Enhancing Adverse Drug Event Detection with Multimodal Dataset: Corpus Creation and Model Development

Pranab Sahoo[1], **Ayush Kumar Singh**[1], **Sriparna Saha**[1], **Aman Chadha**[2,3*], **Samrat Mondal**[1]
[1]Department of Computer Science And Engineering, Indian Institute of Technology Patna
[2]Stanford University, [3]Amazon GenAI
pranab_2021cs25@iitp.ac.in, ayush_2211ai27@iitp.ac.in, sriparna@iitp.ac.in, samrat@iitp.ac.in
hi@aman.ai

Abstract

The mining of adverse drug events (ADEs) is pivotal in pharmacovigilance, enhancing patient safety by identifying potential risks associated with medications, facilitating early detection of adverse events, and guiding regulatory decision-making. Traditional ADE detection methods are reliable but slow, not easily adaptable to large-scale operations, and offer limited information. With the exponential increase in data sources like social media content, biomedical literature, and Electronic Medical Records (EMR), extracting relevant ADE-related information from these unstructured texts is imperative. Previous ADE mining studies have focused on text-based methodologies, overlooking visual cues, limiting contextual comprehension, and hindering accurate interpretation. To address this gap, we present a MultiModal Adverse Drug Event (MMADE) detection dataset, merging ADE-related textual information with visual aids. Additionally, we introduce a framework that leverages the capabilities of LLMs and VLMs for ADE detection by generating detailed descriptions of medical images depicting ADEs, aiding healthcare professionals in visually identifying adverse events. Using our MMADE dataset, we showcase the significance of integrating visual cues from images to enhance overall performance. This approach holds promise for patient safety, ADE awareness, and healthcare accessibility, paving the way for further exploration in personalized healthcare. The code and dataset used in this work are publicly available [1].

Disclaimer: The article features images that may be visually disturbing to some readers.

1 Introduction

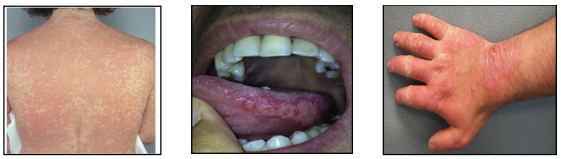
An adverse drug event (ADE) encompasses any harm resulting from medication use, whether it's unintended, off-label, or due to medication errors. Adverse drug reactions (ADRs) are a specific type of ADE, denoting unexpected harm arising from the proper use of medication at the prescribed dosage. Injuries from inappropriate or off-label use are not classified as ADRs (Karimi et al., 2015b). ADEs pose significant public health concerns, contributing to numerous fatalities, serious injuries, millions of hospitalizations, and prolonged hospital stays. Consequently, they impose substantial financial burdens, costing healthcare systems billions of dollars globally. Despite advancements in healthcare, ADE detection remains a significant challenge. Implementing effective detection and monitoring strategies can substantially mitigate the adverse impacts on patients and healthcare systems (Hakkarainen et al., 2012), (Yadav et al., 2018b), (Sultana et al., 2013). Most of the previous ADE detection works are based on text data only (D'Oosterlinck et al., 2023), (Sarker and Gonzalez, 2015), (Yadav et al., 2019), (Sarker et al., 2016), (Chowdhury et al., 2018), (Yadav et al., 2018a), which presents a significant disadvantage due to its subjective nature and lack of specific details of visual cues, leading to potential inaccuracies and incomplete categorization of ADE detection. Despite extensive research, the potential of integrating textual data with visual information, such as images, has been largely overlooked. Visual aids are essential in ADE detection for numerous reasons. A substantial proportion of the population lacks proficiency in medical jargon, hindering accurate symptom descriptions. Moreover, certain symptoms are inherently challenging to express through text alone. Patients may struggle to differentiate between similar symptoms, like skin rash, eczema, peeling, and blister. As depicted in Fig. 1, sample images may present confusion to individuals lacking adequate medical expertise. Integrating both text and images in these scenarios can enhance the accuracy and effectiveness of ADE detection,

*Work does not relate to position at Amazon.
[1]https://github.com/singhayush27/MMADE.git

Findings of the Association for Computational Linguistics ACL 2024, pages 11214–11226
August 11-16, 2024

offering a comprehensive understanding of the patient's current medical condition. To the best of our knowledge, ADE detection using both image and text data has not been explored previously, and we take this opportunity to introduce a MultiModal Adverse Drug Event (MMADE) dataset comprising ADR images paired with corresponding textual descriptions.

Large Language Models (LLMs) and Vision Language Models (VLMs) have exhibited remarkable skills in generating human-like text, prompting their integration into various medical applications, including tasks such as chest radiography report generation, summarization, and medical question answering (Thawkar et al., 2023), (Ghosh et al., 2024b), (Sahoo et al., 2024b), (Ghosh et al., 2024a). However, their potential in ADE detection, which involves both text and images, has yet to be explored. Leveraging LLMs and VLMs for this task presents inherent limitations as they are predominantly trained on generic natural images sourced from databases like ImageNet, Wikipedia, and the internet. Generic models may not possess the specialized medical knowledge required for comprehensive caption generation, potentially leading to oversimplified descriptions that overlook essential details like symptoms and medical intricacies. Furthermore, while VLMs have excelled in traditional visual-linguistic tasks, their application to medical imaging presents unique challenges that may hinder the accurate interpretation and description of complex medical images (Sahoo et al., 2024a). Specialized models such as XrayGPT (Thawkar et al., 2023) and SkinGPT4 (Zhou and Gao, 2023), which are trained on chest X-ray and skin disease images, exemplify the domain specificity required for accurate medical image analysis. This has led us to explore ADR detection within a multimodal framework. To support this exploration, we introduce MMADE, a carefully curated dataset crafted for this specific purpose. MMADE consists of 1500 instances of patient-reported concerns regarding drugs and associated side effects, each paired with both textual descriptions and corresponding images. In our study, we have employed InstructBLIP (Dai et al., 2023), which builds upon the strong foundation of BLIP-2 (Li et al., 2023), a pre-trained model with high-quality visual representation and strong language generation capabilities. The meticulous fine-tuning process enables it to bridge the disparity between general-purpose models and the specialized demands of ADE-specific tasks. Moreover, our exploration of BLIP (Li et al., 2022) and GIT (Wang et al., 2022) reveals that these models exhibit insufficient performance before fine-tuning. Nevertheless, upon fine-tuning with domain-specific data, their performance experiences notable improvement.



