



# OPEN AI caption generation model for digital pathology of adenocarcinoma in endoscopic histopathology using multi-instance attention mechanisms

Youngseop Lee<sup>1</sup>, Kyungah Bai<sup>2</sup>, Young Jae Kim<sup>1,3</sup>, Jisup Kim<sup>4</sup> & Kwang Gi Kim<sup>1,5</sup>✉

Gastric adenocarcinoma is a leading cause of cancer related mortality worldwide, and histopathologic examination of endoscopic biopsy samples remains essential for its diagnosis and grading. In this study, we propose a novel AI based caption generation model, termed MIAC (Multi-instance Attention Captioning), designed to produce descriptive diagnostic reports from digital pathology images. The model leverages a Multi-instance learning framework with permutation-invariant self attention to aggregate features from multiple histopathology image patches into a unified representation, effectively capturing whole slide characteristics. Using the publicly available PatchGastricADC22 dataset for training and validation, and an External Test dataset from Gil Hospital of Gachon University for clinical testing, the model demonstrated strong performance across standard natural language generation metrics (BLEU@4, ROUGE-L, METEOR, CIDEr). Notably, MIAC maintained high captioning accuracy even when evaluated on previously unseen data, particularly after color normalization using the Macenko method. These results underscore the model's robustness, generalizability, and potential for integration into routine digital pathology workflows to assist pathologists in generating structured diagnostic reports.

**Keywords** Deep learning, Digital pathology, Captioning model, Multi-instance learning, Endoscopic histopathology

Gastric cancer ranks fifth in incidence and fourth in mortality worldwide, with the highest incidence rates observed in eastern Asia, accounting for the fourth highest incidence and the third highest mortality<sup>1</sup>. Early detection through screening and accurate diagnosis is crucial for improving clinical outcomes<sup>2</sup>, with common procedures including endoscopy, ultrasound, CT, and pathological examination<sup>3</sup>.

Artificial intelligence (AI) has been increasingly explored in various aspects of clinical practice for gastric cancer, including detection in endoscopic images<sup>4–6</sup>, staging<sup>7,8</sup>, treatment response assessment<sup>9</sup>, and prediction of prognosis<sup>10–12</sup>. Pathologic examination, a critical step for confirming diagnosis, has transitioned from traditional glass slides microscope examination to digital and computational pathology with the advent of whole slide imaging (WSI)<sup>13–15</sup>. WSI has demonstrated non-inferiority to traditional microscopy for primary diagnosis in surgical pathology and has become a foundational technology for integrating artificial intelligence into routine pathology workflows<sup>15–17</sup>.

In histopathology, convolutional neural networks (CNNs) have significantly advanced AI applications for gastric cancer by directly learning features from raw data without requiring pathologist input<sup>18–20</sup>. AI based models have been developed for tasks such as segmentation of gastric cancer images<sup>21,22</sup>, cancer identification<sup>23,24</sup>, and classification of adenomatous lesions in biopsy specimens<sup>25–27</sup>. Combining AI generated insights with

<sup>1</sup>Medical Devices R&D Center, Gachon University Gil Medical Center, Incheon 21565, South Korea. <sup>2</sup>Department of Pathology, Graduate School of Medicine, College of Medicine, Seoul National University, Seoul 03080, South Korea.

<sup>3</sup>Gachon Biomedical & Convergence Institute, Gachon University Gil Medical Center, Incheon 21936, South Korea.

<sup>4</sup>Department of Pathology, Gil Medical Center, Gachon University College of Medicine, Incheon 21565, South Korea. <sup>5</sup>Department of Biomedical Engineering, College of IT Convergence, Seongnam, Seongnam 13120, South Korea. ✉email: jsopath@gilhospital.com; kimkg@gachon.ac.kr

pathologist judgment has been shown to improve diagnostic accuracy and reduce false negatives, ultimately supporting pathologists in making more efficient and accurate diagnoses across diverse surgical pathology domains and cytopathology<sup>28–32</sup>.

Beyond diagnosis, AI plays a vital role in determining the histological type and grade of gastric cancer, essential for effective clinical decision making<sup>3</sup>. Several studies have highlighted the role of AI in histologic classification, grading, and structured or subtype-specific diagnostic reporting based on histopathological findings<sup>33–35</sup>. However, only a limited number of studies explicitly address free-text caption generation directly from biopsy WSIs without predefined subtype supervision. While previous studies have applied deep learning to histologic classification or lesion level segmentation, few have addressed automated caption generation from biopsy slides. For instance, SGMT (Subtype guided Mask Transformer)<sup>36</sup> introduced a subtype conditioned mask transformer for captioning, integrating subtype features as global context. However, SGMT requires subtype annotation, limiting its scalability and generalizability.

In contrast to prior pathology captioning frameworks that adopt subtype-conditioned or sequence-dependent modeling, MIAC is designed as a subtype-free, permutation-invariant captioning framework based on multi-instance learning. Rather than treating patch features as ordered sequences or relying on predefined diagnostic categories, This design enables MIAC to accommodate variability in biopsy patch sets encountered in routine histopathology workflows, distinguishing it conceptually from existing MIL-based captioning or reporting approaches.

Despite recent progress, automated caption generation from biopsy WSIs remains substantially underexplored compared with classification and segmentation in digital pathology. Existing captioning approaches often depend on subtype annotations and/or order-sensitive modeling, which limits scalability in routine biopsy preprocessing where patch counts and patch ordering are inherently variable. In addition, externally validated captioning frameworks on independent clinical datasets are scarce, creating a gap between methodological development and real-world clinical applicability. These gaps motivate the development of a subtype-free, permutation-invariant captioning framework that can robustly aggregate heterogeneous biopsy patches and be validated beyond the training institution under realistic cross-institutional conditions.

This study aims to develop a subtype-free, permutation-invariant multi-instance captioning framework for generating descriptive diagnostic reports from gastric adenocarcinoma biopsy WSIs, and to evaluate its generalizability using an independent external clinical dataset.

## Methods

In this study, we developed a model to generate diagnostic reports for pathology images of endoscopic biopsy pathology of gastric adenocarcinoma. For model training and initial validation, In this study, we developed MIAC, a multi-instance attention based captioning model to generate WSI level diagnostic descriptions from gastric adenocarcinoma biopsy WSIs. The model was trained and internally evaluated on the public PatchGastricADC22 dataset and externally tested on an independent clinical dataset from Gil Hospital of Gachon University. External Test validation was performed using an independent dataset of gastric pathology images.

