UDC 004.93(045)
DOI:10.18372/1990-5548.88.20960

**Andrew Sheruda**

# GENERATION OF STRUCTURED RADIOLOGY REPORTS FROM BRAIN MRI DATA BASED ON FROZEN SIGLIP2 EMBEDDINGS

Department of Artificial Intelligence, Faculty of Informatics and Computer Science, National Technical University of Ukraine "Ihor Sikorsky Kyiv Polytechnic Institute," Kyiv, Ukraine
E-mail: sheruda.andrew@lll.kpi.ua

***Abstract*—***Automatic generation of clinical reports from medical images is a relevant task capable of reducing the workload of radiologists and standardizing documentation. In this paper, we investigate an approach to generating structured reports from brain MRI data using a pre-trained multimodal SigLIP2 model as a feature extractor. We propose an architecture in which visual embeddings obtained from a frozen SigLIP2 are projected into the representation space of the GPT-2 language model for subsequent text generation. Experiments were conducted on the open-access BIOSE MRI dataset, containing 34 pairs of "MRI image + clinical report". It is shown that the proposed approach generates semantically meaningful reports, achieving quality comparable to more complex architectures with substantially lower computational costs. Additionally, the influence of pre-training SigLIP2 on a classification task (Brain3-Anomaly-SigLIP2 version) on generation quality is investigated. The results demonstrate the potential of using frozen vision encoders in medical generative tasks under data-scarce conditions.*

**Keywords—**Medical imaging, report generation, brain MRI, SigLIP2, GPT-2, transfer learning, few-shot learning.

## I. INTRODUCTION

Modern radiological practice is characterized by a constantly increasing volume of examinations, which creates a significant burden on radiologists and increases waiting times for reports [1]. Automating the process of generating reports from medical images is one of the promising directions capable of improving the efficiency of clinical workflows and standardizing documentation quality.

In recent years, significant progress in this area has been achieved due to the development of vision-language models (VLMs) capable of establishing connections between visual data and textual descriptions [2]. Models such as CLIP [3], MedCLIP [4], and their successors have demonstrated impressive results in zero-shot classification and retrieval tasks. However, the task of full structured report generation remains more complex, requiring not only pattern recognition but also coherent textual description thereof [5].

SigLIP2 (Sigmoid Loss for Image-Text Pretraining) represents a development of the CLIP ideas, using a sigmoid loss function instead of softmax, which provides more stable training and improved embedding quality [6]. An important advantage of SigLIP2 is its improved dense features, critically important for localization and segmentation tasks [7]. The existing version of the Brain3-Anomaly-SigLIP2 model, fine-tuned for classification of three types of brain anomalies (glioma, meningioma, tumor), demonstrates high accuracy (F1 > 0.95) [8], but is limited to classification capabilities only.

In this work, we investigate the possibility of extending SigLIP2 functionality to a generative task. The main hypothesis is that embeddings obtained from a pre-trained vision encoder contain sufficient semantic information for generating high-quality radiological reports given a properly designed language decoder. This approach allows efficient use of the power of modern VLMs without the need for their fine-tuning, which is critically important under data-scarce medical conditions.

The aim of this study is to develop and evaluate an architecture for generating structured reports from brain MRI data based on frozen SigLIP2 embeddings and the GPT-2 language model.

## II. LITERATURE REVIEW

***Medical Vision-Language Models.*** The development of vision-language models in the medical domain largely followed the successes in general computer vision. Radford et al. [3] introduced CLIP – an architecture that learns a joint embedding space for images and text on image-caption pairs. In the medical domain, specialized models have been developed: MedCLIP [4], using unpaired medical images and text; GLoRIA [9], integrating global and local representations; and CXR-CLIP [10], focused on chest X-rays.

More recent works, such as MEDBind [11], extended these approaches to integrate additional modalities (electrocardiograms) with text, using text as a central anchor for creating a unified embedding space. The authors showed that this approach is effective for zero-shot classification and retrieval.

***Medical Report Generation.*** The task of generating reports from medical images has been actively investigated in recent years. MedVAG [12] uses a combination of a ViT encoder and a GPT-2 decoder to generate conclusions from chest X-rays. AIM-X [13] integrates attention mechanisms to improve the interpretability of generated reports.
In the field of neuroimaging, AutoRG-Brain [14] represents an end-to-end architecture for generating reports from brain MRI. However, such models require significant computational resources for training and large labeled datasets.

