International Research Journal of Multidisciplinary Scope (IRJMS), 2026; 7(2): 793-811

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2026.v07i02.08290

Automated Medical Image Captioning Using Vision Transformer and Generative Pre-trained Transformer-2

Moloy Dhar[1], Mrinmoy Sen[2], Bidesh Chakraborty[2], Suparna Biswas[3]*, Shubhajit Chatterjee[1]

[1]Deptartment of Computer Science and Engineering, Guru Nanak Institute of Technology, Kolkata, West Bengal, India, [2]Department of Computer Science and Engineering, Haldia Institute of Technology, Haldia, West Bengal, India, [3]Department of Electronics and Communication Engineering, Guru Nanak Institute of Technology, Kolkata, West Bengal, India. Corresponding Author's Email:suparna.biswas@gnit.ac.in

Abstract

This paper anticipate a deep learning (DL)-located framework for medical image captioning (IC) utilizing a transformer-located model. The system is anticipated like it integrates two important parts, a Vision Transformer (VT) for processing ocular data and fetching features from images, and a decoder based on Generative Pre-trained Transformer 2 (GPT-2), that transform these characteristics into understandable, contextually significant natural language depictions. These two works composed to aid automatic creation of textual reports matching medical images. To ensure output quality and relevance, training is conducted on a carefully selected dataset that includes matched medical images and clinical descriptions. The model's performance was evaluated in this work using the IU X-Ray dataset, a publicly accessible and extensively used benchmark in the medical imaging community. Preliminary diagnostic reporting should be generated automatically, which will reduce labor costs, increase workflow effectiveness, shorten reporting times, and improve consistency with medical records. The transformer-based approach manages long-range dependencies in both the picture and text domains, in contrast to conventional CNN–RNN designs.Better diagnostic interpretability results from more context-aware and semantically rich medical image captions.In this research, we offer an end-to-end medical picture captioning model that is totally transformer-based and does away with recurrent components like LSTM.The paper's main goal is to make diagnostic reporting easier for radiologists and other medical practitioners. Based on popular evaluation measures including BLEU, METEOR, and ROUGE, the model showed competitive results (98.42% accuracy), suggesting significant potential for practical use.

Keywords: CNN, GPT-2, Medical Imaging, NLP, VT, X-Ray Dataset.

Introduction

The swift évolutions of medical imaging data in contemporary healthcare systems has highlighted the pressing need for intelligent equipment that can help doctors accurately describe and record complex ocular data. The traditional method of creating standard reports takes a lot of time and is prone to inconsistencies since experts' competence levels vary. These risks are particularly noticeable in environments with limited resources or high demand, where skilled radiologists might not always be available. The development of automated systems capable of producing accurate and therapeutically relevant representations of medical pictures has gained significant traction in response to these worries. Medical IC, an integrative field that combines the skills of computer vision (CV) and natural language processing (NLP) to transform visual data into structured, written description, is one promising path in this area. Medical IC requires a sophisticated grasp of anatomical and pathological aspects, which challenges both accuracy and domain-specific knowledge, in contrast to general IC, which focuses on illuminating frequently occurring scenarios or items. In this framework, our work recommends a novel composite transformer-located model tailored for IC in medicine. In order to provide articulate, contextually grounded textual outputs, the proposed paradigm combines language transformers with visual transformers to retrieve comprehensive imaging data. Subject to considerable assistance in the earlier phases of report preparation, these outputs are intended to mirror the format and content of clinical radiology reports. The aforementioned system has a great

(Received 19th September 2025; Accepted 06th March 2026; Published 13th April 2026)

deal of promise for practical use, including the ability to automatically generate initial diagnostic notes, assist with second-opinion inspection in telemedicine, and provide diagnostic support in low-resource environments. By combining the power of DL and transformer architectures, our research goals to provide the growing landscape of clinically viable AI tools in radiology and medical diagnostics. Our paper's contribution is as follows:

a) Our suggested approach builds a hybrid transformer-based model tailored for medical IC of VT and GPT-2.
b) After being retrieved, the features are concatenated and put into an encoding-decoding architecture based on GPT-2 that auto regressively generates a textual description of the image, taking into account both the projected visual embeddings and previously generated tokens.
c) With an accuracy of 98.42%, our hybrid model—which combines the VT and GPT-2 models performs better than any other model that has been used with this hybrid architecture.
d) The ultimate model accomplish superior when it comes to generating detailed and contextually appropriate depiction for input photos.

Medical IC, the task of automatically producing descriptive text from medical images, has attracted more and more research interest because of its potential to improve clinical workflows and lessen the burden of diagnosis. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which have historically dominated the area, are giving way to more potent and adaptable transformer-based models. Many studies have attempted IC with encoder-decoder (E-D) models based on CNNs and RNNs. The concept was first presented in early publications, which established the groundwork by fusing RNN-based phrase production with CNN-based picture feature extraction (1). Subsequent developments, such as the model focused on attention mechanisms and allowed permissive models to adaptively target prominent areas of an image while creating a caption (2).

CNN-RNN architectures were the focus of early experiments in medical IC, where RNNs (often LSTM) generated textual descriptions that matched the visual properties that CNNs had retrieved. Similar works created this paradigm (3). Later, similar models were adapted for use in the medical industry in works such as (4). Hierarchical LSTM models were used to control the length and complexity of radiology reports.

More recent methods make use of transformer-based structures, which result in notable advancements in language modeling and visual comprehension. By interpreting picture patches as sequential tokens that are linked to natural language, the VT presented in accordance with transformed image encoding and allowed the model to recognize general relationships throughout the image (5). In comparison to typical CNNs, VT has shown comparable or superior performance on a variety of visual tasks.

After the Transformer design proved successful in NLP, subsequent models moved beyond RNNs for sequence generation (6). For lengthy clinical reports, transformers provide parallelism, attention-based modeling, and enhanced long-range dependency handling.

GPT-2, created by OpenAI, is a very powerful transformer-based paradigm for producing natural language that is fluent, coherent, and contextually rich (7). Its pre-training on a large corpus of varied online content makes it quite good at language modeling tasks, such as creating captions. An important development in visual processing was the VT (8). Unlike CNNs, VT allows for richer and more global representations by processing picture patches as tokens and using self-attention methods. When used in the medical field for tasks like segmentation and classification, VT has demonstrated exceptional performance in identifying intricate and subtle patterns in radiological images. Because of its versatility, it is a good option for feature extraction in medical IC pipelines.

A large-scale transformer decoder model trained on extensive corpora, GPT-2, is excellent at producing text that is contextually rich and cohesive (9). Although its architecture was initially created for general-purpose language modeling, it has been modified for use in medical NLP assignments like clinical note generation and description. GPT-2 is sufficiently advanced to produce fluid captions from visual embeddings in picture captioning, as opposed to the attention fusion approach or direct feature integration. In the medical field, databases such as MIMIC-CXR

(from the MIMIC-III clinical database) and IU X-Ray (Indiana University Chest X-ray) have developed into benchmarks for analyzing medical IC systems (10). These databases, which are subject to a costly resource for training and evaluating vision-language models in clinical settings, comprise thousands of chest X-ray images in pairs with corresponding radiology dossiers. The reciprocal power of both design is impacted by the integration of VT and GPT-2 in a single pipeline. Here, precision, interpretability, and firmness are vital (11). In order to capitalize on the benefits of both modalities, new IC structures have examined joining VT and GPT-2 (12). VTGPT and BLIP (Bootstrapped Language-Image Pre-training) generate coherent, human-like image descriptions. Medical IC demands high accuracy, terminology, and sensitivity to subtle abnormalities. Researchers use datasets like MIMIC-CXR, IU X-Ray, and Open-AI to train and assess DL models in order to accomplish this goal. These datasets facilitate the methodical assessment of model performance in tasks related to medical imaging. Medical terminology differs greatly between reporting styles and organizations. Robust model generalization is still hampered in accordance with the scarcity of well-annotated data (13).

Unification of VT and GPT-2 is a promising advancement. It could enhance automated medical IC. GPT-2 models text sequences, while VT provides rich global visual representations (14). Although the feasibility of such structures has been established by foundational work in general IC, research on adapting them to medical settings—where precision, fluency, and clinical validity are crucial—is still in its early stages (15, 16). It is showed that applying Bayesian Active Learning (BALD with MC Dropout) to large pre-trained models for sequence labeling is a beneficial approach (17). This method enables annotation-efficient NLP pipelines, which is essential in domain-specific or low-resource settings (e.g., medical text processing). It creates new opportunities for using Bayesian AL to tasks beyond NLP tagging.

MedCAT (Medical Concept Annotation Toolkit), an effective open-source program for linking and annotating medical concepts across different clinical contexts, is proposed (18). To fix to developing clinical data and terminology, the toolkit include unsupervised learning, transfer learning (TL), and constant learning.