### Training dataset

The Training dataset used in this study is the publicly available PatchGastricADC22 dataset<sup>37</sup>, which contains histopathologic patch images extracted from H&E-stained WSIs of gastric adenocarcinoma endoscopic biopsy samples. This dataset comprises 991 H&E-stained WSIs obtained from individual patients in the surgical pathology archives of the Mita Hospital School of International Health and Welfare (Tokyo, Japan). The WSIs were digitized at 20× magnification and subsequently divided into non-overlapping patch images of size 300×300 pixels, excluding empty regions. Each WSI is paired with a single WSI-level diagnostic caption extracted from the corresponding pathology report.

The diagnostic captions in PatchGastricADC22 were extracted from routine pathology reports associated with each WSI, following the dataset construction described by Tsuneki et al.<sup>37</sup>. Each WSI is paired with a single WSI level caption written in standardized pathology terminology, primarily describing histologic type and differentiation-related features. The original reports were generated by multiple pathologists at a single institution and retrospectively translated into English, resulting in moderate linguistic variability with partially overlapping phrasing across cases. The caption vocabulary consists of 277 words, with a maximum sentence length of 50 words, and does not contain patient-identifiable information.

The dataset was divided into training, validation, and test sets in an approximate ratio of 8: 1: 1, based on WSI<sup>38</sup>. Specifically, the split consisted of 800 slides for training, 91 for validation, and 100 for testing, resulting in a total of 262,777 patch images. On average, there are 265 patch images per slide, with the number of patch images ranging from as little as 2 to as many as 1,860, leading to a highly imbalanced distribution with a large standard deviation. The composition of the dataset is shown in Table 1, and the distribution of patch images per WSI is visualized in Fig. 1.

### External test dataset

The External Test dataset used in this study was collected from Gil Hospital of Gachon University and consisted of histopathological patch images extracted from H&E-stained WSIs of gastric adenocarcinoma endoscopic biopsy samples. The dataset included 105 H&E-stained WSIs, each obtained from an individual patient. Ethical approval for the use of this dataset was granted by the Institutional Review Board (IRB) of the Gil Hospital of Gachon University under protocol number GBIRB2024-121. The WSIs were digitized at 20x magnification and divided into non overlapping patch images, each measuring 300 × 300 pixels, with blank areas excluded. In total, 12,788 patch images were generated, with an average of 121.79 patches per WSI.

| Dataset        | WSIs (Captions) | Patch Images | Patch Images / WSI (P/W) | P/W Std. Dev. |
|----------------|-----------------|--------------|--------------------------|---------------|
| Train Set      | 800             | 210,011      | 262.764                  | 244.178       |
| Validation Set | 91              | 26,170       | 287.582                  | 269.998       |
| Test Set       | 100             | 26,596       | 265.960                  | 220.367       |
| Total          | 991             | 262,777      | 265.163                  | 244.019       |

**Table 1.** PatchGastricADC22 Dataset Composition.

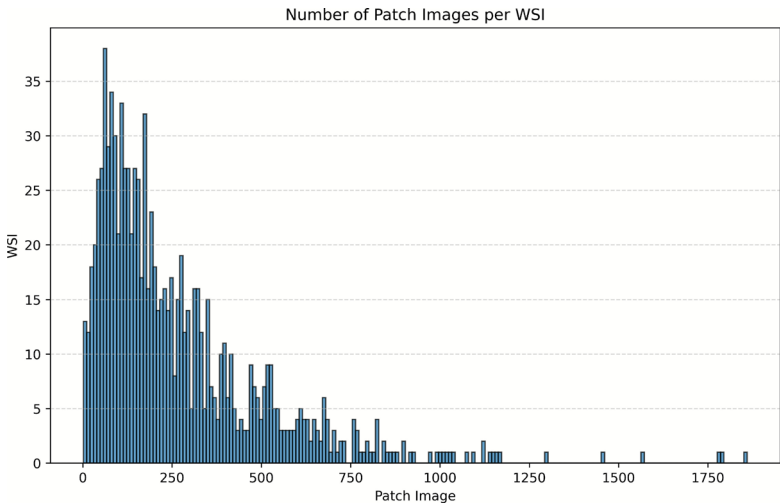


**Fig. 1.** Distribution of the number of patch images per WSI in the Training dataset.

Each WSI was paired with a diagnostic caption written independently by a pathologist following routine clinical reporting practice. Captions were produced as free-text descriptions rather than fixed templates; however, their diagnostic scope and terminology were intentionally aligned with those of the Training dataset to enable meaningful quantitative comparison. In particular, the caption word usage patterns and sentence structures were guided by the PatchGastricADC22 dataset definition<sup>37</sup>, which is derived from routine pathology reports using standardized histopathologic terminology with limited vocabulary and constrained sentence length. The pathologist was fully blinded to MIAC outputs during caption creation to prevent annotation bias. All External Test data were used exclusively for testing to evaluate cross-institutional generalization under real-world domain shift.

To support reliable evaluation under fixed-count random patch sampling, all WSIs in the External Test dataset were curated to be epithelial-region–dominant biopsy samples, with epithelial tissue comprising more than 80% of the tissue area. This design choice reduces the likelihood of sampling non-informative background regions during stochastic patch selection.

The patch images in the External Test dataset exhibited noticeable differences in color compared to those in the Training dataset. These discrepancies were attributed to variations in staining protocols, scanner settings, and other technical factors, which are common in multi-institutional histopathology datasets. Such inconsistencies in the intensity and color of the stain can affect the performance of computational models in image analysis tasks.

To standardize the color distribution across datasets, the H&E-stained images were preprocessed using the Macenko normalization method<sup>39</sup>. The process involved:

1. *Stain separation.* Decomposing each image into Hematoxylin and Eosin components in the optical density (OD) space.
2. *Normalization.* Aligning the stain vectors of each image with the stain vectors of a reference image using principal component analysis (PCA).
3. *Reconstruction.* Converting normalized OD values back into the RGB color space.

This preprocessing step minimized inter-dataset variability by aligning the color characteristics of both Training and External Test datasets to a common color space. Figure 2 shows a comparison of patch images from External Test and Training datasets before and after applying the Macenko normalization. The figure demonstrates the effectiveness of the normalization process in standardizing the appearance of the H&E-stained images.

