the broader domain of AI-driven diagnostic tools, including multimodal models. As we continue to push the boundaries of AI in healthcare, rigorous testing and continual improvement of these tools will be paramount to fully realizing their potential in improving patient care.

References

[1] Ali, S., Zhou, F., Bailey, A., Braden, B., East, J., Lu, X., and Rittscher, J. A deep learning framework for quality assessment and restoration in video endoscopy, 2019.

[2] Ali, S., Zhou, F., Daul, C., Braden, B., Bailey, A., Realdon, S., East, J., Wagnières, G., Loschenov, V., Grisan, E., Blondel, W., and Rittscher, J. Endoscopy artifact detection (ead 2019) challenge dataset, 2019.

[3] Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M., and Asari, V. K. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *ArXiv abs/1802.06955* (2018).

[4] Asyrofi, M. H., Yang, Z., Shi, J., Quan, C. W., and Lo, D. Can differential testing improve automatic speech recognition systems? *2021 IEEE International Conference on Software Maintenance and Evolution (ICSME)* (2021), 674–678.

[5] Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., and Vilariño, F. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics 43* (2015), 99–111.

[6] Brody, H. Medical imaging. *Nature 502*, 7473 (2013), S81–S81.

[7] Carlini, N., Mishra, P., Vaidya, T., Zhang, Y., Sherr, M. E., Shields, C., Wagner, D. A., and Zhou, W. Hidden voice commands. In *USENIX Security Symposium* (2016).

[8] Chen, S., Jin, S., and Xie, X. Testing your question answering software via asking recursively. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)* (2021), IEEE, pp. 104–116.

[9] Chen, S., Jin, S., and Xie, X. Validation on machine reading comprehension software without annotated labels: A property-based method. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (2021), pp. 590–602.

[10] Chen, T. Y., Cheung, S. C., and Yiu, S.-M. Metamorphic testing: A new approach for generating next test cases. *ArXiv abs/2002.12543* (2020).

[11] Chen, T. Y., Ho, J. W. K., Liu, H., and Xie, X. An innovative approach for testing bioinformatics programs using metamorphic testing. *BMC Bioinformatics 10* (2008), 24 – 24.

[12] Chen, Y., Juttukonda, M., Su, Y., Benzinger, T., Rubin, B. G., Lee, Y. Z., Lin, W., Shen, D., Lalush, D., and An, H. Probabilistic air segmentation and sparse regression estimated pseudo ct for pet/mr attenuation correction. *Radiology 275*, 2 (2015), 562–569.

[13] Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 8789–8797.

[14] Chornyy, R. Artificial intelligence in healthcare: Market size, growth, and trends. `https://binariks.com/blog/artificial-intelligence-ai-healthcare-market/`, 2023. Accessed: 2023-11-01.

[15] Dray, X., Li, C., Saurin, J.-C., Cholet, F., Rahmi, G., Le Mouel, J., Leandri, C., Lecleire, S., Amiot, X., Delvaux, J.-M., et al. Cad-cap: une base de données française à vocation internationale, pour le développement et la validation d'outils de diagnostic assisté par ordinateur en vidéocapsule endoscopique du grêle. *Endoscopy 50*, 03 (2018), 000441.

[16] Dwarakanath, A., Ahuja, M., Sikand, S., Rao, R. M., Bose, R. P. J. C., Dubash, N., and Podder, S. Identifying implementation bugs in machine learning based image classifiers using metamorphic testing. *Proceedings of the 27th ACM SIGSOFT International Symposium on Software Testing and Analysis* (2018).

[17] Fan, D.-P., Ji, G.-P., Zhou, T., Chen, G., Fu, H., Shen, J., and Shao, L. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention* (2020), Springer, pp. 263–273.

[18] Gao, X., Saha, R. K., Prasad, M. R., and Roychoudhury, A. Fuzz testing based data augmentation to improve robustness of deep neural networks. *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)* (2020), 1147–1158.

[19] Günther, E., Gong, R., and Van Gool, L. Style adaptive semantic image editing with transformers. In *European Conference on Computer Vision* (2022), Springer, pp. 187–203.

[20] Guo, J., Zhang, Z., Zhang, L., Xu, L., Chen, B., Chen, E., and Luo, W. Towards variable-length textual adversarial attacks. *arXiv preprint arXiv:2104.08139* (2021).

[21] Gupta, S., He, P., Meister, C., and Su, Z. Machine translation testing via pathological invariance. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (2020), pp. 863–875.