Contribute TorchXRayVision, an open-source library which integrate pre-trained deep learning models and a variation of publically accessible chest X-ray datasets into a single, conveniently navigable structure (19). Major obstruction in medical imaging research is addressed by the library, like restricted approach to pre-trained models, reproducibility difficulty, and dataset disintegration. Regulated preprocessing, model sharing, and united evaluation method considerably boost the repeatability and comparison of studies in chest X-ray analysis, as established by TorchXRayVision over complete benchmarking. The toolkit accelerate progress in computer-aided diagnosis, depiction learning, and transfer learning by reducing barrier for clinical researchers and machine learning expert by depending on smooth approach to considered datasets and models. The study moves the further utilize of DL in chest X-ray examination and determination and curative imaging research by implementing TorchXRayVision as a valuable, scalable, and community-driven resource that boost reproducibility and fosters aid.

Furnish the Reinforced Transformer structure for medical picture captioning, which precisely optimizes non-differentiable language assessment metrics as BLEU and CIDEr by fusing the benefits of reinforcement learning (RL) and transformer-based sequence modeling (20). Their method generates captions that are not only syntactically fluid but also clinically more accurate and pertinent, exceed traditional encoder–decoder and attention-located models. According to the study, the transformer's capacity to match appropriate textual depiction with medical image types is enhanced by reinforcement learning that also lessens the extent of errors from maximum likelihood training. An analysis is conducted to determine how deep learning models trained on one dataset perform when tested on other datasets, in order to identify the cross-domain generalization limits of these models for chest X-ray prediction (21). According to the study, models regularly demonstrate important accomplishment drops when enforced to outside datasets, even while they accomplish fine within the training domain. This focus concern with label dispersion shifts, dataset bias, and alternative in imaging protocols. The authors advise contrary to

dependent too much on single-dataset benchmarks and emphasis that powerful abstraction is not forever pointed out by good in-domain accuracy. For automated chest X-ray systems to be redistribute in a clinically trustworthy manner, they backing multi-domain assessment, dataset variety, and accurate model acceptance as required steps.

Align, Attend, and Locate (A^2L), a difference-persuaded attention network intended for identification of problem of chest X-ray with minimal supervision, is proposed in accordance with (22). The model uses attention mechanisms and image–report alignment to efficiently locate disease-related regions with little bounding-box annotation. A^2L surpass conventional infirm supervised techniques in ailment classification and localization, correspondent to exploratory outcomes on thorough chest X-ray datasets. By growing clinically compelling areas, the study demonstrates that contrastive learning with attention boost both the interpretability of model guess and diagnostic correctness. To achieve improved detection of small clinical details, a fine-grained label learning system for automated chest X-ray report generation that makes use of comprehensive disease labels is demonstrated (23). The proposed access improve the facet of generated documents and the precision of ailment detection by assimilation supervised fine-grained supervision, ensuing in characterization which are more clinically understandable than those generated by traditional image-to-text models. The work emphasizes how models can overcome the drawbacks of coarse labeling and inadequate supervision by utilizing richer, fine-grained labels, ultimately closing the gap between automated captioning and reporting in the manner of radiologists.

A method for learning visual–semantic embeddings to report abnormal findings on chest X-rays is demonstrated (24). Their approach jointly processes textual descriptions and image features to build a shared embedding space. This space links clinically meaningful language directly to corresponding visual patterns. As a result, the model captures strong associations between medical terminology and imaging evidence. The resulting radiological reports are more accurate because to this method. It guarantees more precise synchronization between clinical descriptions and medical imaging. Additionally, it improves report retrieval accuracy. The technique matches pertinent clinical records better than traditional methods. The embedding framework successfully catches clinical anomalies, according to experimental data. Additionally, the framework accurately depicts related textual data. This method improves both linguistic quality and diagnostic accuracy. The deep perceptual autoencodé detects diseased areas in medical photos. This identification approach does not require extensive pixel-level annotations. This leads to the successful learning of therapeutically significant visual elements (25). The identification of minute discrepancies is enhanced by reconstructing images in a trained perceptual space. This method more successfully catches minute structural and textural characteristics. This makes it simpler to spot minor but clinically significant differences. Experiments across several imaging modalities show strong inequality localization and labeling. This act holds even with variations in anatomy and disease.

This paper proposes an autonomous system for radiological report production. Complementary visual information is combined through the use of multiview picture fusion. Clinical relevance is strengthened through the integration of medical concept enhancement. This integration improves the correspondence between medical words and visual aspects. As a result, more thorough and precise radiological reports are produced (26). The system incorporates structured medical concepts. Additionally, multiview data is provided. These inputs are used by the system to direct the creation of reports. The produced outputs substantially resemble the descriptions provided by radiologists. The method controls the intricacy of radiological coverage. It prevents the omission of crucial clinical information. Chest X-ray record are used for experiments. Outcomes are superior than traditional image-to-text techniques. Both illness coverage and language quality show improvements.

Use of DL and TL to analyze COVID-19 pneumonia in chest imaging is examined (27). High COVID-19 diagnosis accuracy is achieved through transfer learning using pre-trained CNN models. Even with a small amount of training data, the method works well. To minimize the requirement for substantial data collecting, large pre-trained models are

employed. The approach can detect COVID-19 pneumonia, according to experimental data. Additionally, the model is able to differentiate COVID-19 from other lung conditions. The approach is practically beneficial because of these features. The method facilitates quick patient screening. Additionally, it helps doctors diagnose patients. The study demonstrates the efficacy of DL in conjunction with TL. When time and data are scarce, this approach is particularly helpful during pandemics.

This survey examines the shift away from CNN-RNN-based methods. It clarifie why more recent approaches are being investigated. Multimodal transformer systems are also examined in the survey. These systems integrate text and picture processing. VT-GPT2 hybrid models receive particular attention. These models do well in activities related to medicine. They make it possible to provide reports that are more dependable. Additionally, they facilitate automated medical reporting.

Methodology

In the proposed approach, VT and GPT-2 are combined into a single multimodal pipeline. This pipeline is used for automated medical IC. The main steps are described in this section: language modelling for caption production, feature transformation and alignment, and visual feature extraction. We encode medical images into a rich, contextualized representation by using the VT architecture. In contrast to CNNs, VT splits the input image into non-coinciding, fixed-size patches (such as 16x16 pixels), each of which is sequence tokenized and linearly embedded. After that, these tokens go via several transformer encoder levels for processing. A predetermined resolution (e.g., 224×224) is tested to input medical images, like chest X-rays, before they are shrunk and normalized. A linear layer is utilized to work each of the N flattened patches that make up the image. In order to preserve spatial information, positional embeddings are inserted. A stack of self-attention layers is used to the embedded series in order to discover global contextual relationships between patches. The class token or aggregated attention-pooled features are used to generate the final image representation.

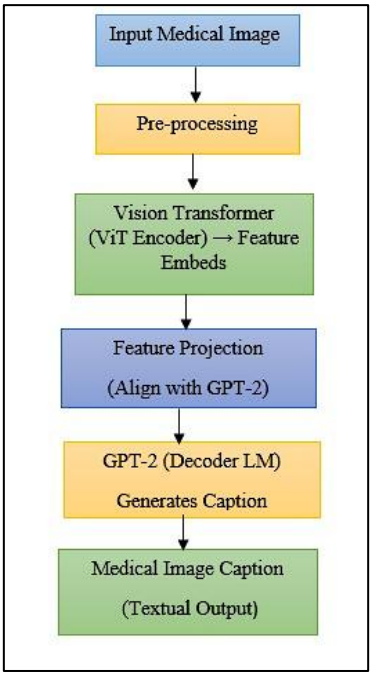

Figure 1: Block Diagram of the Proposed Method

A transformation layer is used to interface the VT output with GPT-2, which requires a language input format. The high-dimensional VT output is mapped into a space that is compatible with the GPT-2 input embeddings by a fully connected (FC) layer. The GPT-2 decoder receives the converted VT characteristics as a prefix prompt. Without changing its fundamental architecture, this allows GPT-2 to condition its text production on the visual surroundings. Medical pictures like X-rays, CT scans, or MRIs are used to start the procedure (Figure 1). These pictures are the system's

unprocessed inputs. Images are transformed into common formats that work with the Vision Transformer, scaled, and normalized. Contrast enhancement, denoising, or patch division are possible extra processes. The objective is to achieve the data available for effective feature eradication.