**Caption generation approach**

The training data used in this study are summarized in Table 1. For each model architecture, 10, 25, and 50 patch images were randomly sampled from each WSI to construct a fixed-size multi-instance input. These patch

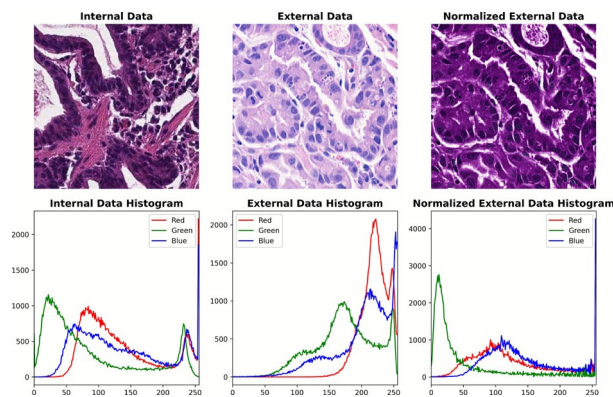


**Fig. 2.** Comparison of patch images and their color histograms from Training and External Test datasets before and after applying Macenko normalization.

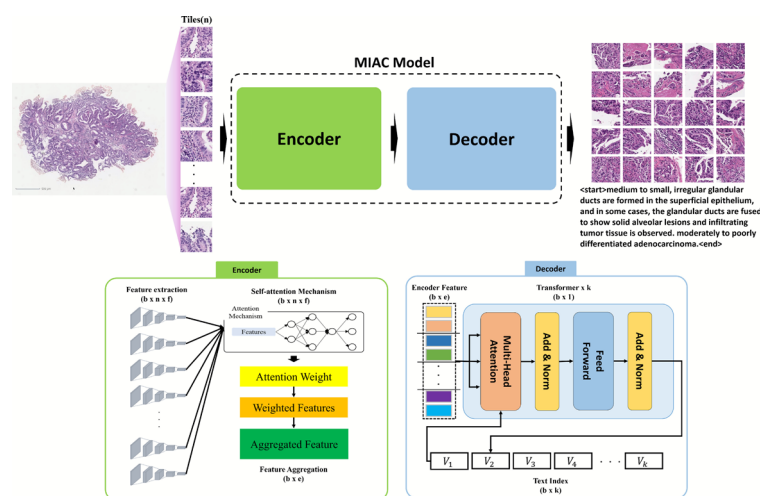


**Fig. 3.** Architecture of the Multi-instance self attention Captioning model (MIAC). The encoder extracts features from patch images and aggregates them, while the decoder uses these features to generate captions.  $b$  batch size,  $n$  number of patch images,  $f$  patch image features,  $e$  embedding size,  $k$  sequence size.

counts were selected as controlled design points to examine the trade-off between slide-level tissue coverage and computational feasibility within a fixed-size multi-instance learning framework. When the number of available patches was fewer than the required input size, sampling with replacement was applied. This method standardized the input sizes across WSIs while retaining the characteristics of the original data distribution.

Each input was subjected to a preprocessing stage before being used in the model. The pixel intensity values of the patch images were normalized from integers in the range of 0–255 to real numbers between  $-1$  and  $1$ . Furthermore, each patch image was randomly rotated by  $0^\circ$ ,  $90^\circ$ ,  $180^\circ$ , or  $270^\circ$  during input processing. These steps adjusted the data to the required scale for the model and introduced controlled variability in spatial orientation. This preprocessing took into account the varying perspectives and orientations commonly observed in histopathology slides, ensuring that the input data were prepared for effective feature extraction<sup>40</sup>.

The output for caption generation in this study was represented as tokenized text. The sentences extracted from each slide were tokenized using the Punkt tokenization method, which segments text based on statistical patterns rather than predefined rules. This method identified sentence boundaries and differentiated abbreviations, numerical expressions, and punctuation frequently present in pathology reports (e.g., abbreviations, alphanumeric expressions, and punctuation frequently appearing in diagnostic pathology reports). Processed text features such as capitalization and mixed alphanumeric content, segmenting input text into structured tokenized sentences suitable for further processing in caption generation tasks<sup>41</sup>.

### Model architecture

The Multi-instance self attention Captioning model (MIAC), illustrated in Fig. 3, comprises two primary components: an encoder and a decoder. The encoder is responsible for extracting features from patch images, while the decoder uses these features to generate diagnostic caption. Using a Multi-instance learning (MIL) approach, the model aggregates features from multiple patch images into a single representative feature vector,

effectively capturing the general characteristics of the original WSI<sup>42,43</sup>. This aggregated feature enables efficient processing of the endoscopic biopsy pathology dataset for gastric adenocarcinoma.

The encoder, as shown in Fig. 3, extracts features from WSI derived patch images. According to MIL principles, the features of each patch are extracted using the EfficientNetV2-S model<sup>44</sup>, which ensures a robust and efficient representation of the image data. To eliminate order dependency among patch images, a self attention mechanism without positional embeddings is applied. This mechanism effectively neutralizes any bias that arises from the sequence of patch images<sup>45,46</sup>. The extracted features are subsequently processed through a Feature Aggregation layer, which consolidates them into a single unified feature vector with a preset embedding size. This aggregated characteristic encapsulates the critical characteristics of WSI, ensuring accurate and efficient downstream analysis<sup>47,48</sup>.

The decoder, as shown in Fig. 3, generates diagnostic caption based on the characteristics extracted by the encoder. The decoder generates captions in an autoregressive manner, predicting one token at a time conditioned on the previously generated tokens and the encoded WSI representation. The encoder's output is an embedding vector of size 1024, which is divided into 16 blocks and passed into a Transformer network. In addition, the value of the previous character is included as an input to the decoder. The first input,  $V_1$ , is fixed to the token '<start>'; and the model begins to generate word indices from  $V_2$ . To accommodate diagnostic caption requirements, the maximum number of characters ( $k$ ) was set to 200, ensuring that the generated output is an array of words containing up to 200 characters.