[22] Hall, K. K., Shoemaker-Hunt, S., Hoffman, L., Richard, S., Gall, E., Schoyer, E., Costar, D., Gale, B., Schiff, G., Miller, K., et al. Making healthcare safer iii: a critical analysis of existing and emerging patient safety practices.

[23] He, P., Meister, C., and Su, Z. Structure-invariant testing for machine translation. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering* (2020), pp. 961–973.

[24] He, P., Meister, C., and Su, Z. Testing machine translation via referential transparency. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)* (2021), IEEE, pp. 410–422.

[25] Huang, W., Sun, Y., Zhao, X.-E., Sharp, J., Ruan, W., Meng, J., and Huang, X. Coverage-guided testing for recurrent neural networks. *IEEE Transactions on Reliability* (2021).

[26] Humbatova, N., Jahangirova, G., and Tonella, P. Deepcrime: mutation testing of deep learning systems based on real faults. *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis* (2021).

[27] Ionescu, B., Müller, H., Drăgulinescu, A., Yim, W., Ben Abacha, A., Snider, N., Adams, G., Yetisgen, M., Rückert, J., García Seco de Herrera, A., Friedrich, C. M., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Hicks, S. A., Riegler, M. A., Thambawita, V., Storås, A., Halvorsen, P., Papachrysos, N., Schöler, J., Jha, D., Andrei, A., Radzhabov, A., Coman, I., Kovalev, V., Stan, A., Ioannidis, G., Manguinhas, H., Ştefan, L., Constantin, M. G., Dogariu, M., Deshayes, J., and Popescu, A. Overview of ImageCLEF 2023: Multimedia Retrieval in Medical, Social Media, and Recommender Systems Applications. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction* (Thessaloniki, Greece, September 18-21 2023), Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023), Springer Lecture Notes in Computer Science (LNCS), p. To be updated.

[28] Jha, D., Ali, S., Emanuelsen, K., Hicks, S. A., VajiraThambawita, Garcia-Ceja, E., Riegler, M. A., de Lange, T., Schmidt, P. T., Johansen, H. D., Johansen, D., and

Halvorsen, P. Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy, 2020.

[29] Jha, D., Riegler, M., Johansen, D., Halvorsen, P., and Johansen, H. D. Doubleu-net: A deep convolutional neural network for medical image segmentation. *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)* (2020), 558–564.

[30] Jha, D., Smedsrud, P. H., Riegler, M. A., Halvorsen, P., de Lange, T., Johansen, D., and Johansen, H. D. Kvasir-seg: A segmented polyp dataset, 2019.

[31] Jia, R., Raghunathan, A., Göksel, K., and Liang, P. Certified robustness to adversarial word substitutions. *arXiv preprint arXiv:1909.00986* (2019).

[32] Jie, Z., Zhiying, Z., and Li, L. A meta-analysis of watson for oncology in clinical application. *Scientific Reports 11* (2021).

[33] Jin, D., Jin, Z., Zhou, J. T., and Szolovits, P. Is bert really robust? a strong baseline for natural language attack on text classification and entailment, 2020.

[34] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. Segment anything, 2023.

[35] Koulaouzidis, A., Iakovidis, D. K., Yung, D. E., Rondonotti, E., Kopylov, U., Plevris, J. N., Toth, E., Eliakim, A., Johansson, G. W., Marlicz, W., et al. Kid project: an internet-based digital video atlas of capsule endoscopy for research purposes. *Endoscopy international open 5*, 06 (2017), E477–E483.

[36] Levin, S. Tesla fatal crash: 'autopilot' mode sped up car before driver killed, report finds [online]. `https://www.theguardian.com/technology/2018/jun/07/tesla-fatal-crash-silicon-valley-autopilot-mode-report`, 2018. Accessed: 2018-06.

[37] Li, D., Zhang, Y., Peng, H., Chen, L., Brockett, C., Sun, M.-T., and Dolan, B. Contextualized perturbation for textual adversarial attack. *arXiv preprint arXiv:2009.07502* (2020).

[38] Lin, Y., Wu, J., Xiao, G., Guo, J., Chen, G., and Ma, J. Bsca-net: Bit slicing context attention network for polyp segmentation. *Pattern Recognition 132* (2022), 108917.

[39] Litjens, G. J. S., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J., van Ginneken, B., and Sánchez, C. I. A survey on deep learning in medical image analysis. *Medical image analysis 42* (2017).