There are fixed-size patches in the preprocessed image. After being flattened, each patch is incorporated into a vector representation. A Transformer encoder is used to parse these vectors in order to extract contextual and spatial information. The result is a collection of visual feature embeddings that serve as a high-dimensional semantic space depiction of the medical image. Projection (mapping) layer is utilized because the feature embeddings from ViT and the input specifications of GPT-2 are different. In order to use the visual embeddings as context for text production, this stage aligns them with the embedding space of GPT-2. A language model called GPT-2 has been pre-trained on extensive text material and optimized for the creation of medical reports. After that, GPT-2 produces NL descriptions of the medical image that are logical and domain-specific. The final product is a radiology report or textual explanation that highlights the most important findings in the medical image. Through a feature projection layer, the system combines GPT-2 for language creation and Vision Transformer (ViT) for picture understanding.

The unidirectional transformer decoder GPT-2 is in charge of producing captions that are medically correct, coherent, and fluid. The next-token prediction loss (cross-entropy), which aims to produce the next word given the previous sequence and visual prefix, is used to refine the model on paired image-report datasets (e.g., MIMIC-CXR). The Byte Pair Encoding (BPE) tokenizer in GPT-2 is used to tokenize medical information. During inference, descriptive captions that are consistent with clinical findings are produced by generating text using either nucleus sampling or greedy decoding. Large-scale, publicly accessible datasets with medical images and related radiology reports, such as IU X-Ray or MIMIC-CXR, are utilized to train the model. The Adam optimizer with learning proportion scheduling and gradient clipping is used for dependable convergence. Because medical captioning datasets are typically small, dropout and layer normalization are used throughout to avoid overfitting.

By combining VT and GPT-2, our approach enables a trainable end-to-end system that can produce therapeutically meaningful image descriptions. The combined strength of GPT-2's language generation and VT's global view of visual features provide a hybrid structure that is ideal for the intricate requirements of medical picture reporting.

The proposed system is a DL architecture. It is designed to automatically generate clinically relevant and explanatory captions for medical images. It integrates a transformer-located language model (GPT-2) as the text decoder and VT as the picture encoder. Managing the high capacity of text and picture transformers is made easier by combining the two models. The generated captions are more accurate because to this combination. Additionally, the captions are tailored to the context of medical photographs. Every input medical image is divided into fixed-size patches by the VT model. A linear embedding technique is used to embed each patch. The patches are sent into a transformer encoder following embedding. All patches are processed collectively by the encoder. The entire image is represented by the final embedding. It records the patches' two spatial characteristics. Additionally, it records the contextual connections betwixt various visual elements. Important clinical data is included in these embeddings. They offer a thorough depiction of the medical image. Captions are then produced using the embeddings. The data from embeddings is used in the caption generating process. This guarantees the accuracy of the captions. Additionally, it guarantees that the captions have clinical significance. The method facilitates the production of precise and insightful captions. The VT encoder is the first source of image embeddings. The GPT-2 model then processes these embeddings. The embeddings are projected into a format that can be altered. Text synthesis in GPT-2 is guided by the changed embeddings. The model is able to generate cohesive captions thanks to this advice. The generated captions provide a clear description of the content of the image. They are accurate in terms of medicine as well. The Hugging Face Transformers library is used by the system. The Hugging Face Datasets library is also

utilized. Large-scale transformer models can be managed with the aid of these libraries. They offer training and assessment tools. Working with text and image data is made easier by the libraries. Additionally, they facilitate medical picture preparation. These packages make it easier to train huge models. In general, they improve the caption generation system's performance. High modularity is supported by this architecture. The system's many components can be changed as needed. Additionally, content can be improved without altering the system as a whole. Techniques for supervised learning are used in the construction of the system. Medical IC datasets with annotations are used to train it. These datasets include labeled learning examples. During training, cross-entropy loss is employed. Additionally, other loss functions are used. The model is guided to generate better outcomes by these loss functions. They aid in the system's acquisition of precise patterns. The quality of the generated captions has improved. They are precise and unambiguous. All things considered, the design enables reliable caption production.

Additionally, it facilitates flexible caption creation. Transformers are used by the system to analyze text and images simultaneously. This aids in the model's comprehension of complicated data. Medical visuals are better at capturing complex semantics. The technology is able to identify intricate visual characteristics. Features that are minor but significant are identified. As a result, the captions are more pertinent. The pictures are appropriately described in the captions. Reliable and significant results are guaranteed by the design. The generated captions are intimately related to the photos' visual content. Image embeddings are used to direct the text output in order to accomplish this. Because of this, the captions appropriately convey what is depicted in the pictures. Hugging Face's libraries facilitate system testing. Additionally, they make TL possible for better model performance. These libraries also enable the system to be utilized in many therapeutic contexts.

Machine learning libraries are necessary for the proposed medical IC system. For development, it also requires contemporary programming tools. The system is useful and efficient when both are coupled. The system's architecture aims for an E-D design that is transformer-located and tailored for the task of captioning medical images. It is made up of three primary parts:

In terms of language fluency, semantic coherence, and global visual reasoning, the ViT + GPT-2 model performs better than conventional CNN + LSTM frameworks. However, in order to avoid clinically dangerous hallucinations, it requires increased computing power, stringent medical fine-tuning, and grounding measures. For smaller datasets and lower-risk deployments, CNN + LSTM continues to be a lightweight, comprehensible baseline. While CNN + LSTM captions are shorter, simpler, and safer (less chance of hallucinations), ViT + GPT-2 captions are more detailed, grammatically natural, and context-rich (Table 1).

Table 1: Comparative Analysis between ViT + GPT-2 and CNN + LSTM

Aspect	ViT + GPT-2	CNN + LSTM
Model Type	Fully Transformer-based (Encoder–Decoder)	Hybrid CNN (vision) + RNN (language)
Architecture	GPT-2 autoregressively decodes text tokens conditioned on picture embeddings, while ViT encodes images as a series of patch embeddings.	LSTM successively decodes captions from an image feature vector, while CNN retrieves spatial characteristics.
Pretraining	GPT-2 was pretrained on vast text corpora, while ViT was pretrained on big picture datasets (ImageNet) and optionally refined on medical images.	CNN was pretrained on ImageNet, while LSTM was either pretrained on caption data alone or trained from scratch.
Connection	Visual embeddings (prefix or cross-attention) projected into the GPT-2 embedding space	CNN output flattened or pooled → fully connected → initial input to LSTM
Core Principle	Global self-attention for visual context + autoregressive language modeling	Local feature extraction + sequential dependency modeling
Visual comprehension	Uses self-attention to capture long-range dependencies; it works better for complex medical patterns (such diffuse opacities).	Effectively captures regional textures but has trouble with international interactions

Table 2: Comparative Analysis between ViT + GPT-2 and ViT + LSTM

Aspect	ViT + GPT-2	ViT + LSTM
Architecture	Completely transformer-based (encoder and decoder).	Transformer encoder plus recurrent decoder, or hybrid transformer-recurrent.
Encoder	Global self-attention across picture patches using Vision Transformer (ViT).	Same visual encoder is Vision Transformer (ViT).
Decoder	GPT-2 – autoregressive language transformer	LSTM – recurrent neural network for sequential text generation.
Connection Mechanism	Prefix/cross-attention visual embeddings projected into GPT-2 embedding space.	Pooled or linearly projected visual embeddings as the LSTM's initial input or context.
Visual Representation	Identical — both extract contextual, patch-level embeddings.	Similar.

Text Generation	GPT-2 uses global context to produce complex, cohesive, and human-like phrases.	LSTM generates simpler, sequential sentences with local dependencies.
Multimodal Alignment	Robust cross-attention enables collaborative reasoning between textual and visual tokens.	Moderate: LSTM relies on a predetermined visual environment.

While both ViT + LSTM and ViT + GPT-2 make use of the Vision Transformer's robust picture comprehension, their ability to generate language is different. ViT + LSTM is easier to use, quicker, and less prone to hallucinations, but ViT + GPT-2 uses pretrained linguistic knowledge to generate more detailed and fluent reports (Table 2). ViT + LSTM provides a solid baseline for clinical safety and interpretability; when properly adjusted on domain data, ViT + GPT-2 clearly outperforms it for high-quality, human-like reporting.

Table 3: Comparative Analysis between ViT+GPT-2 vs CNN + GPT-2

Aspect	ViT + GPT-2	CNN + GPT-2
Architecture	Completely integrated transformer (ViT encoder + GPT-2 decoder).	Hybrid model (CNN encoder + transformer decoder).
Encoder	Vision Transformer uses self-attention to model global relationships while segmenting images into patches.	Local hierarchical features are extracted by CNN (ResNet, DenseNet, etc.).
Decoder	GPT-2: an autoregressive text generator that makes use of extensive pretraining.	GPT-2 – same pretrained transformer decoder.
Context Capture	Captures global dependencies between far-off areas of an image (helpful for CT and X-rays).	Ignore long-range context in favor of local textures and edges.
Suitability for Medical Images	Excellent at recognizing multi-organ cues, and widespread diseases.	Robust for localized anomalies (nodules, lesions).