### Training specifics

The training process used a combination of an efficient image feature extractor and a Transformer based caption decoder to generate accurate captions from the dataset. EfficientNetV2-S, pre-trained on ImageNet, was used as the encoder, providing a feature dimension of 1280. The Transformer based decoder was configured with an embedding size of 1024, a hidden size of 256, four layers, and 16 attention heads. The Adam optimizer was used with a learning rate of  $1 \times 10^{-5}$ ,  $\beta_1 = 0.5$ , and  $\beta_2 = 0.999$ . Cross entropy loss was used to train the sequence generation task, with a batch size of 4. Training was performed for up to 1,000 epochs with early stopping based on validation loss to prevent overfitting. A teacher forcing ratio was applied to the decoder and decayed per epoch according to the equation<sup>49</sup>:

$$\text{Teacher Forcing Ratio} = 0.98^{\text{epoch}}$$

Key hyperparameters were selected to balance slide level coverage, caption completeness, and computational feasibility. Patch counts (10/25/50) were used to examine the effect of increasing WSI coverage under a controlled MIL input size. The maximum caption length was set to  $k = 200$  to cover the typical length of diagnostic style descriptions in the Training dataset while preventing overly long or degenerate generations. The decoder embedding size (1024) and attention configuration were chosen to provide sufficient capacity for pathology specific terminology while maintaining stable optimization with the specified batch size and GPU resources. The training workflow included the following steps. First, the image features were extracted from the augmented image batches using the EfficientNetV2-S encoder. These extracted features were then passed to the Transformer decoder, which generated captions based on the features and the given captions. The generated captions were compared with the target captions and the cross entropy loss was calculated. Gradients were backpropagated to update the weights of both the encoder and decoder. After each epoch, the model was evaluated on the validation dataset and the encoder and decoder checkpoints were saved based on the lowest validation loss observed.

The models were implemented using PyTorch and trained on a NVIDIA A100-SXM4-80GB GPU. To monitor the training process, visual progress bars were used to record the epoch wise training and validation performance, providing real time insights into the trends in loss and the accuracy of the model.

### Performance assessment techniques

Because patch selection is stochastic, especially for WSIs with heterogeneous patch counts, all evaluations were performed using repeated random sampling. For WSIs with fewer patches than the required input size, replacement sampling was applied to maintain a fixed-size MIL input without excluding clinically valid slides. Model performance was therefore reported as mean  $\pm$  standard deviation across repeated sampling runs.

The performance of the caption generation model was evaluated using several standard natural language processing metrics to compare the generated captions with the reference captions. The following metrics were used:

1. **BLEU@4 (Bilingual Evaluation Understudies)**: Evaluates the 4-gram precision of the generated captions in terms of grams compared to the reference captions, taking into account the penalty of brevity and length<sup>50</sup>. The BLEU score is defined as

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

where BP is the penalty for brevity,  $p_n$  is the precision of the  $n$  grams and  $w_n$  is the weight of the precision of the  $n$  gram.

2. **ROUGE-L (Longest Common Subsequence):** Measures the longest common subsequence (LCS) between the generated and reference captions, focusing on capturing structural similarity and the order of words<sup>51</sup>. ROUGE-L is particularly useful for evaluating the fluency and coherence of generated captions. The score is defined as follows:

$$\text{ROUGE-L} = \frac{\text{LCS}}{\text{Length of reference caption}}, \quad (2)$$

where LCS represents the longest common subsequence of words shared between the generated and reference captions. ROUGE-L emphasizes the alignment of words in their proper order, which is crucial to capture the readability and contextual consistency of generated text.

3. **METEOR (Metric for Evaluation of Translation with Explicit Ordering):** Measures caption similarity using unigram matching, incorporating stemming, synonymy matching, and a weighted F-measure<sup>52</sup>. The METEOR score is defined as

$$\text{METEOR} = F_{\beta} \cdot (1 - \text{Penalty}), \quad (3)$$

where  $F_{\beta}$  is the weighted harmonic mean of precision and recall, and Penalty accounts for fragmentation.

4. **CIDEr (Consensus based Image Description Evaluation):** Evaluates the consensus between the generated captions and the reference captions using a weighted n-gram similarity of the frequency inverse document frequency (TF-IDF)<sup>53</sup>. CIDEr emphasizes capturing human consensus and is calculated as:

$$\text{CIDEr} = \frac{1}{N} \sum_{i=1}^N \frac{\sum_{n=1}^N g_n(c_i, r_i)}{\sqrt{\sum_{n=1}^N g_n(c_i, c_i) \cdot \sum_{n=1}^N g_n(r_i, r_i)}}, \quad (4)$$

where  $g_n(c_i, r_i)$  represents the weighted similarity of TF-IDF between the generated caption  $c_i$  and the reference caption  $r_i$  for each n-gram.

These metrics were calculated on the test split of the Training dataset and on the External Test dataset to evaluate the model's ability to generate captions that closely match the reference captions in terms of fluency, accuracy, and relevance.

## Results

In this study, the performance of the proposed MIAC model was evaluated on both the training and External Test datasets, as outlined in the Performance Evaluation Techniques section. The results for the External Test dataset were generated using the model trained exclusively on the Training dataset, without any additional fine tuning or parameter adjustments. Given the inherent characteristics of the model, some variability in the results was observed. This variability arises primarily from stochastic patch sampling during External Test inference, where a fixed number of patches is randomly selected per WSI for caption generation. Detailed information on this extraction process is provided in the Methods section.

To account for the potential variability introduced by this randomness, the MIAC model was evaluated 100 times. The results of these evaluations were averaged to provide stable performance metrics, and the standard deviation (std.) was calculated to quantify the variability between trials. This approach enabled a more detailed and reliable comparison of the performance of the model. The averaged values reflect the central tendency of the model's performance, while the standard deviation indicates the consistency and reproducibility of the results under varying random sampling conditions.

## Training dataset

The Training dataset was evaluated using the set of tests described in Table 1 of the Training dataset section in the methods. The performance of the MIAC model was assessed directly in this test set, while the results for Subtype-guided Masked Transformer (SGMT) and other models were sourced from the paper Subtype-guided Mask Transformer for Pathological Image Captioning<sup>36</sup>, which represents the state-of-the-art (SOTA) for this dataset. Due to differences in test sets, direct comparison under identical conditions was not feasible. The performance metrics for each model are summarized in Table 2.

Table 2 summarizes the performance metrics of various models, including the proposed MIAC model, evaluated in the PatchGastricADC22 dataset. The MIAC model achieved the highest performance across all metrics, particularly when trained with a larger number of patches per WSI.