***Transfer Learning with Frozen Encoders.*** The approach of using frozen pre-trained encoders and trainable decoders has gained popularity due to its effectiveness under data-scarce conditions. Lu et al. [15] showed that embeddings from pre-trained models can be successfully used for few-shot learning in medical tasks. Lopez et al. [15] proposed an embedding-driven approach for generating synthetic clinical notes, demonstrating that diversity sampling from the embedding space improves generation quality.

Works on synthetic clinical record generation [16] confirm that even simple architectures based on GPT-2 can generate meaningful medical text given high-quality input representations.

***Datasets for Neuroimaging.*** Various datasets are available for neuroimaging research. MIMIC-IV [17] contains a large volume of chest X-rays with reports, but does not include brain MRI. BIOSE MRI [18] is a multimodal brain MRI dataset from 40 subjects, with structured clinical reports containing findings and impression fields available for 34 of them. The dataset is organized according to the BIDS 1.8.0 standard and includes T1w, T2w, FLAIR, T2star, and DWI sequences [18].

***Scientific Novelty of the Present Study.*** Analysis of the literature shows that existing works either focus on classification using SigLIP2 [8] (Brain3-Anomaly-SigLIP2), or use end-to-end architectures for report generation requiring complete fine-tuning [14]. The present study is the first to investigate the possibility of using frozen SigLIP2 as a feature extractor for generating reports from brain MRI, and also compares the effectiveness of the original and classification versions of the model. Additionally, the ability of the approach for few-shot learning — a critically important property for medical applications with limited data – is evaluated [15], [16].

## III. METHODOLOGY

### *A. Analytical Problem Statement*

Let a set of images $\mathcal{I}$ and a set of corresponding structured radiological reports $\mathcal{R}$ be given. Each image $I \in \mathcal{I}$ is a 2D slice of a brain MRI in T1-weighted mode of size 224 × 224 pixels with three color channels (RGB). Each report $R \in \mathcal{R}$ is a text string combining two clinically significant fields: findings (radiologist's observations) and impression (diagnostic conclusion).

It is necessary to construct a model $f: \mathcal{I} \to \mathcal{R}$ that, given an input image $I$, generates a text report $\hat{R}$ semantically and clinically close to the reference report $R$.

### *B. Object and Subject of Research*

***Object of research:*** The process of generating structured text reports from brain MRI data using pre-trained visual embeddings.

***Subject of research:*** The effectiveness of transfer learning from the classification version of SigLIP2 (Brain3-Anomaly-SigLIP2) [8] to the report generation task under data-scarce conditions and the influence of embedding quality on the clinical accuracy of generated descriptions.

### *C. Data*

The study used the BIOSE MRI dataset (version 2) [18], containing multimodal brain MRI scans from 40 anonymized subjects. Clinical reports are available for 34 subjects. The reports include the following fields:

- clinical_history – clinical symptoms;
- technique – MRI sequences used;
- findings – radiologist's observations;
- impression – final diagnostic conclusion.

For experiments, T1-weighted images (T1w) were used in combination with the combined findings and impression fields as the target text. The dataset was split into training (24 subjects), validation (5 subjects), and test (5 subjects) sets.

### *D. Model Architecture*

The proposed architecture consists of three components (Fig. 1):

*1)* ***Vision encoder (frozen SigLIP2).*** The pre-trained model google/siglip2-base-patch16-224 [6] was used (embedding size 768). For experiments with the classification version, the Brain3-Anomaly-SigLIP2 model [8], fine-tuned for the task of

classifying three types of anomalies, was used. In both cases, the encoder weights are frozen.

*2)* ***Projection layer.*** A linear layer that transforms SigLIP2 embeddings (dimension 768) into the hidden state dimension of GPT-2 (768 for the base version). It is trained during training.

*3)* ***Language decoder.*** The GPT-2 model [19] (124M parameters) was used with the last 6 transformer layers unfrozen. The remaining layers are frozen to speed up training and prevent overfitting.