After 1 hour of taking the penicillins!! I get this.

I found this after my doctor prescribed Etoricoxib for post-surgical dental pain.

This is very unusual to me after taking amoxicillin for pharyngitis.

Figure 1: Samples from the dataset highlight the significance of visual cues in understanding adverse drug events, particularly in cases where patients are unaware of a specific medical condition.

Our key contributions are as follows:

- A novel approach to ADE detection in multimodal settings greatly assists medical professionals, such as doctors, nurses, and pharmacists, by delivering detailed descriptions of ADE cases, enhancing precision in diagnosis, treatment planning, and patient care.
- Introduction of a novel multimodal dataset MMADE for further research on ADE detection area.
- The proposed dataset demonstrates promising potential for various applications, including ADE classification, caption generation, and summarization tasks.
- We have utilized InstructBLIP and experimented with two other pre-trained VLMs, and reported a detailed analysis.
- The ADE-specific model holds promise for enhancing patient safety, ADE awareness, and healthcare communication. It aims to provide individuals seeking information about ADEs with understandable and informative captions accompanying medical images to improve their comprehension of potential medication risks.

2 Related Works

This section details the related ADE detection tasks based on the data sources.

Works Based on Biomedical Text and Electronic Medical Record: Various techniques have been developed for extracting ADEs from Electronic Medical Records (EMRs) (Aramaki et al., 2010), (Wang et al., 2009), as well as from medical case reports (MCRs) (Gurulingappa et al., 2011). Gurulingappa et al. (2012a) utilized machine learning methods to identify and extract potential ADE relations from MEDLINE case reports. Unlike random data sources such as social media, both EMRs and MCRs offer significant advantages by providing comprehensive records of a patient's medical history, treatment, conditions, and potential risk factors. Moreover, these records are not limited to patients who have experienced ADRs (Harpaz et al., 2013). Sarker and Gonzalez (2015) conducted a study by taking data from MEDLINE case reports and Twitter and reported how combining different datasets increases the performance of identifying ADRs. Huynh et al. (2016) explored various neural network frameworks for ADE classification, utilizing datasets from both MCRs and Twitter. DISAE is another corpus (Gurulingappa et al., 2010), which consists of 400 MEDLINE articles with annotations for disease and adverse effect names without drug-related information. D'Oosterlinck et al. (2023) introduced the BioDEX dataset for biomedical ADE extraction in real-world pharmacovigilance. This dataset comprises 65,000 abstracts, 19,000 full-text biomedical papers, and 256,000 document-level safety reports crafted by medical professionals. However, to the best of our knowledge, there is currently no publicly accessible annotated multimodal (Image and Text) corpus suitable for identifying drug-related adverse effects.

Works Based on Social Media Datasets: Social media has become vital for accessing vast amounts of real-time information, making it valuable for identifying potential ADEs. Leaman et al. (2010) conducted a pioneering study that analyzed user comments from social media posts, comprising a dataset of 6890 comments. The research demonstrated the significant value of user comments in identifying ADEs, highlighting their crucial role in this context. Several other authors (Gurulingappa et al., 2012b), (Yadav et al., 2020), (Benton et al., 2011) employed lexicon-based approaches to extract ADEs. However, these methods are limited to a specific set of target ADEs. Nikfarjam and Gonzalez (2011) employed a rule-based technique instead of a naive lexicon-based approach on the same dataset, enabling the detection of ADEs not covered by lexicons. Several authors utilized supervised machine learning techniques like Support Vector Machines (SVM) (Sarker and Gonzalez, 2015), Conditional Random Fields (CRF) (Nikfarjam et al., 2015), and Random Forests (Zhang et al., 2016) for ADE detection. Sarker et al. (2016) introduced one corpus by collecting data from social media, focusing on adverse drug reactions. Tasks included automatically classifying user posts, extracting specific mentions, and normalizing mentions to standardized concepts. With the availability of annotated data, in recent times, the rise of deep learning techniques has significantly influenced research methodologies, leading to the adoption of deep learning models for predicting ADEs. Tutubalina et al. (2017) explored the synergy between CRF and Recurrent Neural Networks (RNN), demonstrating that CRF enhances the RNN model's ability to capture contextual information effectively. Chowdhury et al. (2018) developed a multi-task architecture that simultaneously tackled binary classification, ADR labeling, and indication labeling, using the PSB 2016 Social Media dataset (Sarker et al., 2016).

Table 1: Keywords used for the data collection.

Keywords: ADE, ADR, adverse drug reaction, adverse drug event, adverse reaction, adverse drug event reporting, side effects, drug reactions, drug side effects, type of infection and reaction, medicine, drugs, skin rashes, red patches, eczema, ulcer, acne, skin irritation, edema, rosacea, alopecia, lip swelling.

3 Corpus Development

Our study began with a thorough literature review to identify existing ADE-related datasets. We discovered four text-only datasets: PSB 2016 social media shared task (Sarker et al., 2016) comprising 572 tweets, Medline ADE corpus (Gurulingappa et al., 2012b) with 4,272 sentences, CADEC (Karimi et al., 2015a) containing 1,248 sentences, and recently released BioDEX dataset (D'Oosterlinck et al., 2023). This revealed a notable gap in multimodal ADE datasets, where images complement textual data. We take this opportunity to introduce a multimodal corpus consist-

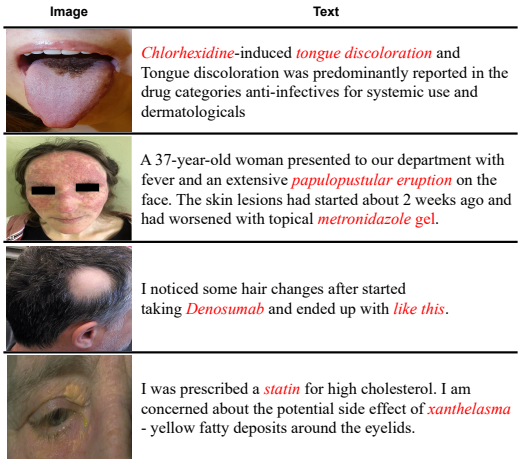


Figure 2: Some samples from the dataset containing images and the corresponding descriptive text.

ing of 1,500 ADE images with corresponding English sentences to facilitate further research. While preparing this corpus, we carry out the following steps.