With 50 training patches per WSI, the MIAC model recorded a BLEU@4 score of 0.617(±0.029), a ROUGE-L score of 0.731(±0.021), a METEOR score of 0.506(±0.021) and a CIDEr score of 5.588(±0.330). These metrics are higher than those achieved by SGMT, which reached its best performance with BLEU @ 4 of 0.551, ROUGE-L of 0.697, METEOR of 0.432 and CIDEr of 4.836 using 32 training patches and upsampling of inference to 64 patches.

| Model( $\times$ Training Patch)        | BLEU@4                      | ROUGE-L                     | METEOR                      | CIDEr                       |
|----------------------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| SGMT $\times$ 1                        | 0.226                       | 0.453                       | 0.325                       | 0.486                       |
| SGMT $\times$ 4                        | 0.299                       | 0.565                       | 0.349                       | 2.013                       |
| MIAC $\times$ 10 (our)                 | 0.398( $\pm$ 0.024)         | 0.576( $\pm$ 0.016)         | 0.369( $\pm$ 0.015)         | 3.183( $\pm$ 0.250)         |
| SGMT $\times$ 16                       | 0.346                       | 0.600                       | 0.353                       | 2.634                       |
| DenseNet121 $\times$ 20 p3x3           | 0.336                       | 0.481                       | 0.284                       | 2.030                       |
| EfficientNetB3 $\times$ 20 p3x3        | 0.364                       | 0.510                       | 0.302                       | 2.154                       |
| DenseNet121 $\times$ 20 p3x3 aug       | 0.347                       | 0.485                       | 0.292                       | 2.024                       |
| EfficientNetB3 $\times$ 20 p3x3 aug    | 0.414                       | 0.540                       | 0.327                       | 2.820                       |
| MIAC $\times$ 25 (our)                 | 0.499( $\pm$ 0.033)         | 0.649( $\pm$ 0.024)         | 0.402( $\pm$ 0.023)         | 4.296( $\pm$ 0.338)         |
| SGMT $\times$ 32                       | 0.528                       | 0.689                       | 0.425                       | 4.656                       |
| SGMT $\times$ 32 Inference $\times$ 64 | 0.551                       | 0.697                       | 0.432                       | 4.836                       |
| SGMT $\times$ 64                       | 0.383                       | 0.594                       | 0.357                       | 3.165                       |
| MIAC $\times$ 50 (our)                 | <b>0.617</b> ( $\pm$ 0.029) | <b>0.731</b> ( $\pm$ 0.021) | <b>0.506</b> ( $\pm$ 0.021) | <b>5.588</b> ( $\pm$ 0.330) |

**Table 2.** Experiments on the PatchGastricADC22 dataset ( $\pm$ std. indicates standard deviation).

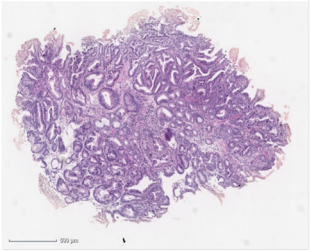
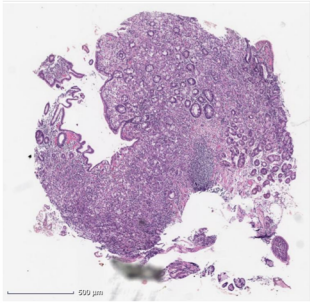
| WSI thumbnail                                                                       | Reference                                                                                                                                                                                                                                 | SGMT                                                                                                                                                                                                                                                           | MIAC                                                                                                                                                                                                                                                             |
|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|   | Tumor tissue consisting of medium sized cord like or irregular glandular infiltration is observed in the superficial epithelium. poorly differentiated adenocarcinoma non solid type or moderately differentiated tubular adenocarcinoma. | In the superficial epithelial layer, infiltrative tumor structures composed of irregular or cord-forming glands are identified, consistent with either a non-solid poorly differentiated adenocarcinoma or a moderately differentiated tubular adenocarcinoma. | Tumor tissue consisting of medium sized cord like or irregular glandular infiltration is observed in the superficial epithelium, compatible with poorly differentiated adenocarcinoma of the non solid type or moderately differentiated tubular adenocarcinoma. |
|  | A tumor tissue consisting of signet ring cell carcinoma cells is observed in the lamina propria. Signet ring cell carcinoma.                                                                                                              | Clusters of malignant cells with signet ring like morphology are identified within the lamina propria, findings suggestive of signet ring cell type adenocarcinoma.                                                                                            | Tumor tissue composed of signet ring cell carcinoma cells is observed in the lamina propria, consistent with signet ring cell carcinoma.                                                                                                                         |

**Fig. 4.** Qualitative comparison of captions generated by different models on the Internal Test dataset..

The MIAC model also showed competitive performance at lower patch counts. For example, with 10 patches per WSI, it achieved a CIDEr score of 3.183( $\pm$ 0.250), which exceeded the score of 2.634 obtained by SGMT with 16 patches. This suggests that the MIAC model captures meaningful features even with a reduced number of training patches, maintaining reliable performance.

The standard deviation (std.) values reported along with the mean scores provide additional information on the consistency of the models. Lower std. values, such as 0.029 for BLEU@4 of MIAC with 50 patches, indicate consistent performance across multiple evaluations. Higher std. values, such as 0.338 for MIAC CIDEr with 25 patches, may reflect variability due to the randomized patch extraction process or differences in data composition during training and evaluation. To complement the quantitative results, we additionally provide a qualitative comparison across models on the Internal Test dataset (Fig. 4). Compared with SGMT, MIAC demonstrates superior preservation of pathology-report-level fidelity in several aspects. First, MIAC more consistently retains standardized diagnostic terminology aligned with established classification systems such as Japanese Gastric Cancer Association<sup>54</sup>, whereas SGMT occasionally omits or alters critical subtype qualifiers (e.g., “non-solid type” in poorly differentiated adenocarcinoma). Second, MIAC better preserves fine-grained morphologic descriptors present in reference reports, while SGMT-generated captions tend to lose specific

details or replace them with broader terms (e.g., loss of “medium-sized” descriptors and substitution of tissue-level expressions with non-specific terms such as “tumor structures”). Third, MIAC maintains diagnostic label accuracy without unnecessary reformulation, whereas SGMT sometimes introduces non-standard or modified diagnostic expressions (e.g., “signet ring cell type adenocarcinoma” instead of the guideline-consistent term “signet ring cell carcinoma”)<sup>54</sup>.