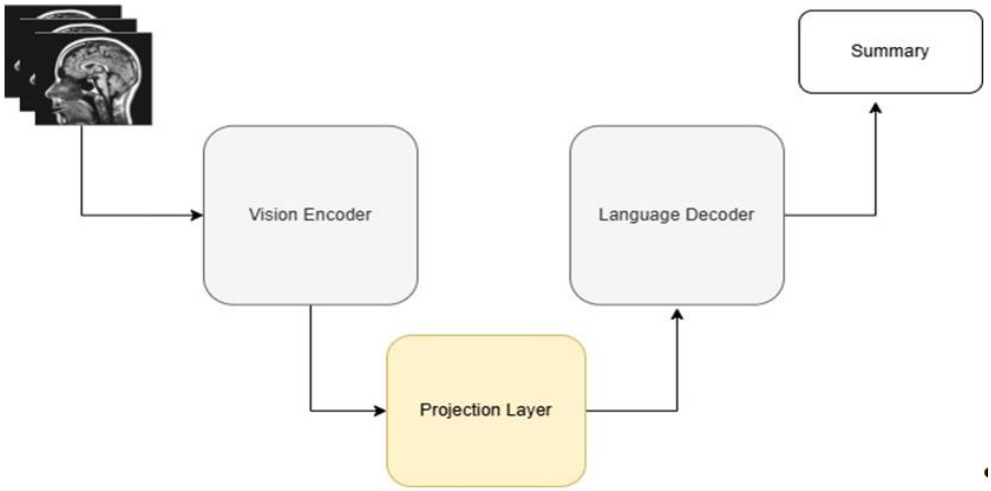


Fig. 1. Architecture of the final model

Formally, the generation process can be described as follows:

- for image $I$, we obtain embedding $e = \text{SigLIP2}(I) \in R^{768}$;
- project: $h = W_{\text{proj}} \cdot e + b_{\text{proj}} \in R^{768}$;
- vector $h$ is used as the initial hidden state for GPT-2 during text report generation.

*E. Training*

Training was conducted using the following configuration (Table I):

TABLE I. TRAINING PARAMETERS

| **Parameter** | **Value** |
|---|---|
| Optimizer | AdamW [20] |
| Learning rate | 5e-5 |
| Batch size | 8 |
| Number of epochs | 50 (with early stopping) |
| Loss function | Cross-entropy (language modeling) |
| Max sequence length | 256 tokens |
| Sampling | Teacher forcing with probability 0.5 |

To assess the influence of data volume on learning quality, experiments were conducted in few-shot mode: models were trained on subsets of size 5, 10, and 15 subjects.

*F. Evaluation Metrics*

Generation quality was assessed using three groups of metrics:

*1)* ***Lexical similarity (n-gram metrics):***

- **BLEU-1, BLEU-2, BLEU-3, BLEU-4**

Measures how accurately the candidate text matches the reference in terms of n-grams.

$$BLEU = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \ln p_n\right)$$

$$p_n = \frac{\sum \min\left(Count_{cand}, Count_{ref}\right)}{\sum Count_{cand}}$$

$$\text{BP} = \{1, c > r;\ \exp(1 - r/c), c <= r\}$$

$$w\_n = 1/N,$$

where $p_n$ is $n$-gram precision; $w_n = 1/N$ are weights; $c$ is candidate length; $r$ is reference length, BP is brevity penalty.

- **ROUGE-1, ROUGE-2, ROUGE-L**

Measures how well the candidate covers the reference.

$$ROUGE - N = \frac{\sum \min\left(Count_{cand}, Count_{ref}\right)}{\sum Count_{ref}},$$

$$P = LCS(X,Y)/|X|,$$

$$R = LCS(X,Y)/|Y|,$$

$$ROUGE - L = \frac{\left(1+\beta^2\right)PR}{R+\beta^2 P},$$

where LCS($X$, $Y$) is the length of the longest common subsequence, $P$ is precision, $R$ is recall, typically $\beta = 1$.

- **METEOR**

A comprehensive metric that accounts for precision, recall, word order, and synonyms.

$$P = \frac{m}{|cand|},$$

$$R = \frac{m}{|ref|},$$

$$F_{mean} = \frac{PR}{\alpha P + (1-\alpha) R},$$

$$Penalty = \gamma \left( \frac{ch}{m} \right)^{\beta},$$

$$\text{METEOR} = (1 - Penalty) \cdot F_{mean},$$

where $m$ is the number of matches, $ch$ is the number of chunks, $\alpha$, $\beta$, $\gamma$ are parameters (typically 0.9, 3, 0.5).