[40] Luo, Y., Meghjani, M., Ho, Q. H., Hsu, D., and Rus, D. Interactive planning for autonomous urban driving in adversarial scenarios. *2021 IEEE International Conference on Robotics and Automation (ICRA)* (2021), 5261–5267.

[41] Ma, S., Liu, Y., Lee, W.-C., Zhang, X., and Grama, A. Y. Mode: automated neural network model debugging via state differential analysis and input selection. *Proceedings of the*

2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (2018).

[42] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *ICLR abs/1706.06083* (2018).

[43] Makary, M. A. Medical error—the third leading cause of death in the us. `https://www.bmj.com/content/353/bmj.i2139`, 2016. Accessed: 2023-11-01.

[44] Merative. Ibm watson health's product has been equipped in nine of the top 10 us hospitals and decreased 32% in ed visits by high utilizers. `https://www.merative.com/`, 2023. Accessed: 2023-11-01.

[45] Newman-Toker, D. E., Nassery, N., Schaffer, A., Yu-Moe, C. W., Clemens, G. D., Wang, Z., Zhu, Y., Tehrani, A. S. S., Fanai, M., Hassoon, A., and Siegal, D. Burden of serious harms from diagnostic error in the usa. *BMJ Quality & Safety* (2023).

[46] OpenAI. Gpt-4 technical report, 2024.

[47] Pei, K., Cao, Y., Yang, J., and Jana, S. S. Deepxplore: Automated whitebox testing of deep learning systems. *Proceedings of the 26th Symposium on Operating Systems Principles* (2017).

[48] Pham, H. V., Kim, M., Tan, L., Yu, Y., and Nagappan, N. Deviate: A deep learning variance testing framework. *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)* (2021), 1286–1290.

[49] Rajan, R. E., and Prasadh, D. K. Spatial and hierarchical feature extraction based on sift for medical images. *International Journal of Computer Engineering & Technology (IJCET) 3*, 2 (2012), 308–322.

[50] Reid, D. Google's deepmind a.i. beats doctors in breast cancer screening trial. `https://www.cnbc.com/2020/01/02/googles-deepmind-ai-beats-doctors-in-breast-cancer-screening-trial.html`, 2020. Accessed: 2023-11-01.

[51] Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), pp. 4902–4912.

[52] Riccio, V., Jahangirova, G., Stocco, A., Humbatova, N., Weiss, M., and Tonella, P. Testing machine learning based systems: a systematic mapping. *Empir. Softw. Eng. 25* (2020), 5193–5254.

[53] S., A., F., Z., and et al., B. B. An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. *Scientific Reports 10* (2020), 2748.

[54] Schmidhuber, J. Deep learning in neural networks: An overview. *Neural networks 61* (2015), 85–117.

[55] Segura, S., Fraser, G., Sánchez, A. B., and Cortés, A. R. A survey on metamorphic testing. *IEEE Transactions on Software Engineering 42* (2016), 805–824.

[56] Shao, Y., Gao, Y., Guo, Y., Shi, Y., Yang, X., and Shen, D. Hierarchical lung field segmentation with joint shape and appearance sparse learning. *IEEE transactions on medical imaging 33*, 9 (2014), 1761–1780.

[57] Shen, Z., Fu, H., Shen, J., and Shao, L. Modeling and enhancing low-quality retinal fundus images. *IEEE transactions on medical imaging 40*, 3 (2020), 996–1006.

[58] Suk, H.-I., Lee, S.-W., Shen, D., and Initiative, A. D. N. Deep sparse multi-task learning for feature selection in alzheimer's disease diagnosis. *Brain Structure and Function 221* (2016), 2569–2587.

[59] Sun, Z., Zhang, J. M., Harman, M., Papadakis, M., and Zhang, L. Automatic testing and improvement of machine translation. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering* (2020), pp. 974–985.

[60] Tajbakhsh, N., Gurudu, S. R., and Liang, J. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging 35*, 2 (2015), 630–644.

[61] Tang, C., Dong, Y., and Su, X. Automatic registration based on improved sift for medical microscopic sequence images. *2008 Second International Symposium on Intelligent Information Technology Application 1* (2008), 580–583.

[62] Tao, G., Ma, S., Liu, Y., Xu, Q., and Zhang, X. Trader: Trace divergence analysis and embedding regulation for debugging recurrent neural networks. *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)* (2020), 986–998.

[63] Team, G. Gemini: A family of highly capable multimodal models, 2024.