3.1 Data Collection

Utilizing a diverse array of keywords, we have curated a comprehensive dataset from social media, healthcare blogs, and MCRs (refer to Table 1). The inclusion of various sources ensures a broad representation of the population, enriching the dataset's diversity.

3.1.1 Social Media

Social media data is invaluable for ADE-related tasks due to its real-time nature and diverse user-generated content (Sarker et al., 2016). In this study, we have utilized the official X (Twitter) and scraper API to gather Tweets[2],[3] employing diverse keywords related to ADE. The data collection phase, conducted between June 2023 and October 2023, collected a total of 20,000 tweets using specified keywords presented in Table 1. From this pool, 3,000 tweets were meticulously identified as pertinent to ADEs, featuring either images, text, or a combination of both. Notably, 142 tweets included relevant images accompanied by textual descriptions of the adverse drug events.

[2] https://twitter.com/
[3] https://www.scraperapi.com/

3.1.2 Healthcare Blog

We utilized a public healthcare-related blog, healthdirect[4], a government-funded virtual health service that provides access to health advice and information via a website. We used Python's BeautifulSoup library to scrape the data and collected 1,150 unique images with corresponding text. Among these, 54 relevant images depicting adverse drug events were manually curated along with corresponding texts.

3.1.3 Medical Case Reports

Data from published medical case reports is crucial for constructing comprehensive datasets to analyze ADE. These articles provide structured and verified information, forming a reliable foundation for in-depth analysis and research in pharmacovigilance. We have extracted data from the New England Journal of Medicine[5], Science Direct[6]. A precise Science Direct query performed is as follows:

```
(("adverse drug event") AND
(Languages=English) AND (Article
type=Case Reports) AND
(Years=2000 to 2023))
```

This approach retrieved approximately 2,907 documents from ScienceDirect, from which we manually selected 1390 relevant images with corresponding texts.

3.2 Data Annotation

To ensure meticulous annotation aligned with ethical standards, we enlisted the assistance of two medical students and one Ph.D. student, selected based on specific criteria. These criteria included being at least 25 years old, proficient in English (reading, writing, and speaking), and willing to handle sensitive content. The process was finalized within a span of five months, and participants received compensation for their involvement [7]. Annotators were tasked with meticulously assessing each image and its corresponding text based on the annotation manual. Sentences accurately depicting adverse drug events, including the drug's

[4] https://www.healthdirect.gov.au/
[5] https://www.nejm.org
[6] https://www.sciencedirect.com
[7] The medical students received compensation in the form of gift vouchers and honorarium amounts in accordance with https://www.minimum-wage.org/international/india.

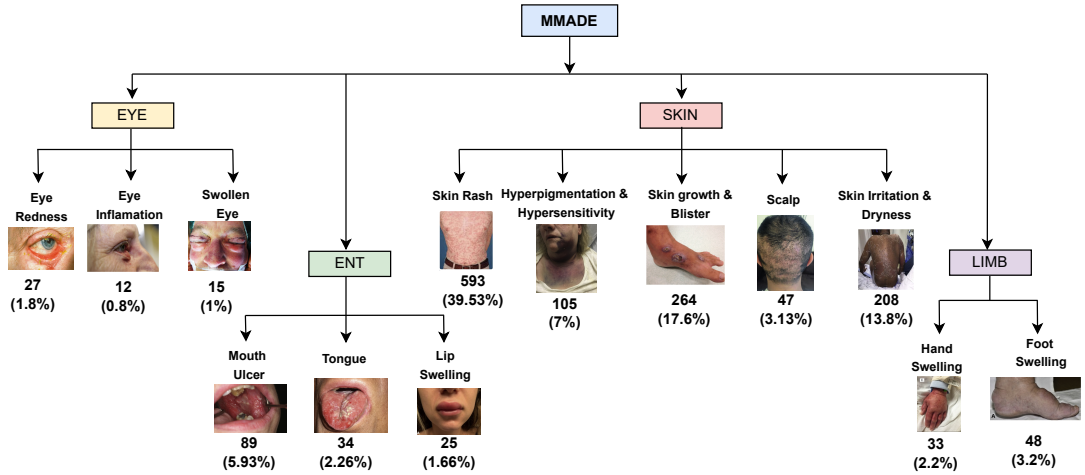


Figure 3: Distribution of different body parts in the curated MMADE dataset. The number of data points and the percentage corresponding to each category have been provided.

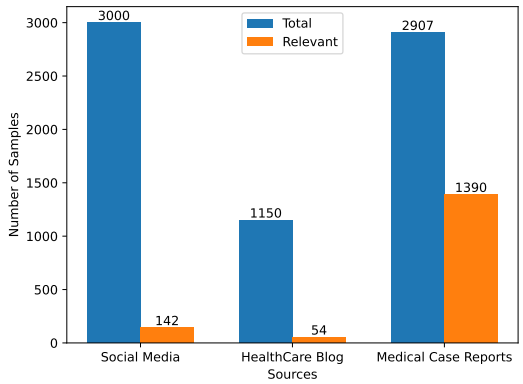


Figure 4: Data distribution statistics from various sources, illustrating the total ADE data (images, text, and image-text pairs), and the relevant image-text pairs.

name and associated side effects, were chosen for inclusion in the corpus development process, while all other instances were deemed irrelevant and removed. To maintain consistency and consensus among annotators, final rationale labels were determined through a majority voting approach. Annotators were explicitly instructed to annotate posts without biases related to demographics, religion, or other extraneous factors. The quality of annotations was assessed by measuring inter-annotator agreement (IAA) using Cohen's Kappa score (Viera et al., 2005). The resulting agreement scores of 0.78 affirm the acceptability and high quality of the annotations.

3.3 Annotation Manual

We initially provided annotators with an instruction manual detailing different instances, accompanied by examples (as depicted in Fig. 5).