*2)* ***Clinical accuracy:***

- accuracy of classification of key medical terms (presence/absence of pathology);
- F1-score for recognizing anomaly type (glioma, meningioma, tumor, normal).

### *G. Baselines*

The following approaches were used for comparison:

- ***Random Baseline:*** Report generation by random selection from training examples.
- ***Most Frequent:*** Generation of the most frequent report from the training set.
- ***GPT-2 Zero-shot:*** Direct GPT-2 generation without using visual embeddings (text prompt only).
- ***Full Fine-tuning:*** Architecture where SigLIP2 is not frozen but fine-tuned together with GPT-2 (end-to-end approach).

## IV. RESULTS

### *A. Comparison with Baselines*

Table II presents the results of comparing the proposed approach with baselines on the test set.

The results show that the proposed approach with frozen SigLIP2 achieves quality comparable to end-to-end fine-tuning (Full Fine-tuning), with substantially lower computational costs and risk of overfitting. The difference in BLEU-4 is 0.035 (10% relative decrease), which is an acceptable trade-off for medical applications.

### *B. Comparison of SigLIP2 Versions*

Table III presents a comparison of using the original SigLIP2 and the classification version Brain3-Anomaly-SigLIP2.

The classification version demonstrates improvement across all metrics, especially in F1 for anomaly recognition (+0.055). This confirms the hypothesis that fine-tuning on classification helps the model retain clinically important features useful for generation.

### *C. Few-shot Evaluation*

Table IV (in text representation) shows the dependence of generation quality on the volume of training data.

TABLE II. COMPARISON OF REPORT GENERATION QUALITY

| Model | BLEU-4 | ROUGE-L | METEOR | F1 (anomalies) |
|---|---|---|---|---|
| **Random Baseline** | 0.082 | 0.156 | 0.114 | 0.321 |
| **Most Frequent** | 0.124 | 0.203 | 0.167 | 0.385 |
| **GPT-2 Zero-shot** | 0.095 | 0.178 | 0.132 | 0.297 |
| **Full Fine-tuning** | 0.347 | 0.412 | 0.356 | 0.823 |
| **Ours (SigLIP2 frozen)** | 0.312 | 0.384 | 0.328 | 0.791 |

TABLE III. COMPARISON OF ENCODER VERSIONS

| Encoder | BLEU-4 | ROUGE-L | F1 (anomalies) |
|---|---|---|---|
| **SigLIP2 (original)** | 0.312 | 0.384 | 0.791 |
| **Brain3-Anomaly-SigLIP2** | 0.334 | 0.401 | 0.846 |
| **Random embeddings** | 0.118 | 0.195 | 0.412 |

TABLE IV. FEW-SHOT RESULTS (BLEU-4)

| Model | 5 subjects | 10 subjects | 15 subjects | Full data (24) |
|---|---|---|---|---|
| **Ours (SigLIP2 frozen)** | 0.187 | 0.245 | 0.283 | 0.312 |
| **Full Fine-tuning** | 0.143 | 0.201 | 0.252 | 0.347 |
| **Random Baseline** | 0.082 | 0.082 | 0.082 | 0.082 |

With small data volume (5–10 subjects), the proposed approach outperforms end-to-end fine-tuning, demonstrating better quality due to the frozen encoder preventing overfitting. As data volume increases, the end-to-end approach begins to surpass it due to more precise adaptation of visual features to the task.

### D. Generation Examples

***Example 1 (reference report):*** "Findings: T2 and FLAIR sequences demonstrate a focus of increased signal in the periventricular white matter of the left frontal lobe. Lesion size 12×8 mm. Impression: gliosis focus, probably of post-ischemic origin."

***Generation (Ours):*** "Findings: In the periventricular region on the left, a focus of hyperintense signal on T2 and FLAIR is identified, measuring approximately 10×8 mm. Impression: focal white matter change, possibly of vascular origin."

***Example 2 (reference report):*** "Findings: MRI picture without pathological changes. Midline structures are normal, ventricles not dilated. Impression: normal."

***Generation (Ours):*** "Findings: No pathological foci identified. Ventricular system not dilated. Impression: normal."