External test dataset

The performance of the MIAC model on the External Test dataset is summarized in Table 3. The External Test dataset, which was used exclusively for testing, evaluated the generalizability of the model trained solely on the Training dataset. To quantify the model’s ability to generate accurate captions for unseen data, the evaluation was conducted both on raw External Test data and on normalized data preprocessed with the Macenko normalization method. Because MIAC inference depends on randomly sampled patches per WSI, all External Test evaluations were repeated 100 times, and mean ± standard deviation values are reported to reflect expected performance under stochastic patch selection rather than a single sampling instance.

When evaluated on raw External Test data, the MIAC model demonstrated a BLEU@4 score of 0.158 (±0.065), a ROUGE-L score of 0.412 (±0.058), a METEOR score of 0.208 (±0.053) and a CIDEr score of 1.853 (±0.750) with 10 patches per WSI. As the number of patches increased to 25 and 50, the model’s performance improved across all metrics, achieving a BLEU@4 score of 0.357 (±0.092), a ROUGE-L score of 0.550 (±0.080), a METEOR score of 0.318 (±0.077), and a CIDEr score of 4.150 (±1.120) at 50 patches per WSI.

On the External Test dataset, MIAC performance increased with the number of sampled patches per WSI for both the raw and Macenko normalized settings (Table 3). For the raw External Test data, MIAC achieved BLEU@4 of 0.158 (±0.065) and CIDEr of 1.853 (±0.750) at 10 patches, improving to BLEU@4 of 0.357 (±0.092) and CIDEr of 4.150 (±1.120) at 50 patches. For the Macenko normalized External Test data, MIAC achieved BLEU@4 of 0.171 (±0.058) and CIDEr of 2.115 (±0.710) at 10 patches, improving to BLEU@4 of 0.375 (±0.080) and CIDEr of 4.382 (±1.010) at 50 patches.

To improve interpretability beyond quantitative metrics, we provide representative qualitative success and failure cases on the External Test dataset (Fig. 5). In successful cases (BLEU@4 0.601), MIAC preserves the overall histologic structure and guideline-consistent diagnostic labels, with only minor loss of fine-grained descriptors such as size modifiers (e.g., omission of “small” in cord-like patterns), while maintaining correct differentiation (grade) and tumor type. In contrast, failure cases (BLEU@4 0.190) are characterized by loss of morphologic specificity and diagnostic precision. Complex architectural features such as glandular duct fusion and solid alveolar patterns are frequently generalized into non-specific descriptions (e.g., “poorly organized architectural patterns”), accompanied by substitution of standardized grading terminology (e.g., “moderately differentiated”) with non-standard expressions (e.g., “intermediate differentiation”). These observations suggest that, under substantial domain shift and architectural heterogeneity, MIAC may favor semantically plausible but diagnostically less precise descriptions, highlighting an important area for further improvement.

Figure 5 presents representative qualitative examples illustrating the range of captioning outcomes observed on the External Test dataset. The top example shows a case in which the generated caption closely aligns with the reference report in terms of overall histologic interpretation and descriptive scope. The middle example represents a partially successful case, where the generated caption is semantically consistent with the reference diagnosis but provides a more generic and less detailed description. In contrast, the bottom example illustrates a failure case, in which the generated caption does not adequately reflect the key diagnostic content of the reference report at the WSI level.

Metrics analysis and comparison

The evaluation metrics BLEU@4, ROUGE-L, METEOR and CIDEr were used to assess the performance of the MIAC model in Training and External Test datasets. Each metric provides unique insights into the quality and fluency of the captions generated, reflecting different aspects of the performance of the model. The detailed results for both datasets are summarized in Tables 2 and 3.

BLEU@4 is a widely used metric for evaluating *n*-gram overlap between generated and reference captions. For interpretability, we follow the rough guideline provided in the official Google Cloud documentation for BLEU scores (reported as percentages): scores below 10 are considered almost useless, 10–19 indicate that it is hard to get the gist, 20–29 suggest that the gist is clear but with significant errors, 30–40 correspond to understandable to good quality, 40–50 to high-quality, 50–60 to very high-quality and fluent outputs, and scores above 60 are often considered better than human translations<sup>55</sup>. Because BLEU values in this study are reported in decimal form (0–1), the above percentage ranges can be interpreted by multiplying our scores by 100.

| DataSet              | Model     | BLEU@4               | ROUGE-L              | METEOR               | CIDEr                |
|----------------------|-----------|----------------------|----------------------|----------------------|----------------------|
| External Test        | MIAC × 10 | 0.158(±0.065)        | 0.412(±0.058)        | 0.208(±0.053)        | 1.853(±0.750)        |
|                      | MIAC × 25 | 0.239(±0.087)        | 0.480(±0.072)        | 0.245(±0.065)        | 2.956(±0.940)        |
|                      | MIAC × 50 | <b>0.357(±0.092)</b> | <b>0.550(±0.080)</b> | <b>0.318(±0.077)</b> | <b>4.150(±1.120)</b> |
| External Test (Nor.) | MIAC × 10 | 0.171(±0.058)        | 0.429(±0.052)        | 0.224(±0.047)        | 2.115(±0.710)        |
|                      | MIAC × 25 | 0.267(±0.068)        | 0.501(±0.064)        | 0.261(±0.058)        | 3.206(±0.890)        |
|                      | MIAC × 50 | <b>0.375(±0.080)</b> | <b>0.577(±0.071)</b> | <b>0.336(±0.069)</b> | <b>4.382(±1.010)</b> |

Table 3. Experiment on External Test dataset (±std. indicates standard deviation).