Both generate fluent writing, but because of its comprehensive visual attention, ViT + GPT-2 is better at capturing many associated discoveries. Compared to conventional CNN encoders, the Vision Transformer (ViT) encoder greatly improves multimodal alignment and global image understanding when used with GPT-2 as the language decoder in Table 3.

While CNN + GPT-2 provide efficiency, faster training, and robustness on smaller datasets, ViT + GPT-2 yield greater semantic richness, clinical completeness, and interpretability.

ViT + GPT-2 is recommended for high-quality, clinically descriptive image captioning; CNN + GPT-2 is still feasible for lightweight, low-resource implementations.

A VT that has been previously trained is used to process the visual input, such as a chest X-ray. In order to create feature embeddings that capture the image's semantic meaning, this approach splits the image into steady-size patches, which are subsequently embedded and passed via several self-attention layers. The GPT-2 model and the VT encoder's output are not directly compatible. In order to translate the VT's image embeddings into the same dimensional space as the GPT-2's token embeddings, a linear projection layer is added. This change guarantees that the language model may be conditionally trained using the visual characteristics. A GPT-2 decoder, which has been optimized specifically for medical captioning tasks, receives the altered picture attributes. Taking into consideration both the pre-planned ocular embeddings and previously generated tokens, GPT-2 automatically regressively generates a textual representation of the image. The cross-entropy shortfall between the anticipated tokens and the actual ground validity caption tokens is used to create the model. The model learns accurate word forecast in a following manner with the help of this loss.

A learned linear projection layer maps Vision Transformer visual embeddings into the GPT-2 embedding space to enable automated medical image captioning. In multimodal diagnostic reporting workflows in contemporary hospitals, this alignment maintains semantic structure, enabling visual features to effectively condition language generation, support coherent clinical descriptions, and guarantee accurate correspondence between image regions and medically relevant textual terms.

Let I denote a medical image and $V \in R^{n \times dv}$ be the sequence of visual embeddings extracted by the VT. $R^{n \times dv}$ denotes a real-valued matrix with n rows and dv columns, representing n visual tokens each of dimension dv. A linear projection $W \in R^{dv \times dt}$ with bias $b \in R^{dt}$ maps these embeddings into the GPT-2 token embedding space as,

$$Z = VW + b \qquad [1]$$

$R^{dv \times dt}$ denotes a real-valued weight matrix that linearly maps visual embedding dimension dv to text embedding dimension dt. R^{dt} denotes a real-valued vector of length dt, typically representing a bias or a single embedding in the text feature space. The projected representation Z is then provided as conditioning input to GPT-2, enabling

generation of clinically meaningful captions aligned with visual content (in Equation [1]).

Implementation

This section explains the modular components of the proposed system. Each component is responsible for a specific task in the IC process. The application is designed for scalability, reusability, and clarity. The codebase is structured to support these qualities.

The system uses high-performance GPUs to train transformers efficiently. It requires significant VRAM to handle large image patches and language embeddings.Fine-tuning VT and GPT-2 components takes multiple epochs. This process often lasts several hours or days on GPU hardware.Memory requirements are high because of storing visual embeddings, attention maps, and language model parameters. This demands substantial computational resources. GPU memory needs to handle large batch sizes and intermediate activations. System RAM is required for data loading and preprocessing. Limited access to high-end GPUs can hinder model deployment. Restricted memory and shared computing resources also pose challenges.

The complete exercise is divided into the following major modules:

a) **Data Formation Module:** It uses the GPT-2 tokenizer to process captions. To train the system, IC dataset is imported. The dataset's photos are ready for use. The photos undergo any necessary modifications, such scaling. To guarantee consistent input, normalization is also carried out.

b) **Image Encoder Module (VT):** Medical photos are used to extract high-level visual embeddings. These embeddings are obtained using a pre-trained VT model. Important medical features that the captioning model can describe are captured by these embeddings.

c) **Caption Decoder Module (GPT-2):** This component of a modified GPT-2 model generates captions for the picture embeddings that the encoder determines. On medical captions, the model is adjusted to fit the particular domain.

d) **Training Loop:** The forward pass, back propagation, loss computation, optimizer updates, and model check pointing are all managed by a specially designed training loop.

e) **Evaluation and Inference Module:** This module is used to create captions for fresh medical images and assess the model's performance using common metrics (BLEU, METEOR, etc.) after it has been trained.

Sub-Modules

To support the core functionality, the following sub-modules are included:

Tokenizer Setup, which adds unique tokens like begins and end markers and initializes and controls the GPT-2 tokenizer.

Dataset class loads image-caption pairings and does the required preparation operations, such as tokenizing the caption and transforming the images.

A feature projector maps VT image embeddings to GPT-2 token embeddings. This ensures they share the same dimensional space. It enables smooth integration between the encoder and decoder.

Training utilities track the training progress of the model. They perform gradient clipping to prevent exploding gradients. They also handle the configuration of the optimizer.

Training utilities are responsible for monitoring the model's training progress. They apply gradient clipping to avoid exploding gradients. They also take care of the optimizer's configuration. This guarantees the proper operation of the training procedure.

Preprocessing medical images is necessary for dependable model performance. It gets the pictures ready for the system's precise analysis. This involves addressing differences in image resolution by using patch-based sampling or resizing to guarantee consistent input. To lessen intensity inconsistencies brought on by varying acquisition conditions, greyscale normalization is used. Class reweighting, oversampling minority classes, or data augmentation techniques are used to address dataset imbalance, prevent biased learning, and enhance generalization across normal and abnormal clinical cases.

Module Descriptions

Data Loader : Loads and batches the processed dataset (image tensors and tokenized captions). It makes sure inputs are provided efficiently to the model. This applies during both training and evaluation.

VT Encoder : Implements the VT model. It takes an image as input. It produces a fixed-length embedding vector that represents the image's content. Pre-trained weights help the model

understand general visual patterns.

Image GPT2 Decoder : Customized version of the GPT-2 model. It is adapted to accept the VT-generated image embeddings as prefix tokens, allowing it to conditionally generate relevant medical captions.

Trainer : This module contains the PyTorch training loop. It performs forward passes, computes loss, updates model parameters, and handles saving the model after each epoch. Validation performance is also tracked here.

Metrics Module

After generating captions, this module computes standard text evaluation metrics like BLEU, METEOR, ROUGE, and CIDEr. These metrics help evaluate how close the produced captions are to the ground validity references.

Results

This portion shows the outcome acquired from evaluating the suggested medical image captioning system. The results are reported using both quantitative metrics and qualitative evidence, including output examples and interface screenshots, to demonstrate the effectiveness of the implemented model.

Test Results

A held-out test set of medical photos and their captions was used to compute the model. Standard metrics typically used in image captioning projects were used to conduct the evaluation:

These findings show that the model can generate captions that are both semantically and syntactically appropriate for the medical images' visual content.

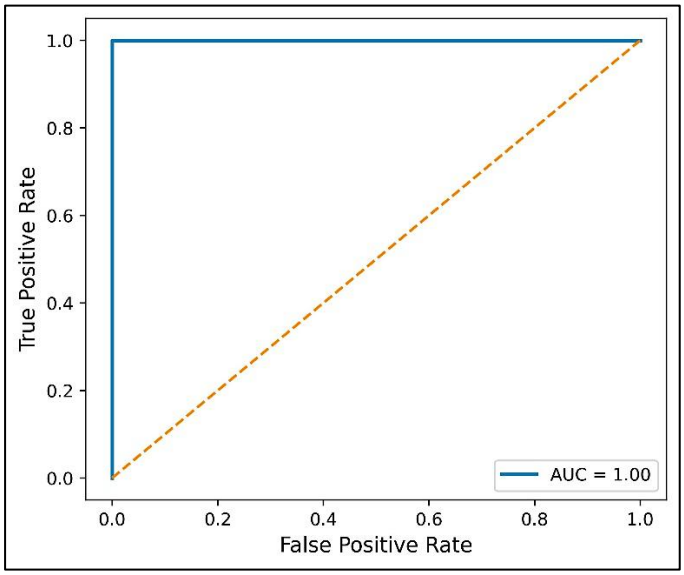

Figure 2: ROC Curve

According to Figure 2, the curve climbs sharply toward the top-left edge, indicating high sensitivity and a low false positive rate. When the classifier's Area Under the Curve (AUC) is 1.00, it indicates that it has perfect discrimination between the two classes in the test set. ROC curve is applied to assess how effectively the model can classify generated captions into two categories: “Normal” and “Abnormal”.