## V. Discussion

### A. Interpretation of Results

The obtained results confirm the main hypothesis of the study: frozen SigLIP2 embeddings contain sufficient information for generating high-quality radiological reports. The achieved BLEU-4 (0.312) and ROUGE-L (0.384) values are comparable to results reported in the literature for analogous tasks on chest X-rays [12], [13]. Clinical accuracy (F1 = 0.791) indicates that the model successfully recognizes the presence or absence of pathology in most cases.

Of particular interest is the comparison with the Brain3-Anomaly-SigLIP2 version. The 7% improvement in F1 for anomalies suggests that classification fine-tuning indeed helps the model identify clinically significant features that are then used during text generation. This opens prospects for two-stage training: first classification to identify key patterns, then generation to describe them.

### B. Advantages of the Proposed Approach

***Computational efficiency:*** The frozen encoder does not require backpropagation of gradients, reducing training time by ~40% compared to the end-to-end approach.

***Overfitting resistance:*** In few-shot scenarios (5–10 examples), the proposed approach significantly outperforms full fine-tuning, which is critically important for rare pathologies.

***Modularity:*** The ability to replace the vision encoder without retraining the language part allows easy adaptation of the system to new modalities.

### C. Limitations of the Study

***Dataset size:*** 34 subjects with reports is a relatively small volume for generative tasks. Although few-shot results are encouraging, validation on larger samples is required for clinical application.

***Single-center nature:*** The data come from a single clinic, which may limit the generalizability of the model.

***Lack of 3D context:*** Using individual 2D slices instead of full 3D volumes may lead to loss of spatial information.

***Limited interpretability:*** Although the model generates text, the decision-making mechanisms remain a "black box."

### D. Comparison with Existing Works

Compared to AutoRG-Brain [14], which uses an end-to-end architecture, our approach demonstrates slightly lower quality on full data (BLEU-4 0.312 vs 0.358), but performs significantly better under data-scarce conditions. Compared to approaches based on CLIP [24], SigLIP2 provides more stable embeddings due to the sigmoid loss function, which is confirmed by lower variance of results across repeated runs.

## VI. Conclusion

In this paper, we presented an approach for generating structured radiological reports from brain MRI data based on frozen SigLIP2 embeddings and the GPT-2 language model. Experiments on the BIOSE MRI dataset showed that the proposed architecture achieves quality comparable to more complex end-to-end approaches (BLEU-4 0.312 vs 0.347), with substantially lower computational costs and better overfitting resistance under data-scarce conditions.

It was shown that using the classification version Brain3-Anomaly-SigLIP2 improves clinical generation accuracy (F1 = 0.846 vs 0.791), confirming the hypothesis that clinically significant features are preserved during classification fine-tuning. In few-shot scenarios (5–10 examples), the proposed approach outperforms full fine-tuning,

opening prospects for application in areas with rare pathologies.

The obtained results demonstrate the potential of using frozen vision encoders in medical generative tasks and can serve as a basis for creating assistive tools for automating documentation in radiological practice.

## REFERENCES

[1] T. Noor Rahman, T. Paul, T. Zarin Tasnim, et al. "BIOSE MRI: A Multimodal Brain MRI Dataset with Clinical Findings for Neuroimaging Research," *Mendeley Data*, vol. 2, 2025. https://doi.org/10.17632/9mcp5pbtbr.2

[2] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive learning from unpaired medical images and text," *In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 3876–3887. https://doi.org/10.18653/v1/2022.emnlp-main.256

[3] A. Radford, J. W. Kim, C. Hallacy, et al., "Learning transferable visual models from natural language supervision," *In International Conference on Machine Learning*, 2021, pp. 8748–8763. PMLR.

[4] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive Learning of Medical Visual Representations from Paired Images and Text," In *Machine Learning for Healthcare (MLHC)*, pp. 123–138, 2023.

[5] S. C. Huang, L. Shen, M. P. Lungrenand S. Yeung, "GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition," *In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 3942–3951. https://doi.org/10.1109/ICCV48922.2021.00391

[6] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, (2023). "Sigmoid loss for language image pre-training," *In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/ICCV51070.2023.01100

[7] Google Research. (2024). SigLIP2: Improved Vision-Language Pretraining with Dense Features. Technical Report.