- For each data instance comprising text and an image, select the data instance if both the text and image indicate concerns about drug side effects.
- Remove data instances where the image does not convey the side effects mentioned in the text.
- Remove data instances where the text does not convey any side effects related to drugs, but some side effects are visually present in the accompanying image.
- Remove instances where both the text and image are unrelated to any drug side effect.
- If a data instance includes a URL link, access the content at that URL address to gain additional insight and context about the data.

3.4 Corpus Analysis

Figure 4 provides an overview of the data sources and the distribution of relevant cases within the dataset. Following the initial collection of 7,057 ADE-related samples encompassing images, text, and image-text pairs, we curated 1,500 pertinent samples. These encompass both image and their associated text descriptions of ADEs, establishing the foundation of our multimodal ADE dataset. Figure 2 depicts some of

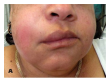
Image	Text	A1	A2	A3	MV	Label
	There have been reported cases of hyperpigmentation associated with benzodiazepines like clonazepam, as observed during patient examinations.	1	1	1	1	1
	After just one week of using the medication, a woman noticed the emergence of dark patches on a part of her face, which may be attributed to a potential side effect of tetracyclines.	1	0	0	0	0
	Challenges in treatment of disseminated nocardiosis in an elderly patient with renal failure on corticosteroids	0	1	0	0	0
	Case Study of Intrapartum Antibiotic Prophylaxis and Subsequent Postpartum Beta-Lactam Anaphylaxis	0	0	0	0	0

Figure 5: Illustration of an annotation process, with A1, A2, and A3 representing individual annotators and MV signifying final majority voting. In this setup, "1" denotes ADR, while "0" is unrelated to ADR.

the samples from the dataset. The first sample shows a person suffering from tongue discoloration after taking a Chlorhexidine drug. Both the text and visual image show a clear indication of drug reaction. The second sample is of a woman with papulopustular eruption on the face, which had worsened with topical metronidazole gel. Adverse drug events can manifest internally or externally, yet acquiring images of internal body parts from public sources is challenging. Thus, we focus on external body parts or symptoms, which can be readily captured and shared with doctors, pharmacovigilance teams, or the public domain for improved consultation and advice. Following dataset analysis, we identified 13 significant adverse effects crucial for multimodal ADE reporting. These effects are categorized into four groups based on their origin: ENT (9.85%), EYE (3.6%), LIMB (5.4%), and SKIN (81.06%). Additional details of the dataset, such as the distribution of samples across different body parts along with corresponding percentages, are illustrated in Fig. 3.

4 Problem Formulation

Each data point in the dataset encompasses a patient's textual description T along with a corresponding image I illustrating their medical issue or concern visually. The textual representation comprises a sequence of words $\{t_1, t_2, \ldots, t_n\}$, while the visual elements are represented by $I \in \mathbb{R}^{3\times W\times H}$, where W and H denote the width and height of the image data, respectively. The objective is to process both the text T and image I for each patient and generate a natural language sequence Y that seamlessly integrates both modalities, expressed as $Y = \{T, I\}$.

5 Methodology

Recent advancements in VLMs, such as BLIP, InstructBlip, and GIT, have showcased remarkable advancements in encoding both textual and visual inputs. These models outperform traditional approaches that rely solely on individual image or text encoders, followed by fusion. By integrating sophisticated mechanisms for joint representation learning, VLMs excel at capturing intricate relationships between textual and visual modalities, thereby enhancing their ability to generate more contextually relevant and coherent outputs (Zhang et al., 2023). In the proposed work, we have leveraged InstructBlip (Dai et al., 2023), known for its exceptional performance across a range of vision-language (VL) tasks including Visual Question Answering (VQA), Image captioning, and Image retrieval. Each patient's inquiry or concern is articulated as a textual sentence along with a visual image where they elaborate on their medical queries or concerns on social media platforms or healthcare blogs to obtain pertinent feedback or advice. InstructBlip integrates two distinct encoders specialized for separate modalities. This architecture includes an image encoder, which utilizes a vision transformer (ViT) to extract visual features, alongside an LLM and a Query Transformer (Q-Former). The Q-Former interacts with the image encoder's output through cross-attention, resulting in K-encoded visual vectors, which are then linearly projected and fed into the frozen LLM for further processing. This representation serves as input for a proficient language model, which generates high-quality text. During instruction tuning, only the Q-Former undergoes fine-tuning, while the image encoder and LLM remain unchanged. Figure 6 illustrates the detailed architecture. We fine-tuned the InstructBLIP to assess its performance on the proposed MMADE dataset. InstructBLIP maintains a consistent image resolution (224×224) during instruction tuning and freezes the visual encoder during fine-tuning. This approach substantially reduces the number of trainable parameters from 1.2 billion to 188 million, improving fine-tuning efficiency. We also employed two more VLMs and fine-tuned them with the proposed MMADE dataset. BLIP (Li et al., 2022), a

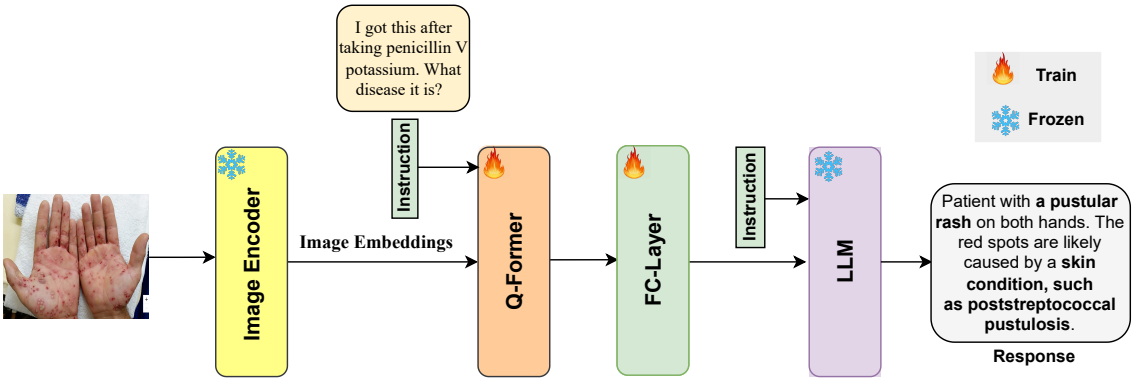


Figure 6: The architectural details are illustrated, depicting the input image and corresponding text. Additionally, the frozen and trainable layers are highlighted for clarity.