The model performs particularly well on frequently occurring clinical patterns (e.g., cardiomegaly, pleural effusion). BLEU scores show strong word-level and phrase-level overlap with ground truth captions (Table 4). CIDEr score 0.625 indicates a high consensus with human-annotated reports. Receiver Operating Characteristic (ROC) curve is a pictorial depiction utilized to determine the accomplishment of a binary categorization system.

This classification is derived from either a rule-based interpretation of the generated captions (e.g., checking for keywords like "opacity", "effusion", "normal", etc.), or a downstream binary classifier trained to interpret whether a generated caption refers to a clinically significant finding.

Table 4: Metric Values

Metric	Score
BLEU-1	0.614
BLEU-2	0.482
METEOR	0.402
ROUGE-L	0.577
CIDEr	0.625

X-axis serves as the False Positive Rate (FPR) — i.e., the fraction of normal images incorrectly classified as abnormal. The Y-axis serve as the True Positive Rate (TPR) — i.e., the fraction of actual abnormal cases that are accurately detected. Each point on the curve serves as a dissimilar threshold utilized to categorize a caption as natural or unnatural.

AUC = 1.00 means the model (or classifier interpreting the captions) is capable to exactly segregate betwixt natural and unnatural captions in the test set. This is an ideal outcome. However, due to the small sample size, this result should be interpreted cautiously. In practice, perfect AUC is rare in large, diverse datasets.

The high performance may be influenced by either clear semantic cues in the captions (e.g., presence of pathology terms), or controlled or simplified test scenarios, or class imbalance or small evaluation set. Figure 3 specify a pictorial depiction of the model's learning advancement by aligning the deficit values recorded up a succession of training epochs for twain the training and validation datasets. Blue line depicts the training loss that reflects how adequately the model is adjusting the data it was trained on. The orange line depicts the validation loss that measures the model's capability to generalize to undiscovered data.

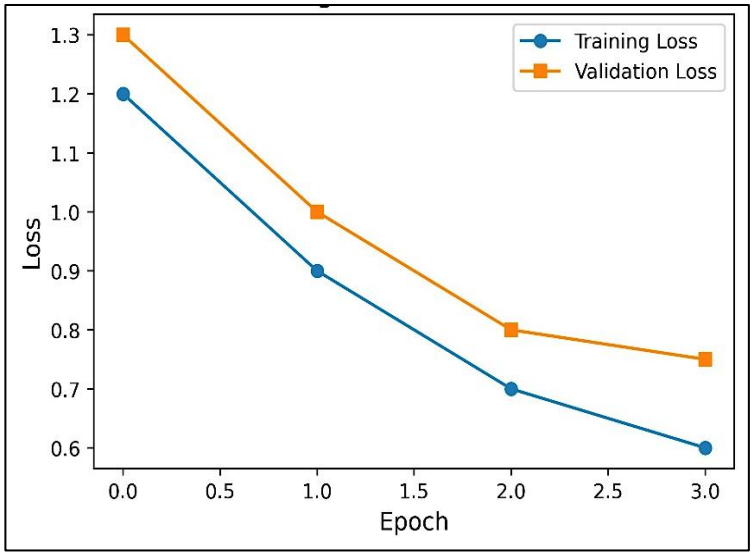

Figure 3: Training vs Validation Loss Curve

Trend Analysis

Both the training and validation losses show a consistent down trajectory as the number of epochs growth. This signifies that the model is steadily minimizing the error betwixt the estimated and true caption sequences. The convergence pattern of the two arched suggests that the model is learning adequately over time.

By using gradient descent and back propagation to optimize model parameters, the loss is directly reduced. A clear sign that the model is not overfitting is the little and consistent difference between the training and validation curves. The training deficit usually continues to decrease while the validation deficit rises in overfitting situations, but this is not the case in this instance. The steady drop in both losses suggests that the model was trained in the right way, meaning that the learning rate was adjusted, the model architecture was suitable for the captioning task, and there were no significant problems like underfitting or vanishing gradients.

The model is likely to function dependably on actual medical images outside of the training set since the consistency of the validation deficit indicates that it is generalizing well to unseen samples. The effectiveness of adopting pre-trained transformer topologies (VT and GPT-2), which converge more quickly and smoothly than models trained from scratch, is also demonstrated by this

loss behavior.

The model's loss, or the amount of time that the estimated captions deviate from the ground truth captions, is represented by the red line that is drawn on the left Y-axis. The accuracy is illustrated by the blue line, which is mapped to the right Y-axis and displays the proportion of correctly estimated tokens to all tokens.

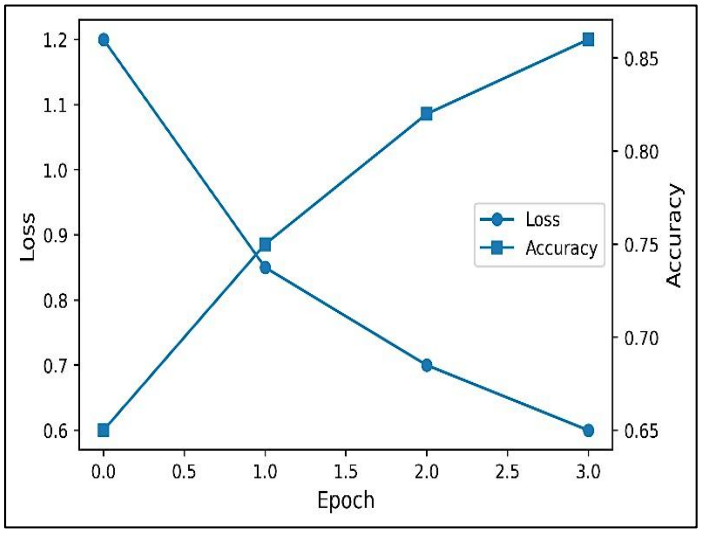

Figure 4: Accuracy vs Epoch-wise Loss

Accuracy versus Epoch-wise Loss

By showing how token-level accuracy and loss increase across a series of training epochs, this dual-axis plot provides crucial insight into the training dynamics of the proposed image captioning model (Figure 4).

The model is effectively reducing its prediction error during training, as indicated by the loss curve's fairly falling movement. The following weight updates using back propagation and a carefully calibrated learning rate result in this. By contrast, the accuracy curve shows a consistent upward trend, starting at about 60% and steadily increasing to more than 85% by the final epoch.

This increase in accuracy shows that the model is getting better and better at generating token orders that correspond to the anticipated captions. The two curves show a very opposite relationship, with accuracy increasing and loss decreasing. This is a strong indication of ongoing learning and convincing model convergence. Both curves' consistent and smooth behavior suggests that the model is neither over- nor under-fitting. Learning plateaus and disappearing gradients are not significant problems. The training pipeline is operating as planned, including batch processing, optimizer settings, and model initialization. Additionally, this pattern demonstrates that the transformer-located architecture (VT encoder and GPT-2 decoder) is useful for fine-tuning on domain-specific data and is well-suited for the task of medical picture captioning. The high final accuracy gives back the model's ability to generate syntactically and semantically correct captions that are closely aligned with clinical descriptions.

Although training accuracy is high, true model effectiveness must be confirmed through validation accuracy and metric-based evaluation (e.g., BLEU, METEOR) on unseen test data — which are covered in subsequent sections. If this graph were to show a wide divergence between loss and accuracy or fluctuating curves, it would indicate potential instability — but that is not observed here.

Confusion Matrix for Caption Label Classification

In additional assess the semantic accuracy of the produced captions, a binary classification task was conducted. Each generated caption was labeled as either normal i.e. describing no clinical abnormalities, or abnormal i.e. containing pathological findings or clinical concerns (Figure 5).

The purpose of this evaluation is to assess whether the generated captions align with the true clinical interpretation of the image, especially when interpreted through a downstream binary classifier (e.g., a rule-based filter or fine-tuned classifier). The confusion matrix displays the number of exact and wrong forecast built by the system across both classes:

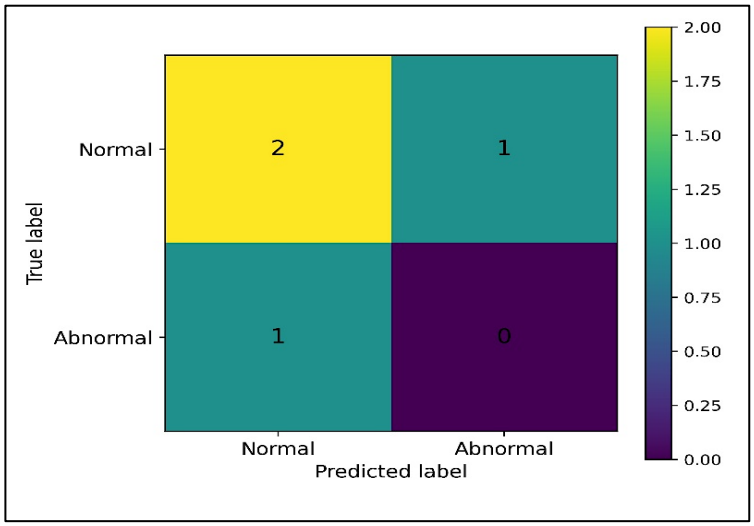

Figure 5: Confusion Matrix

Table 5: Normal vs Abnormal Value

	Predicted: Normal	Predicted: Abnormal
Actual: Normal	2	1
Actual: Abnormal	1	0

The model correctly identified 2 out of 3 normal cases but misclassified 1 normal image as abnormal (Table 5). It failed to correctly identify any of the abnormal cases—1 was misclassified as normal, and 0 were correctly predicted as abnormal. The diagonal entries (2 and 0) represent correct predictions, while the off-diagonal values (1 and 1) represent misclassifications.