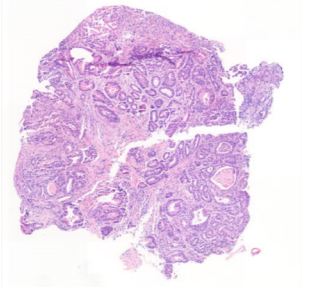
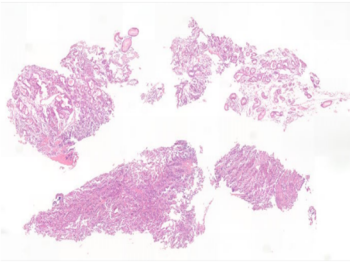
| WSI thumbnail                                                                     | BLEU@4                               | Reference                                                                                                                                                                                                   | MIAC                                                                                                                                                                                                                                      |
|-----------------------------------------------------------------------------------|--------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  | 0.601<br>(overall Match)             | Tumor tissue consisting of medium-sized irregular or small round gland ducts showing a small cord like shape and infiltrating is observed. Well differentiated tubular adenocarcinoma                       | Tumor tissue consisting of medium-sized irregular or small round glandular ducts showing a cord like pattern of infiltration is observed. Well differentiated tubular adenocarcinoma.                                                     |
|  | 0.190<br>(mismatch or partial match) | Medium to small, irregular glandular ducts are formed and the glandular ducts are fused to show solid alveolar lesions and infiltrating tumor tissue is observed. Moderately differentiated adenocarcinoma. | Diffuse epithelial proliferation with poorly organized architectural patterns is present, accompanied by infiltrative growth and compact cellular aggregates. These findings suggest an adenocarcinoma with intermediate differentiation. |

**Fig. 5.** Representative qualitative success and failure cases of MIAC on the External Test dataset..

Although detailed interpretation guidelines for ROUGE-L, METEOR, and CIDEr are not explicitly documented in the existing literature, these metrics provide additional insight into the structural similarity, linguistic fluency, and human like quality of the captions.

Performance differed between the Training and External Test datasets, and Macenko-normalized inputs yielded higher scores than non normalized inputs on the External Test dataset (Tables 2 and 3).

**Discussion**

Understanding the strengths and limitations of the MIAC model is critical to evaluating its potential in real-world applications. The model demonstrated high performance in generating diagnostic captions from Training data, but performance on External Test datasets highlighted challenges associated with domain variability. This discussion focuses on analyzing these results, identifying areas for improvement, and exploring their implications for clinical practice.

**Internal and external test dataset performance and insights**

The evaluation of the MIAC model on the External Test dataset demonstrated both its potential and its limitations in cross-institutional generalization. Although competitive performance was achieved after stain normalization (e.g., BLEU@4 of 0.375(±0.080) and CIDEr of 4.382(±1.010)), these results remained consistently lower than those observed on the Training dataset, underscoring the impact of domain shift arising from inter institutional variability. Although Macenko normalization partially mitigated color related differences, color harmonization alone was insufficient to fully close the performance gap, indicating that additional sources of variability remain.

Beyond quantitative metrics, qualitative analyses clarify how MIAC-generated captions differ from those produced by prior models. As demonstrated by the qualitative comparison on the Internal Test dataset, MIAC consistently generates coherent WSI-level descriptions that emphasize global histologic characteristics and preserve clinically meaningful diagnostic conventions, including standardized terminology, grading, and report-style structuring. In contrast, captions generated by prior models such as SGMT are more often locally focused, partially redundant, and prone to loss of morphologic specificity, frequently relying on non-specific or generalized expressions. These tendencies can lead to subtle but important deviations from guideline-consistent diagnostic terminology. This qualitative advantage of MIAC may stem from its permutation-invariant multi-instance aggregation strategy, which enables robust integration of heterogeneous biopsy patches without reliance on predefined subtype annotations.

Qualitative success and failure cases on the External Test dataset further illustrate these characteristics. In successful cases, MIAC preserves global descriptive consistency and diagnostic terminology despite domain shift, suggesting effective aggregation of diagnostically relevant patterns across patches. In contrast, failure cases are frequently associated with heterogeneous complex tissue architecture or inter institutional differences in diagnostic reporting style, which challenge globally coherent caption generation. These observations highlight both the qualitative strengths of MIAC relative to prior approaches and the practical limitations that remain under real world clinical variability.

In contrast to prior captioning approaches that depend on subtype annotations or order-sensitive sequence modeling, the proposed MIAC framework operates without subtype supervision and explicitly removes positional bias through permutation-invariant self-attention. This design enables robust aggregation of heterogeneous biopsy patches generated during routine preprocessing. Importantly, this study evaluates a pathology caption generation model on an independent External Test clinical dataset curated by pathologists, without additional fine-tuning, supporting its potential for deployment under realistic clinical conditions.

Overall, these findings suggest that MIAC offers meaningful qualitative advantages over existing pathology captioning frameworks through subtype-free, permutation-invariant modeling, while also emphasizing the need for further strategies to improve robustness and consistency across diverse clinical settings.

### Proposed directions for performance improvement

The results underscore the need for further advancements to enhance External Test dataset performance. One promising direction is the enhancement of the Training dataset by incorporating a larger and more diverse collection of WSIs from multiple institutions. This expansion would better capture the variability seen in real-world clinical settings and improve the model's ability to generalize. Furthermore, exploring hybrid approaches that combine multiple normalization techniques or employ advanced domain adaptation methods could further mitigate domain shifts. Model architecture refinements, such as incorporating domain specific regularization, may also help in achieving more robust generalization.

Despite these mitigation strategies, reliance on fixed count random sampling may still result in suboptimal patch sets in rare cases, particularly when tissue composition is highly heterogeneous. To further improve robustness during real-world deployment, future work will explore informed patch selection strategies, such as attention guided sampling to prioritize diagnostically salient patches, region aware sampling based on reduced sensitivity to tissue localization, and diversity aware selection to ensure coverage of multiple histologic patterns within a biopsy.

### Clinical applicability and potential impact

The MIAC model demonstrated its utility in generating diagnostic captions with high accuracy, particularly on the Training dataset. This capability has significant implications for clinical practice, where such models could help pathologists by providing preliminary diagnostic information. Although performance on External Test datasets was less optimal, the model's ability to generate descriptive diagnostic captions, rather than merely identifying diagnostic labels, could still provide valuable reference information for pathologists. This descriptive capability may guide further diagnostic considerations. Furthermore, the model's ability to process data from External Test sources underscores its potential for deployment in multiinstitutional collaborations. However, ensuring reliable performance across diverse datasets is a prerequisite for a real-world application. Developing standardized pre-processing pipelines and creating guidelines for data collection and annotation will be essential for successful implementation.