[64] Tian, Y., Pei, K., Jana, S., and Ray, B. Deeptest: Automated testing of deep-neural-network-driven autonomous cars. In *Proceedings of the 40th international conference on software engineering* (2018), pp. 303–314.

[65] Tian, Y., Pei, K., Jana, S. S., and Ray, B. Deeptest: Automated testing of deep-neural-network-driven autonomous cars. *2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE)* (2017), 303–314.

[66] Tomar, N. K., Jha, D., Bagci, U., and Ali, S. Tganet: Text-guided attention for improved polyp segmentation, 2022.

[67] Tu, J., Li, H., Yan, X., Ren, M., Chen, Y., Liang, M., Bitar, E., Yumer, E., and Urtasun, R. Exploring adversarial robustness of multi-sensor perception systems in self driving. *ArXiv abs/2101.06784* (2021).

[68] Vázquez, D., Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., López, A. M., Romero, A., Drozdzal, M., and Courville, A. A benchmark for endoluminal scene segmentation of colonoscopy images, 2016.

[69] Wang, J., Chen, J., Sun, Y., Ma, X., Wang, D., Sun, J., and Cheng, P. Robot: Robustness-oriented testing for deep learning systems. *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)* (2021), 300–311.

[70] Wang, J., Huang, Q., Tang, F., Meng, J., Su, J., and Song, S. Stepwise feature fusion: Local guides global, 2022.

[71] Wang, L., Chen, K. C., Gao, Y., Shi, F., Liao, S., Li, G., Shen, S. G., Yan, J., Lee, P. K., Chow, B., et al. Automated bone segmentation from dental cbct images using patch-based sparse representation and convex optimization. *Medical physics 41*, 4 (2014), 043503.

[72] Wang, W., Huang, J., Chen, C., Gu, J., Zhang, J., Wu, W., He, P., and Lyu, M. R. Validating multimedia content moderation software via semantic fusion. *ArXiv abs/2305.13623* (2023).

[73] Wang, W., tse Huang, J., Wu, W., Zhang, J., Huang, Y., Li, S., He, P., and Lyu, M. R. Mttm: Metamorphic testing for textual content moderation software. *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)* (2023), 2387–2399.

[74] Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S. K., and Cui, S. Shallow attention network for polyp segmentation, 2021.

[75] Xie, X., Ho, J. W. K., Murphy, C., Kaiser, G. E., Xu, B., and Chen, T. Y. Testing and validating machine learning classifiers by metamorphic testing. *The Journal of systems and software* (2011).

[76] Xing, X., Yuan, Y., and Meng, M. Q.-H. Zoom in lesions for better diagnosis: Attention guided deformation network for wce image classification. *IEEE Transactions on Medical Imaging 39*, 12 (2020), 4047–4059.

[77] Zang, Y., Qi, F., Yang, C., Liu, Z., Zhang, M., Liu, Q., and Sun, M. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), pp. 6066–6080.

[78] Zhang, J., Harman, M., Ma, L., and Liu, Y. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering 48* (2022), 1–36.

[79] Zhang, J., tse Huang, J., Wang, W., Li, Y., Wu, W., Wang, X., Su, Y., and Lyu, M. R. Improving the transferability of adversarial samples by path-augmented method. *ArXiv abs/2303.15735* (2023).

[80] Zhang, J., Wu, W., tse Huang, J., Huang, Y., Wang, W., Su, Y., and Lyu, M. R. Improving adversarial transferability via neuron attribution-based attacks. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 14973–14982.

[81] Zhang, M., Zhang, Y., Zhang, L., Liu, C., and Khurshid, S. Deeproad: Gan-based metamorphic testing and input validation framework for autonomous driving systems. *2018 33rd IEEE/ACM International Conference on Automated Software Engineering (ASE)* (2018).

[82] Zhu, M., Chen, Z., and Yuan, Y. Dsi-net: Deep synergistic interaction network for joint classification and segmentation with endoscope images. *IEEE Transactions on Medical Imaging 40*, 12 (2021), 3315–3325.

[83] Ziegler, C. A google self-driving car caused a crash for the first time. [online]. `https://www.theverge.com/2016/2/29/11134344/google-self-driving-car-crash-report`, 2016. Accessed: 2016-09.

[84] Zou, K. H., Warfield, S. K., Bharatha, A., Tempany, C. M., Kaus, M. R., Haker, S. J., Wells, W. M., Jolesz, F. A., and Kikinis, R. Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports. *Academic Radiology 11*, 2 (2004), 178–189.