versatile vision-language model, utilizes a multimodal mixture of encoder-decoder models during pre-training. This involves bootstrapping a dataset from large-scale noisy image-text pairs, where synthetic captions are injected, and noisy captions are eliminated. Additionally, we utilized GIT (Wang et al., 2022), a VLM model that generates text descriptions of images. Trained on curated datasets of image-text pairs, GIT encodes image features into a latent representation, which is decoded into text descriptions. Its architecture comprises an image encoder using a pre-trained Swin transformer and a text decoder based on a standard transformer decoder, linked by a cross-attention layer for enhanced focus on specific image encoding parts. Additionally, we performed the integration of LSTM networks with VGG16 (Simonyan and Zisserman, 2014) and ResNet50 (He et al., 2016) architectures. In this setup, the LSTM serves as the text encoder, while either VGG16 or ResNet50 acts as the visual encoder. The features extracted from the text and visual encoders are then concatenated to create a joint representation, allowing for comprehensive modeling of both textual and visual information. Please refer to Section 12 of the Appendix for the fine-tuning details.

6 Experimental Results and Analysis

We have utilized four commonly employed evaluation metrics to assess the performance of the models, including BLEU score (Bilingual Evaluation Understudy) (Papineni et al., 2002), ROUGE score (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004), BERTScore (Zhang et al., 2019), and MoverScore (Zhao et al., 2019). Detailed explanations of experimental settings and metrics are given in Section 11 and 13 of the Appendix.

Table 2: The performance evaluation of the models using ROUGE and BLEU Scores in the multimodal dataset setting.

Model	Type	ROUGE			BLEU		
		R1	R2	RL	B1	B2	B3
LSTM+VGG16	-	0.213	0.105	0.201	0.165	0.073	0.041
LSTM+ResNet50	-	0.281	0.086	0.230	0.179	0.058	0.046
BLIP	Base	0.19	0.093	0.185	0.099	0.003	0.001
	Fine-Tune	0.334	0.163	0.225	0.171	0.081	0.058
GIT	Base	0.27	0.11	0.192	0.157	0.014	0.004
	Fine-Tune	0.504	0.285	0.416	0.194	0.10	0.097
InstructBLIP	Base	0.29	0.161	0.212	0.219	0.125	0.008
	Fine-Tune	**0.571**	**0.351**	**0.475**	**0.319**	**0.175**	**0.112**

Table 3: The evaluation of the model using BERTScore and MoverScore in the multimodal dataset setting.

BERTScore								
Evaluation Metrics	LSTM+ VGG16	LSTM+ ResNet50	BLIP		GIT		InstructBLIP	
			Base	Fine-Tune	Base	Fine-Tune	Base	Fine-Tune
Precision	0.821	0.847	0.819	0.841	0.826	0.866	0.832	**0.896**
Recall	0.797	0.811	0.783	0.819	0.791	0.831	0.781	**0.891**
F1-score	0.809	0.821	0.801	0.829	0.812	0.852	0.805	**0.893**
MoverScore								
	0.441	0.487	0.482	0.513	0.496	0.553	0.544	**0.622**

6.1 Findings from Experiments

The experimental results across Tables 2 and 3 depict the performance of various models in multimodal dataset settings, while Tables 4 and 5 represent the performance in unimodal dataset settings, based on standard evaluation metrics. From the

Table 4: The evaluation of the models using BERTScore and MoverScore in the unimodal dataset setting.

BERTScore					
Evaluation Metrics	LSTM+VGG16	LSTM+ResNet50	BLIP	GIT	InstructBLIP
Precision	0.731	0.744	0.753	0.801	0.840
Recall	0.709	0.713	0.724	0.783	0.783
F1-score	0.719	0.729	0.738	0.790	0.810
MoverScore					
	0.121	0.143	0.151	0.184	0.253

results, several observations are:

- Table 2 presents the ROUGE and BLEU scores, indicating the superior performance of fine-tuned InstructBLIP compared to other models. The higher ROUGE and BLEU scores achieved by fine-tuned InstructBLIP suggest its proficiency in capturing relevant information from the input and generating text. This superiority can be attributed to its training on more extensive and diverse datasets, enabling it to capture intricate data patterns effectively.

- Tables 3 and 4 present BERTScores and MoverScores in the multimodal and unimodal settings, respectively. Fine-tuned InstructBLIP achieves the highest BERTScore and MoverScores, demonstrating its efficacy in capturing contextual similarity and superior ability to convey meaningful content effectively.

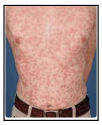
- Table 4 and 5 show the decline in performance when training VLM models with unimodal data. The absence of meaningful visual information likely contributes to the performance degradation.

- One important observation is that InstructBLIP consistently outperforms BLIP and GIT across all metrics, showing its superior capability in effectively integrating textual and visual information. This can be attributed to its innovative architecture, featuring a Query Transformer for instruction-aware feature extraction. Refer to Section 16 of the Appendix for detailed outcomes.

- Fine-tuning with domain-specific ADE data remarkably enhances model performance, reflecting its pivotal role in adapting models to the intricacies of adverse drug event detection.

- Another key finding highlights the substantial performance enhancement achieved by integrating both image and text modalities, emphasizing the critical role of visual information alongside textual data.

Statistical Analysis: We conducted a statistical t-test to compare the performance of the proposed multimodal model, utilizing both image and text data, with that of unimodal models. The analysis yielded a p-value below 0.05, indicating a significant difference in performance between the two. The detailed explanation is given in Section 14 of Appendix.



Actual: I was given penicillin for my ongoing treatment. About 4 hours later, I started feeling really sick. I came down with a high fever, nausea, vomiting, and a terrible headache. My whole body ached, and got this type of rash on my skin. What type of rash it is?

Model	Unimodal (Blank Image and Text)	Multimodal (Image and Text)
BLIP	A man with a large stomach and a large stomach.	A close up of a person's stomach covered in a red rash.
GIT	A middle aged man sits on an exam table in a doctor's office, looking unwell.	A man with a lot of red spots suffering from rashes.
InstructBLIP	A person's covered in a red rash.	The image shows red spots on the abdomen, which is unusual for a person. **The red spots are likely caused by a skin condition, such as syphilis.**

Figure 7: An example case study where a sample image and its corresponding text are provided along with the generated text from all models in both unimodal and multimodal settings.