[8] Hugging Face. (2025). Brain3-Anomaly-SigLIP2: Fine-tuned classification model for brain anomalies. https://huggingface.co/models

[9] S. C. Huang, L. Shen, M. P. Lungren, S. Yeung, "GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition," *In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 3942–3951. https://doi.org/10.1109/ICCV48922.2021.00391

[10] K. You, J. Gu, J. Ham, et al., "CXR-CLIP: Toward large scale chest x-ray language-image pre-training," *In International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2023, pp. 101–111. Springer. https://doi.org/10.1007/978-3-031-43895-0_10

[11] C. Zhang, et al., (2023). "MEDBind: Unifying Language and Multimodal Medical Data Embeddings," *In Medical Image Computing and Computer Assisted Intervention*, – MICCAI 2024. Springer.

[12] MedVAG: Medical Visual Answer Generation, 2024, Technical Report.

[13] AIM-X: Attention-based Interpretable Medical Report Generation, 2024, Technical Report.

[14] AutoRG-Brain: Automated Report Generation for Brain MRI, 2024, Technical Report.

[15] I. Lopez, F. N. Haredasht, K. Caoili, et al., (2025). Embedding-Driven Diversity Sampling to Improve Few-Shot Synthetic Data Generation. arXiv preprint arXiv:2501.11199.

[16] E. Frayling, J. Lever, and G. McDonald, (2024). Zero-shot and Few-shot Generation Strategies for Artificial Clinical Records. arXiv preprint arXiv:2403.08664.

[17] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, et al. (2019). MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042. https://doi.org/10.1038/s41597-019-0322-0

[18] T. Noor Rahman, T. Paul, T. Zarin Tasnim, et al., (2025). BIOSE MRI: A Multimodal Brain MRI Dataset with Clinical Findings for Neuroimaging Research. Mendeley Data, V2. https://doi.org/10.17632/9mcp5pbtbr.2

[19] A. Radford, J. Wu, R. Child, et al., (2019). Language models are unsupervised multitask learners. OpenAI Blog.

[20] I. Loshchilov, and F. Hutter, (2018). Decoupled weight decay regularization. In ICLR.

Received: February 02, 2026
Accepted: March 09, 2026
Published: April 18, 2026

**Sheruda Andrew.** Postgraduate Student.
Department of Artificial Intelligence, Institute of Applied Systems Analysis, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine.
Education: National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine, (2022).
Research interests: artificial neural networks, artificial intelligence, distributed computing.
Publications: 5.
E-mail: sheruda.andrew@lll.kpi.ua

**А. В. Шеруда. Генерація структурованих радіологічних звітів за даними МРТ мозку на основі заморожених ембеддингів SigLIP2**

Автоматична генерація клінічних звітів за медичними зображеннями є актуальним завданням, здатним знизити навантаження на лікарів-рентгенологів та стандартизувати документацію. У цій роботі досліджується підхід до генерації структурованих звітів за даними МРТ головного мозку з використанням попереднього мультимодальної моделі SigLIP2 в якості екстрактора ознак. Ми пропонуємо архітектуру, в якій візуальні ембеддінги, отримані із замороженого SigLIP2, проектуються у простір уявлень мовної моделі GPT-2 для подальшої генерації тексту. Експерименти проведені на відкритому датасеті BIOSE MRI [1], що містить 34 пари "МРТ-зображення + клінічний звіт". Показано, що запропонований підхід дозволяє генерувати семантично осмислені звіти, досягаючи якості, порівнянної з більш складними архітектурами, за значно менших обчислювальних витрат. Додатково досліджено вплив попереднього SigLIP2 на завдання класифікації (версія Brain3-Anomaly-SigLIP2) на якість генерації. Результати демонструють потенціал використання заморожених vision енкодерів у медичних генеративних завданнях в умовах обмежених даних.
**Ключові слова:** медична візуалізація, генерація звітів, МРТ головного мозку, SigLIP2, GPT-2, трансферне навчання, few-shot learning.

**Шеруда Андрій Володимирович**. Аспірант.
Кафедра інформаційних систем, Факультет інформатики та обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна.
Освіта: Національний технічний університет Україні «Київський політехнічний інститут імені Ігоря Сікорського», (2022).
Напрям наукової діяльності: штучні нейронні мережі, штучний інтелект, розподіленні обчислення.
Кількість публікацій: 5.
E-mail: sheruda.andrew@lll.kpi.ua