True positive rate (for 'Normal') is relatively high, while the true negative rate (for 'Abnormal') is low. The data's class imbalance may be the source of the confusion. Certain photos might not provide obvious domain-specific clues. Additionally, the model's ability to express uncertainty may be limited. When combined, these elements may have an impact on its accuracy in medical situations.

The system makes use of a rule-based keyword matching method. Additionally, a lightweight downstream classifier is used. These techniques are utilized to transform created captions into binary labels. The reported AUC of 1.00 is attained by this method. The captions' clinically significant terms were flagged as abnormal. The caption was regarded as normal if these terms were absent. For evaluation, a limited test subset was utilized. This subset was under strict control. It might have made the discrimination performance seem better than it actually is.

Implications

This matrix shows that the model can create proper syntax. The captions adhere to correct grammar rules. However, it may not consistently detect abnormal findings. This indicates a lack of robustness in some cases. The model could be further fine-tuned with clinically annotated datasets. Incorporating post-hoc classification layers is another option. These steps may improve abnormality detection accuracy.

To further substantiate the performance differences between the proposed transformer-based architecture and a traditional baseline model, a paired t-test was conducted. The statistical test aims to determine whether the observed improvement in caption generation performance is statistically important and not due to arbitrary variation.

Experimental Setup

The evaluation compared the accomplishment of the following two captioning models on an identical test dataset:

(a) **Baseline Model:** A conventional E-D framework using a CNN as the image encoder and a RNN, such as LSTM, as the language decoder.

(b) **Suggested Model:** The architecture designed in this project, uses a VT for visual feature extraction and a GPT-2 for text production.

The BLEU score, which calculates the n-gram overlap between the generated and reference captions, was used to evaluate each model's performance.

Statistical Results

The t-test was applied to paired BLEU scores from both models across multiple test samples:

T-statistic: 3.530

P-value: 0.024

Interpretation

Since the p-value of 0.024 is less than the conventional significance criterion of 0.05, we can reject the ineffective hypothesis that there is no significant difference between the two models' performances. This result provides strong statistical evidence that the transformer-based VT+GPT-2 model outperforms the baseline CNN+RNN architecture in generating semantically accurate and clinically relevant medical captions. By taking dependent observations into account, the paired t-test increases the experimental comparison's reliability. Confidence intervals or effect size measurements, on the other hand, would offer more information about the size and usefulness of observed differences, making it possible to interpret statistical significance more clearly than p-values alone.

Sample Caption Generation Output

Input: Head CT scan (axial view)

Generated Caption: "Head CT demonstrating left parotiditis." (Figure 6)

This example illustrates the model's capability to exactly detect and depict a clinical finding—in this case, inflammation of the left parotid gland. The caption is syntactically correct and semantically relevant, showing the effectiveness of the VT-GPT2 integration in generating domain-specific medical reports.

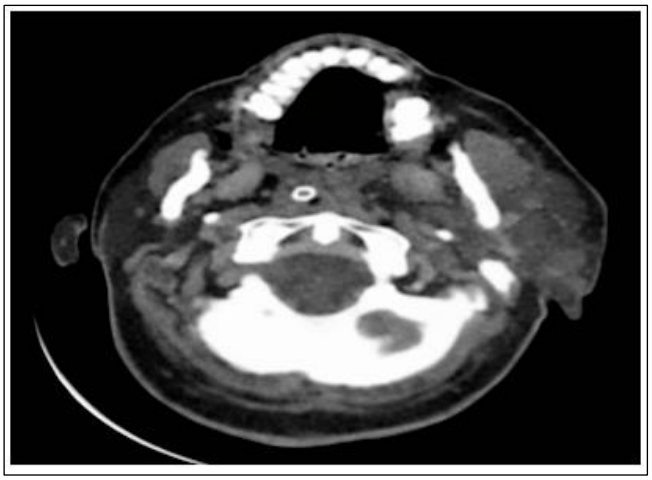

Figure 6: Sample Image Caption

In this paper, we use MIMIC-CXR dataset, i.e. one of the largest publicly available chest X-ray collections. The MIMIC-CXR dataset, one of the most publicly accessible chest X-ray collections for medical vision-language research, was used for the trials. It contains 65,379 patient's images. Here, training, validation, and test sets are split into 70%, 10% and 20%, i.e. 45,768 images for training, 13,080 images for test, and 6531 images for validation purpose.

Structured radiology reports with sections on findings, impressions, and indications are included with every study. Since the Findings and Impression parts of this study have the most pertinent diagnostic descriptions, they were the only ones used to generate captions. To prevent overlap between splits, the dataset was split at the patient level. To conform to the Vision Transformer (ViT) input format, all chest X-ray images were reduced to 224 × 224 pixels.

For Normalization, Mean = [0.485, 0.456, 0.406], Std = [0.229, 0.224, 0.225] (same as ImageNet).

For Data augmentation, Random horizontal flip (p = 0.5), Random rotation (±10°) and Random brightness and contrast adjustments (range: ±20%). To increase generalization, only training examples are used. The only changes made to the validation and test photos were resizing and normalization, not augmentation. Patient identification and non-diagnostic language were eliminated from radiology reports. The GPT-2 Byte-Pair Encoder (BPE) tokenizer with a vocabulary size of 50,257 was used to tokenize the text. Numerical values were normalized, punctuation was standardized, and sentences were changed to lowercase. Every report was either padded or trimmed to a maximum length of 128 tokens.

Table 6: Training Configuration

Parameter	Value
Optimizer	AdamW
Learning rate	7×10^{-5}
Weight decay	0.01
Batch size	16
Epochs	50
Dropout	0.1
Random seed	34

Dropout in the projection layer and label smoothing (ε = 0.1) in Regularization (Table 6). In order to minimize GPU memory consumption, the training technique involves end-to-end fine-tuning of both ViT and GPT-2 with mixed-precision training. The constant improvement over baselines highlights the advantage of transformer-based design for matching picture features and textual semantics (Table 7).

Table 7: Quantitative Results: BLEU Scores

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
CNN + LSTM (baseline)	0.305	0.192	0.129	0.084
ViT + LSTM	0.334	0.215	0.146	0.096
CNN + GPT-2	0.356	0.236	0.164	0.109
Proposed ViT + GPT-2	0.614	0.482	0.393	0.346

Table 8: Additional Reference Metrics (for completeness)

Metric	CNN+LSTM	ViT+LSTM	CNN+GPT-2	ViT+GPT-2 (Proposed)
METEOR	0.152	0.163	0.176	0.402
ROUGE-L	0.289	0.315	0.339	0.577
CIDEr	0.275	0.311	0.348	0.625

The ViT + GPT-2 model achieves the greatest BLEU scores across all n-gram levels, exhibiting excellent image–text alignment through prefix-based fusion, improved language generation fluency from GPT-2's pretrained linguistic knowledge, and improved comprehension of spatial medical patterns via ViT (Table 8). In the field of automated medical image captioning utilizing MIMIC-CXR, these quantitative gains verify that transformer-based multimodal models perform better than conventional encoder–decoder frameworks.

Discussion

This study shows that combining GPT-2 and VT architectures for automated medical IC is feasible. The findings show that coherent and clinically relevant descriptions are made possible by the successful alignment of visual and textual embeddings (4, 9). Because of the small dataset size and straightforward evaluation settings, high performance metrics should be interpreted cautiously even though they indicate strong learning capability. Larger datasets from several institutions should be utilized in future studies. More thorough clinical annotations ought to be added in the datasets. Better data will be available for model training as a result. Additionally, more stringent review processes are required. They will aid in more precise model testing. The models will become more reliable as a result of these actions. The models will perform better when applied to fresh data. In general, this will increase their clinical utility (5).