6.2 Qualitative Analysis

We performed an extensive qualitative analysis of the responses generated by various models in both unimodal and multimodal settings, complemented by several case studies. A detailed case study is depicted in Fig. 7. The analysis has led to the following conclusion: (a) In multimodal settings, all the models demonstrate superior performance and exhibit a greater ability to capture crucial visual information conveyed through images than in unimodal settings (refer to Table 4 and 5). (b) Observations also revealed that models such as BLIP and GIT tended to hallucinate, as evidenced by Fig. 7, occasionally generating facts that were entirely unrelated to the context. (c) InstructBLIP demonstrates superior performance, leading us to conclude that providing instructions during fine-tuning prompts the model to selectively focus on pertinent visual features. This focused attention encourages the model to generate target sequences that closely resemble the desired output. We have presented a case study demonstrating that model performance is not solely determined by the number of samples in the dataset but also by other factors such as the distinct visual characteristics of different types of ADEs present in the dataset. A detailed explanation is provided in the Section 15 of Appendix.

Table 5: The performance evaluation of the models using ROUGE and BLEU Scores in the unimodal dataset setting.

Model	ROUGE			BLEU		
	R1	R2	RL	B1	B2	B3
LSTM+VGG16	0.133	0.018	0.101	0.141	0.012	0.001
LSTM+ResNet50	0.146	0.029	0.112	0.152	0.016	0.002
BLIP	0.158	0.032	0.129	0.161	0.019	0.002
GIT	0.169	0.035	0.137	0.172	0.020	0.003
InstructBLIP	0.211	0.087	0.172	0.197	0.089	0.006

7 Risk Analysis

While our multimodal model demonstrates promise, it is essential to have medical experts and pharmacovigilance teams validate the findings, considering other critical factors. Our model and dataset are intended to support medical professionals rather than replace them.

8 Conclusion and Future Work

In this paper, we present the task of ADE detection within pharmacovigilance mining, leveraging multimodal datasets. In order to solve this task, we have created a multimodal ADE dataset, MMADE, containing images and corresponding descriptions, enhancing decision-making with the inclusion of visual cues. We have employed InstructBLIP, fine-tuned with the proposed dataset, and compared it with other models. Our findings suggest that domain-specific fine-tuning significantly enhances overall performance, emphasizing the importance of multimodal visual cues. We envision MMADE as a pivotal resource for advancing research in multimodal ADE detection. Moreover, our fine-tuned architecture holds promise as a valuable tool for pharmacovigilance teams, clinicians, and researchers, facilitating more effective ADE monitoring and ultimately improving patient safety and outcomes. In addition to expanding the dataset, future investigations could explore the potential of this multimodal dataset in tasks such as ADE severity classification and summarization.

9 Limitations

While our effort aimed to develop an ADE detection framework and introduce the novel MMADE dataset, comprising textual descriptions of drug events paired with images, it is crucial to acknowledge certain limitations inherent in the dataset. Specifically, our dataset primarily focuses on drug events associated with external body parts, omitting data about internal conditions such as liver infections, kidney stones, or psychological ailments like depression and migraine. Acquiring a substantial volume of image-text pairs within the ADE domain presents inherent challenges, including data privacy concerns, regulatory constraints, and the specialized nature of ADE occurrences. Despite these obstacles, our research breaks new ground by integrating images with text, solving a real-world challenge where individuals affected by ADE may resort to image sharing for communication when verbally expressing their symptoms is difficult. Moving forward, we aim to enhance the dataset by incorporating more ADE-related images and expanding its utility through additional tasks such as complaint identification and text summarization.

10 Ethics and Broader Impact

User Privacy.

Our dataset contains AD images and corresponding text tweets with annotation labels and no personal user information.

Biases.

Any biases detected in the dataset are inadvertent, and we have no intention of harming anyone or any group. We acknowledge that evaluating whether a tweet is ADE can be subjective, so we have taken agreement from all the annotators before selecting the data.

Intended Use.

We share our data to promote more research on Adverse drug event detection. We only release the dataset for research purposes and do not grant a license for commercial use.

References

Eiji Aramaki, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Masuichi, Kayo Waki, and Kazuhiko Ohe. 2010. Extraction of adverse drug effects from clinical records. In *MEDINFO 2010*, pages 739–743. IOS Press.

Adrian Benton, Lyle Ungar, Shawndra Hill, Sean Hennessy, Jun Mao, Annie Chung, Charles E Leonard, and John H Holmes. 2011. Identifying potential adverse effects using the web: A new approach to medical hypothesis generation. *Journal of biomedical informatics*, 44(6):989–996.

Shaika Chowdhury, Chenwei Zhang, and Philip S Yu. 2018. Multi-task pharmacovigilance mining from

social media posts. In *Proceedings of the 2018 World Wide Web Conference*, pages 117–126.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Karel D'Oosterlinck, François Remy, Johannes Deleu, Thomas Demeester, Chris Develder, Klim Zaporojets, Aneiss Ghodsi, Simon Ellershaw, Jack Collins, and Christopher Potts. 2023. Biodex: Large-scale biomedical adverse drug event extraction for real-world pharmacovigilance. *arXiv preprint arXiv:2305.13395*.

Akash Ghosh, Arkadeep Acharya, Raghav Jain, Sriparna Saha, Aman Chadha, and Setu Sinha. 2024a. Clipsyntel: CLIP and LLM synergy for multimodal question summarization in healthcare. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 22031–22039. AAAI Press.

Akash Ghosh, Arkadeep Acharya, Prince Jha, Sriparna Saha, Aniket Gaudgaul, Rajdeep Majumdar, Aman Chadha, Raghav Jain, Setu Sinha, and Shivani Agarwal. 2024b. Medsumm: A multimodal approach to summarizing code-mixed hindi-english clinical queries. In *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part V*, volume 14612 of *Lecture Notes in Computer Science*, pages 106–120. Springer.

Harsha Gurulingappa, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2011. Identification of adverse drug event assertive sentences in medical case reports. In *First international workshop on knowledge discovery and health care management (KD-HCM), European conference on machine learning and principles and practice of knowledge discovery in databases (ECML PKDD)*, pages 16–27.

Harsha Gurulingappa, Roman Klinger, Martin Hofmann-Apitius, and Juliane Fluck. 2010. An empirical evaluation of resources for the identification of diseases and adverse effects in biomedical literature. In *2nd Workshop on Building and evaluating resources for biomedical text mining (7th edition of the Language Resources and Evaluation Conference)*, pages 15–22.

Harsha Gurulingappa, Abdul Mateen-Rajpu, and Luca Toldo. 2012a. Extraction of potential adverse drug events from medical case reports. *Journal of biomedical semantics*, 3(1):1–10.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012b. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45(5):885–892.

Katja M Hakkarainen, Khadidja Hedna, Max Petzold, and Staffan Hägg. 2012. Percentage of patients with preventable adverse drug reactions and preventability of adverse drug reactions–a meta-analysis. *PloS one*, 7(3):e33236.

Rave Harpaz, Santiago Vilar, William DuMouchel, Hojjat Salmasian, Krystl Haerian, Nigam H Shah, Herbert S Chase, and Carol Friedman. 2013. Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. *Journal of the American Medical Informatics Association*, 20(3):413–419.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Trung Huynh, Yulan He, Alistair Willis, and Stefan Rueger. 2016. Adverse drug reaction classification with deep neural networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 877–887, Osaka, Japan. The COLING 2016 Organizing Committee.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015a. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81.

Sarvnaz Karimi, Chen Wang, Alejandro Metke-Jimenez, Raj Gaire, and Cecile Paris. 2015b. Text and data mining techniques in adverse drug reaction detection. *ACM Computing Surveys (CSUR)*, 47(4):1–39.

Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang, and Graciela Gonzalez. 2010. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts in health-related social networks. In *Proceedings of the 2010 workshop on biomedical natural language processing*, pages 117–125.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Azadeh Nikfarjam and Graciela H Gonzalez. 2011. Pattern mining for extraction of mentions of adverse drug reactions from user comments. In *AMIA annual symposium proceedings*, volume 2011, page 1019. American Medical Informatics Association.

Azadeh Nikfarjam, Abeed Sarker, Karen O'connor, Rachel Ginn, and Graciela Gonzalez. 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024a. Unveiling hallucination in text, image, video, and audio foundation models: A comprehensive survey. *arXiv preprint arXiv:2405.09589*.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024b. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Abeed Sarker and Graciela Gonzalez. 2015. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics*, 53:196–207.

Abeed Sarker, Azadeh Nikfarjam, and Graciela Gonzalez. 2016. Social media mining shared task workshop. In *Biocomputing 2016: Proceedings of the Pacific Symposium*, pages 581–592. World Scientific.

Ryoma Sato, Makoto Yamada, and Hisashi Kashima. 2022. Re-evaluating word mover's distance. In *International Conference on Machine Learning*, pages 19231–19249. PMLR.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Janet Sultana, Paola Cutroneo, and Gianluca Trifirò. 2013. Clinical and economic burden of adverse drug reactions. *Journal of Pharmacology and Pharmacotherapeutics*, 4(1_suppl):S73–S77.

Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. 2023. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*.

Elena Tutubalina, Sergey Nikolenko, et al. 2017. Combination of deep recurrent neural networks and conditional random fields for extracting adverse drug reactions from user reviews. *Journal of healthcare engineering*, 2017.

Anthony J Viera, Joanne M Garrett, et al. 2005. Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5):360–363.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.

Xiaoyan Wang, George Hripcsak, Marianthi Markatou, and Carol Friedman. 2009. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *Journal of the American Medical Informatics Association*, 16(3):328–337.

Shweta Yadav, Asif Ekbal, and Sriparna Saha. 2018a. Feature selection for entity extraction from multiple biomedical corpora: A pso-based approach. *Soft Computing*, 22(20):6881–6904.

Shweta Yadav, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya. 2019. A unified multi-task adversarial learning framework for pharmacovigilance mining. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5234–5245, Florence, Italy. Association for Computational Linguistics.

Shweta Yadav, Asif Ekbal, Sriparna Saha, Pushpak Bhattacharyya, and Amit Sheth. 2018b. Multi-task learning framework for mining crowd intelligence towards clinical treatment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 271–277, New Orleans, Louisiana. Association for Computational Linguistics.

Shweta Yadav, Srivatsa Ramesh, Sriparna Saha, and Asif Ekbal. 2020. Relation extraction from biomedical and clinical text: Unified multitask learning framework. *IEEE/ACM transactions on computational biology and bioinformatics*, 19(2):1105–1116.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2023. Vision-language models for vision tasks: A survey.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhifei Zhang, JY Nie, and Xuyao Zhang. 2016. An ensemble method for binary classification of adverse drug reactions from social media. In *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*, volume 1.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.

Juexiao Zhou and Xin Gao. 2023. Skingpt: A dermatology diagnostic system with vision large language model. *arXiv preprint arXiv:2304.10691*.

Appendix

11 Experimental Settings

This section provides the hyperparameters and experimental setups utilized in the study. All experiments were conducted using multiple RTX 2080Ti GPUs. The dataset was partitioned, allocating 80% for training and 20% for testing. For InstructBLIP, the learning rate was set to 1e-5, executed for 50 epochs, with a batch size of 2. Similarly, for the BLIP model, a learning rate of 0.0001 was utilized and executed for 50 epochs, with a batch size of 2. For GIT fine-tuning, a learning rate of 5e-3 was applied for 50 epochs, with a batch size of 2. All models were implemented using Scikit-Learn [8] and PyTorch as the backend framework [9].

12 Fine-tuning

We have followed several steps to fine-tune BLIP on our multimodal dataset. First, we have prepared the dataset in JSON format, which is compatible with the BLIP framework, and each image-text pair is represented as a dictionary with the following keys: image_path: The path to the image file, text: The caption or other text description of the image. We have fine-tuned BLIP with `learning rate = 0.001`, `number of epochs = 50`, and `batch size = 16`. The training process takes 6 hours for the fine-tuning, and we evaluate the performance on a held-out test set.