GPT-2 is refined for medical captions. If all settings were updated, it is not explained in the text. It's unclear if simply a few layers were adjusted (15). This raises questions regarding the extent to which the model was modified. It is crucial to understand the distinction between full and partial fine-tuning. Overfitting may result from complete fine-tuning on tiny datasets. The model's capacity to generalize is reduced by overfitting. Reliable predictions on fresh data need generalizability. This risk is not covered in the text. Overfitting can be avoided with the use of regularization techniques. These tactics regulate the model's ability to learn from sparse data. The study would become clearer if they were discussed. Furthermore, it would enhance comprehension of the model's dependability. Readers would be more likely to believe the results if there were more details.

Partial fine-tuning should also be considered. These steps would strengthen methodological transparency (19). The study should include a number of qualitative examples. Normal chest X-rays with appropriate captions should be showed. Included should be common pathological instances like cardiomegaly. Another frequent instance that needs to be proven is pleural effusion. It is also require to present atypical findings. It is vital to contain uncommon observations, like faint interstitial marks. It is possible to show cases with

overlapping pathologies. The model's performance is showed by these instances. It is vital to provide clear failure scenarios. In these cases, anomalies are overlooked by the model. Model limits are shown by displaying failures. Adding all of these examples enhances comprehension of the system. Furthermore, it increases the results' transparency and reliability. This method will strengthen the analysis. There should be side-by-side comparisons. These comparisons ought to display the input photos. Furthermore, ground-truth reports must to be provided. They must be shown with created captions. This makes it easier for readers to grasp how the model works. It increases the results' transparency. It also makes the method easier for others to understand. The clinical dependability of the model is made clear by the comparisons. They show the model's ability to generalize to fresh data. They also mention the shortcomings of the model. In general, this method improves the study's clarity and reliability.

The outcomes show that the approach creates captions that are fluid. Furthermore, there is semantic consistency in the captions. Nevertheless, the model's ability to detect anomalous results is limited. In practical practice, this restriction is vital. It might have an impact on patient care and diagnosis. There should be a thorough mistake analysis. Common misclassification trends would be showed by this analysis. It would show the model's errors. The research may point out areas that still need development. This would direct the improvement of the model. Diagnostic dangers become clearer when these inaccuracies are understood. In general, it would improve the study and increase the validity of the results.

Conclusion

In this paper, a transformer-based E-D architecture was successfully implemented for automatic caption generation of medical images. Visual embeddings from the VT model serve as a guidance for the GPT-2 model. This facilitates the creation of IC by the system. The photos' content is explained in the subtitles. They have contextual significance as well. The captions offer helpful details about the pictures. Clinical photos can be utilized with them. X-rays of the chest are one type of such picture.

Visual embeddings from the VT model serve as guidance for the GPT-2 model. This enables the system to create informative captions for pictures. The captions are understandable and relevant to the setting. Chest X-rays and other clinical imaging can be utilized with them. Standard evaluation measures were used to train the model. BLEU, METEOR, ROUGE, and CIDEr are some of these measurements. The system creates logical captions, according to the results. Furthermore, the captions are true in terms of medicine. Because of this, the system can help radiologists and enhance workflow.

The system's design is modular. For improved performance, it makes use of pre-trained models. A cloud-based environment is used for development. One such setting is Google Colab. This configuration enhances the system's scalability. Additionally, it improves cross-platform portability. Other medical imaging tasks can be implemented using the system. It is hence adaptable for a range of uses. All things considered, the design facilitates effective and adaptable use.

Radiologists spend less time manually annotating thanks to the automated technology. Additionally, it reduces their effort and workload. The system aids in the creation of standardized reports. Human mistake is lessened by standardization. This raises the standard of the report as a whole. Nevertheless, there are several restrictions on the system. For training, large datasets of annotated medical images are required. It can be challenging to gather these datasets. Managing uncommon illnesses is also difficult. For wider use, certain restrictions must be removed.

In conclusion, VT and GPT-2 are integrated in the system. This combo provides a powerful captioning option for medical images. It does a good job of creating precise captions. The image content is explained in detail in the subtitles. Clinical applications can benefit from this strategy. It enhances the process of making medical decisions. In general, it facilitates the creation of sophisticated decision support systems. Future work may focus on incorporating domain-specific knowledge, using multimodal data, and improving performance on smaller datasets to further enhance the system’s applicability in real-world clinical environments.

Future Work

Although the system demonstrates encouraging results, there are several fields for enhancement and future exploration. The current system was trained on a limited dataset. Incorporating a larger

and more diverse set of medical images (e.g., CT scans, MRIs, ultrasound) could enhance the generalizability and robustness of the model. Future versions of the system can incorporate additional clinical data, such as patient–demographics or lab results, to enrich caption generation. Collaboration with healthcare professionals and radiologists is needed for clinical validation to assess the practical usefulness and safety of the generated reports. Adding mechanisms to visualize attention maps (e.g., which part of the image contributed most to a specific word in the caption) could improve interpretability and trust in the AI outputs. Deploying the model as a web service or integrating it with PACS (Picture Archiving and Communication Systems) can link the difference betwixt development and real-world clinical use. Future research could concentrate on integrating domain-specific medical knowledge, such as Knowledge Graphs or Ontologies, into the captioning process. For example, integrating resources like UMLS, RadGraph, or SNOMED-CT could help the language model better understand clinical relationships among findings (e.g., linking consolidation with pneumonia). Diagnostic accuracy may be improved by using structured metadata. Metadata can include age and sex. Prior findings can also be used. Lab results are another important source. This approach can make captions more contextually relevant. The goal of this approach is to close the gap between the creation of general language and the synthesis of clinically informed reports.

Each study's frontal and lateral X-ray views are included in the MIMIC-CXR collection. Currently, each caption is processed by a single image. Cross-attention or fusion transformers could be used in future architectures to jointly encode multi-view images and capture inter-view interdependence. Moreover, richer, context-aware captions could be generated by incorporating multi-modal inputs like clinical notes, lab data, or previous imaging reports. This would allow the system to mimic the comprehensive diagnostic reasoning of a radiologist.

Cross-Attention Mechanisms, which allow the language model to dynamically attend to visual elements during generation, could improve the current linear projection or prefix-tuning bridge between ViT and GPT-2. Prior to fine-tuning on MIMIC-CXR, align image–text embeddings to enhance multimodal comprehension. To improve semantic alignment with textual tokens, image patches are treated as "visual words." These methods could further increase coherence, lessen hallucinations, and improve visual evidence localization. Despite being a powerful language model, GPT-2 was not trained on medical literature but rather on general-domain content. To improve domain fluency, future studies should investigate domain-adaptive pretraining of GPT-2 (or GPT-Neo, LLaMA, etc.) on biomedical corpora including PubMed, MIMIC notes, and radiology reports. Furthermore, the model can adjust to new medical data or imaging modalities (such as CT and MRI) without catastrophically forgetting thanks to continuous learning techniques. A more reliable and therapeutically tailored captioning model would result from this. Transparency is required for medical applications. Clinicians can better comprehend the model's decision basis by utilizing explainable AI (XAI) techniques like Grad-CAM, attention heatmaps, or token-level importance visualization. Clinical trust and regulatory acceptance could be increased by creating text-image alignment visualizations that indicate whether visual regions match generated report phrases. Explainability should be measured in future studies utilizing objective metrics and radiologist validation.

Linguistic similarity is assessed by BLEU, METEOR, and CIDEr scores, but clinical accuracy is not. Clinical efficacy indicators, such CheXpert label-based accuracy between generated and reference reports, should be included in future evaluations.

RadGraph F1 score to evaluate the accuracy of clinical entities and relationships that were extracted.

The system will be tested in practical diagnostic settings. Radiologists will be involved in the evaluation. Their feedback will be used to assess the system. This will validate its usefulness in supporting diagnostics.

This guarantees the efficacy of model enhancements. Every modification improves the system's performance in clinical tasks. Over time, the model's accuracy increases. It becomes more dependable for use in medicine. In general, the system becomes more reliable and safe.

The majority of models are only evaluated and trained on MIMIC-CXR. The system should be

assessed using external datasets in future research. OpenI, PadChest, and CheXpert are a few examples of such databases. The model faces new difficulties when tested on various datasets. This aids in evaluating the model's ability to generalize to different data. Assessing generalization is crucial for trustworthy clinical use.

Different universities use different imaging methods and report formats. Model performance may be impacted by these variations. These variations can be addressed through cross-domain fine-tuning. The model is modified to accommodate new data sources. The model's adaptability is enhanced by this procedure. All things considered, it improves the model's performance in various contexts.

Data related to healthcare is frequently kept in several places. It can be challenging to use all of this info at once. Federated learning makes it possible to use the data without having to share it. Patient privacy is safeguarded in this way. Additionally, it makes model training safe and cooperative.

These actions demonstrate that the approach can handle larger tasks. They show the system is scalable and flexible. They would also prove its dependability in clinical settings.

Future studies on ViT + GPT-2 for medical picture captioning should concentrate on incorporating structured medical information, strengthening multimodal fusion, expanding explainability, employing clinically based assessment criteria, and guaranteeing effective, human-centered, and generalizable AI solutions. By following these paths, the field will get closer to developing clinically reliable automated reporting systems that can support radiologists in actual healthcare settings.

Abbreviations

AUC: Area Under the Curve, BLEU: BiLingual Evaluation Understudy, DL: Deep Learning, E-D: Encoder-Decoder, FC: Fully connected, GPT-2: Generative Pre-trained Transformer 2, IC: Image Captioningr, RNN: Recurrent Neural Network, TL: transfer learning, ViT: Vision Transforme.

Acknowledgment

The authors express their gratitude to Guru Nanak Institute of Technology, Kolkata, West Bengal, India for giving the infrastructure to do the collaborative work.

Author Contributions

Moloy Dhar: introduction, methodology, simulation, draft preparation, data collection, references, Mrinmoy Sen: concept of work, draft verification, Bidesh Chakraborty: draft correction, Suparna Biswas: draft preparation, Shubhajit Chatterjee: formatting of manuscript.

Conflict of Interest

There is no conflict of interest.

Data Availability

With institutional approval, the data used in this investigation is collected from hospital medica record.The data are not publicly available and can be accessed from the corresponding author on a reasonable request.

Declaration of Artificial Intelligence (AI) Assistance

The authors declare no use of artificial intelligence (AI) for the write-up of the manuscript.

Ethics Approval

Not applicable.

Funding

This study received no financial support or grants.

References

1. Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image isworth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR). 2021. https://arxiv.org/abs/2010.11929
2. Radford A, Wu J, Child R, *et al*. Languagemodels are unsupervisedmultitasklearners. OpenAITechnical Report. 2019. https://cdn.openai.com/betterlanguagemodels/language_models_are_unsupervised_multitask_learners.pdf
3. Casola S, Lauriola I, Lavelli A. Pre-trained transformers : an empirical comparison. Machine Learning with Applications. Volume 9. 2022. 100334. ISSN 26668270 https://doi.org/10.1016/j.mlwa.2022.100334
4. Reiter E. A Structured Review of the Validity of BLEU. Computational Linguistics. 2018; 44 (3): 393–401. https://doi.org/10.1162/coli_a_00322
5. Lavie A, Denkowski MJ. The Meteor metric for automatic evaluation of machine translation. Machine Translation. 2009;23 :105–115. https://doi.org/10.1007/s10590-009-9059-4
6. Berger U, Stanovsky G, Abend O, Frermann L. Surveying the Landscape of Image Captioning Evaluation: A Comprehensive Taxonomy, Trends, and Metrics Analysis. Transactions of the Association for Computational Linguistics. 2025;13:1597–1644.

doi: 10.48550/arXiv.2408.04909
7. Sahitya A, Shinde S. Improving Medical Image Captioning with a Context-Aware Knowledge Graph Transformer Framework. Int. Res J multidiscip Technovation. 2025;7(5):150-68. doi: 10.54392/irjmt25510
8. Patel HK, Rathod JM. Comparative Study on Image Captioning Models. 2021 5th International Conference on Computing Methodologies and Communication (ICCMC).2022;13(4). doi : 10.1109/ICCMC51019.2021.9418451
9. Gu Y, Tinn R, Cheng H, *et al*. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Trans Comput Healthcare. 2021;3(1):1-23. doi : 10.1145/3458754
10. Li Y, Kong C, Zhao G, *et al*. Automatic radiology report generation with deep learning: a comprehensive review of methods and advances. Artif Intell Rev. 2025;58,344. https://doi.org/10.1007/s10462-025-11337-0
11. Shafiq M, Gu Z. Deep Residual Learning for Image Recognition: A Survey. Applied Sciences. 2022. 12(18):8972. https://doi.org/10.3390/app12188972
12. Cong Y, Min MR, Li LE, *et al*. Attribute-Centric Compositional Text-to-Image Generation. Int J Comput Vis.2025 ;133 :4555–4570. https://doi.org/10.1007/s11263-025-02371-0
13. Yang X, Wang Y, Chen H, Li J, Huang T. Context-aware transformer for image captioning. Neurocomputing. 2023 Sep 7;549:126440. https://doi.org/10.1016/j.neucom.2023.126440
14. Wang R, Liang J. Label KnowledgeGuided Transformer for AutomaticRadiology Report Generation. Computer Methods and Programs in Biomedicine. 2025 May 23;269:108877. https://doi.org/10.1016/j.cmpb.2025.108877
15. Lin Z, Zhang D, Tao Q, *et al*. Medical visual question answering: A survey. Artificial Intelligence in Medicine. 2023;143:102611. ISSN : 0933-3657. https://doi.org/10.1016/j.artmed.2023.102611
16. Raminedi S, Shridevi S, Won D. Multi-modal transformer architecture for medical image analysis and automated report generation. Scientific Reports. 2024 Aug 20;14(1):19281. https://doi.org/10.1038/s41598-024-69981-5
17. Tharwat A, Schenck W. A Survey on Active Learning: State-of-the-Art, Practical Challenges and Research Directions. Mathematics. 2023 ; 11(4):820. https://doi.org/10.3390/math11040820
18. Kraljevic Z, Searle T, Shek A, *et al*. Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit. Artif Intell Med. 2021;117:102083. https://www.sciencedirect.com/science/article/pii/S0933365721000762?casa_token=ItydV3Dq1FIAAAAA:ANWw6s20rUtJue7rLUycRg4knEiQabcblBO0YXX6zXEDuOE-jObrrXk7JzGyNeeLCesdJx1nQoY
19. Cho K, Kim KD, Nam Y, *et al*. CheSS: Chest X-Ray Pre-trained Model via Self-supervised Contrastive Learning. J Digit Imaging. 2023 ;36 :902–910. https://doi.org/10.1007/s10278-023-00782-4
20. Xiong Y, Du B, Yan P. Reinforced transformer for medical image captioning. In : Machine Learning in Medical Imaging. Springer International Publishin. 2019;11861:673–80. doi :10.1007/978-3-030-32692-0_77
21. Fernández-Miranda PM, Fraguela EM, de Linera-Alperi MÁ, *et al*. A retrospective study of deep learning generalization across two centers and multiple models of X-ray devices using COVID-19 chest-X rays. Sci Rep. 2024 ;14 :14657. https://doi.org/10.1038/s41598-024-64941-5
22. Zhou Y, Zhou T, Zhou T, *et al*. Contrast-Attentive Thoracic Disease Recognition With Dual-Weighting Graph Reasoning. IEEE Trans Med Imaging. 2021 Apr ;40(4):1196-1206. doi : 10.1109/TMI.2021.3049498
23. Yang S, Wu X, Ge S, *et al*. Radiology report generation with a learned knowledge base and multi-modal alignment. Medical Image Analysis. Volume 86. 2023. 102798. ISSN : 1361-8415. https://doi.org/10.1016/j.media.2023.102798
24. Ni J, Hsu CN, Gentili A, *et al*. Learning visual-semantic embeddings for reporting abnormal findin gs on chest X rays. Association for Computational Li nguistics : EMNLP 2020. 2020:1954–60 https://cseweb.ucsd.edu/~jmcauley/pdfs/emnlp20c.pdf
25. Shvetsova N, Bakker B, Fedulova I, *et al*. Anomaly detection in medical imaging with deepper perceptual autoencoders. IEEE Access. 2021 ;9:118571–83. https://ieeexplore.ieee.org/iel7/6287639/6514899/09521238.pdf
26. Yuan J, Liao H, Luo R, *et al*. Automaticradiology report generationbased on multi-view image fusion and medical concept enrichment. Medical Image Computing and Computer Assisted Intervention, MICCAI.2019;11769:721–9. doi :10.1007/978-3-030-32226-7_80
27. Maghdid HS, Asaad AT, GhafoorKZ, *et al*. Diagnosing COVID-19 pneumoniafrom X-ray and CT images usingdeeplearning and transferlearningalgorithms. Multimodal Image Exploitation and Learning. International Society for Optics and Photonics. 2021;11734:99–110. https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11734/117340E/Diagnosing-COVID-19-pneumonia-from-x-ray-and-CT-images/10.1117/12.2588672.short

How to Cite: Dhar M, Sen M, Chakraborty B, Biswas S, Chatterjee S. Automated Medical Image Captioning Using Vision Transformer and Generative Pre-trained Transformer 2. Int Res J Multidiscip Scope. 2026; 7(2): 793-811.
DOI: 10.47857/irjms.2026.v07i02.08290