To fine-tune the GIT model on the proposed MMADE dataset, we utilized the Hugging Face Transformers library and followed a systematic process. First, we prepared the data using a PyTorch dataset, converting it into the required format using the `GitProcessor` class. Finally, we have utilized the GIT-base model, which is pre-trained on a substantial dataset of image-text pairs. We have utilized Parameters such as a `learning rate = 5e-3`, `epochs = 30`, and `batch size = 2`. The optimization was performed using the Adam optimizer with the cross-entropy loss function, and the fine-tuning process took 3 hours.

To fine-tune the CNN-LSTM model, we leverage the CNN models (VGG16 and ResNet50) to extract image features, followed by an LSTM for sequence generation. Model compilation involves the use of categorical cross-entropy loss and the Adam optimizer. Training proceeds for 40 epochs, utilizing a batch size of 32.

[8] https://scikit-learn.org/stable/
[9] https://pytorch.org/

13 Evaluation Metrics

We have utilized BLEU score (Papineni et al., 2002), ROUGE score (Lin, 2004), BERTScore (Zhang et al., 2019) and MoverScore (Zhao et al., 2019). BLEU score is utilized to assess the quality of machine-generated text by comparing it to human-generated reference text. It measures the similarity in word sequences between the machine-generated and human reference texts using n-grams, penalizing shorter machine-generated texts to provide a quantitative measure of translation accuracy. The ROUGE score is also used to evaluate the quality of machine-generated summaries compared to human-written summaries. It works by calculating the overlap of n-grams between the machine-generated summary and the reference summaries. ROUGE and BLEU metrics evaluate text quality based on syntactic overlap, considering unigrams and bigrams, lacking the ability to decode semantic meaning effectively. However, BERTScore focuses on understanding the semantic meaning of generated text compared to the intended text, enabling a more nuanced and accurate comparison. MoverScore is a way to measure how similar a machine-generated text is to a human-written text. It does this by using BERT embeddings to understand the meaning of the words and sentences in both texts and then using Word Mover's Distance (Sato et al., 2022) to measure how similar the two texts are.

14 Statistical Analysis

In this study, we employed the paired t-test to assess the statistical significance of differences between the outcomes of unimodal and multimodal models. The null hypothesis (H0) assumes no significant disparity in scores, while the alternative hypothesis (H1) proposes the opposite. We structured our analysis around the assumption of paired data, where each BLEU score corresponds to the same model in different settings (Multimodal Vs. Unimodal). By applying the paired t-test, we aimed to rigorously evaluate performance disparities and offer insights into the models' relative effectiveness. The obtained p-values for ROUGE, BLEU and BERTScore scores were 0.008, 0.0097 and

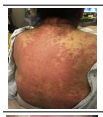
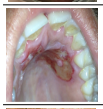
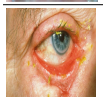
Image	Actual Text	Generated Text
	I have got these red, swollen bumps and patches all over my back, popping up about 13 days after starting vemurafenib and cobimetinib. What type of disease is it?	The image shows a woman who has a **large rash** on her back. This is unusual as it is not common for people to have such a significant rash on their back, the size and intensity of the rash, which appears to be inflamed and red, **possibly indicating an allergic reaction or a skin condition.**
	I have been experiencing oral issues since I took etoricoxib. There's erosion on the roof of my mouth with this weird filmy layer and a fuzzy border. What is this?	In this image, it appears unusual inside the mouth. Looks like a person having **a mouth ulcer.**
	I've been taking medication to manage my high cholesterol levels, and the doctor prescribed statins as the main option. Now I have this. What type of problem it is?	The image is unusual because it shows a close-up view of an eye, specifically the iris and some other parts of the eye related to the man's eye. The appearance of the eye is also **different from a normal, healthy eye, which typically is free of any marks.**

Figure 8: Illustration of a case study showcasing the performance of the InstructBLIP model from the dataset

0.0316, respectively, both below the 0.05 threshold, leading to the rejection of the null hypothesis.

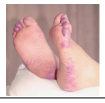
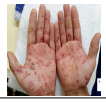
Image		
Actual Text	Twelve hours after urgent coronary angiography, the appearance of this patient's feet had changed.	A 47-year-old man presented to the emergency department with a 3-day history of a pustular rash on both hands. One week before presentation, he had started treatment with penicillin V potassium for pharyngitis.
BLIP	A person with a pinkish foot and a white background.	A person holding their hands with red spots on them.
Fine-tune BLIP	A noticeable transformation in this patient's feet is evident.	A case of pustular rash with a man on hands.
GIT	The purple spots on the feet of a woman.	The hands of my husband, with a little bit of red on the fingers.
Fine-tune GIT	The patient's feet had turned blue and mottled.	Man with a pustular rash on hands.
InstructBLIP	The image shows a person's foot with a purple rash on it, which is unusual for a photo of a person's feet. The rash appears to be caused by a fungal infection, and the person's feet appear to be infected with the fungus.	The image shows a pair of hands with red spots on them, which is unusual for a person's hands. The red spots are likely caused by a skin condition, such as eczema or psoriasis.
Fine-tune InstructBLIP	The image displays a person's feet, with covered in blue and purple spots which is unusual for a person.	The image depicts the hands of a person who has experienced a rash or spots on their hands, featuring small red bumps and dots. The skin appears red and slightly irritated, possibly due to an allergic reaction or other cause of inflammation.

Figure 9: Example of the sample image and corresponding text from the dataset along with generated text from the Proposed model with two other models.

15 Case study

The case study depicted in Fig. 8 reveals intriguing insights across various body parts. In the first row, depicting rashes, fine-tuned InstructBLIP accurately identifies distinct body parts and key findings, potentially attributed to the diverse visual characteristics of different rash types prevalent in the dataset with the largest portion of 81.06%. However, performance in identifying mouth-related ADEs is less satisfactory despite comprising 5.93% of the dataset, likely due to potential confusion with other oral features like the tongue, teeth, or lips. Conversely, despite eye-related problems representing only 1.8% of the dataset, the model performs comparatively better in this category, possibly because it focuses specifically on infected regions, enhancing its ability to identify relevant features accurately.

16 Comparative output

We have added two more examples in Fig. 9 showing the ADR instances, corresponding ground truth, and the model-generated outputs.