From a clinical perspective, MIAC is intended as a decision support tool for pathologists. By generating structured descriptive drafts from biopsy WSIs, the model may assist in reducing reporting time, standardizing documentation, and prioritizing diagnostically complex cases. In resource limited settings where pathologists with expertise in gastric cancer are scarce, such automated descriptive summaries may provide consistent reference information to support diagnostic workflows, while final interpretation and confirmation remain under the supervision of board-certified pathologists.

### Future directions for research and development

Future research should focus on validating the performance of the model in larger and more diverse External Test datasets to better understand its generalizability. To expand our pipeline, we plan to incorporate multi resolution WSI patch sampling, allowing the model to capture both fine grained cellular features and broader tissue architecture. In parallel, we aim to develop an adaptive patch selection strategy that leverages attention based heatmaps to focus on diagnostically relevant regions, rather than relying solely on random sampling.

On the modeling side, we intend to enhance the decoder with hierarchical structure prediction, enabling paragraph level or section level captioning aligned with pathologist reporting styles. In addition, incorporating pathology specific pretraining such as histology domain language models or contrastive learning with WSIs could improve the semantic quality of the generated captions.

These extensions are expected to improve the clinical precision, coherence, and interpretability of generated reports, ultimately facilitating smoother integration into digital pathology workflows. In addition, clinical validation studies are necessary to evaluate the practical utility of captions generated in real-world diagnostic workflows. Incorporating user feedback from pathologists could guide iterative improvements to the model, ensuring alignment with clinical needs. Finally, integrating this model into existing digital pathology systems with seamless workflows could pave the way for its adoption in routine clinical practice.

### Limitations of the study

Despite the promising performance of the MIAC model in generating diagnostic captions for gastric adenocarcinoma biopsy images, several limitations must be recognized.

First, the generalizability of the model remains a challenge, as evidenced by its reduced performance on External Test datasets. Variability in staining protocols, scanner settings, and histological classification systems across institutions introduces domain changes that impact model performance. Although Macenko normalization partially mitigated these differences, more research is needed to develop more robust domain adaptation techniques.

Second, the dataset used for model training was restricted to a single institutional gastric biopsy dataset due to scarcity of publicly available dataset with matched WSI level textual annotations. So, to assess cross site generalizability, we additionally evaluated the trained model on an independent External Test dataset collected at a different institution, with and without stain normalization. The inclusion of more diverse datasets from multiple institutions could enhance the robustness of the model and better reflect the clinical variability of the real-world. In addition, the training captions were derived from routine reports generated within a single institutional reporting practice and were constrained by the dataset's limited vocabulary and sentence length, which may not fully capture the broader linguistic variability observed across institutions and real-world pathology reporting styles.

Third, the model's reliance on patch based MIL presents inherent challenges. Variability in the number of patches per WSI can lead to information loss, and the patch selection process may not always capture the most diagnostically relevant regions. Future work should explore adaptive patch selection strategies or incorporate whole slide contextual information to enhance caption generation.

Finally, while the model performed well on natural language processing (NLP) metrics such as BLEU, ROUGE-L, METEOR, and CIDEr, these automated metrics do not fully capture the clinical relevance and precision of the captions generated. Direct validation by expert pathologists is necessary to assess the practical applicability of reports generated in a clinical setting, although such expert centered evaluation was beyond the scope of the present study, future work will incorporate structured pathologist review protocols, including a checklist or Likert scale assessment of histological descriptor correctness, terminology appropriateness, and overall usefulness for diagnostic reporting.

Addressing these limitations through larger and more diverse datasets, improved domain adaptation techniques, and clinical validation studies will be crucial for the real-world deployment of AI assisted pathology captioning models.

Another limitation of this study is the lack of standardization in histological classification systems for gastric cancer across institutions. Multiple classification frameworks, including the WHO classification<sup>56</sup>, Lauren classification<sup>57</sup>, and the Japanese classification<sup>58</sup>, are used in clinical practice, with different institutions adopting different systems. Although these classifications partially overlap, their fundamental differences can lead to variability in diagnostic terminology and reporting style. Such inconsistencies may contribute to residual domain shift in caption generation tasks and limit the generalizability of AI based pathology captioning models across institutions. Addressing this challenge through standardized reporting or harmonized annotation strategies may further improve cross-institutional robustness in future studies.

## Data availability

The publicly available PatchGastricADC22 dataset used in this study can be accessed at: <https://www.kaggle.com/datasets/sanikapadegaonkar/patchgastricadc22> The code used for model training and evaluation is available at: [https://github.com/Leeyoungsup/histopathology\\_captioning](https://github.com/Leeyoungsup/histopathology_captioning) The clinical test dataset collected from Gachon University Gil Medical Center is not publicly available due to patient privacy and institutional data protection policies. However, data may be available from the corresponding author upon reasonable request and with appropriate institutional approvals.

Received: 30 July 2025; Accepted: 22 January 2026

Published online: 12 March 2026

## References

- Sung, H. et al. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: Cancer J. Clin.* **71**, 209–249 (2021).
- Pasechnikov, V., Chukov, S., Fedorov, E., Kikuste, I. & Leja, M. Gastric cancer: prevention, screening and early diagnosis. *World J. Gastroenterol.* **20**, 13842–13862 (2014).
- Ajani, J. A. et al. Gastric cancer, version 2.2022, nccn clinical practice guidelines in oncology. *J. Natl. Compr. Cancer Netw.* **20**, 167–192 (2022).
- Hirasawa, T. et al. Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. *Gastric Cancer* **21**, 653–660 (2018).
- Luo, H. et al. Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study. *Lancet Oncol.* **20**, 1645–1654 (2019).
- Zhu, Y. et al. Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy. *Gastrointest. Endosc.* **89**, 806–815 (2019).
- Jiang, Y. et al. Radiomics signature on computed tomography imaging: association with lymph node metastasis in patients with gastric cancer. *Front. Oncol.* **9**, 340 (2019).
- Wang, Y. et al. Prediction of the depth of tumor invasion in gastric cancer: Potential role of ct radiomics. *Acad. Radiol.* **27**, 1077–1084 (2020).
- Giganti, F. et al. Pre-treatment mdct-based texture analysis for therapy response prediction in gastric cancer: Comparison with tumour regression grade at final histology. *Eur. J. Radiol.* **90**, 129–137 (2017).
- Jiang, Y. et al. Noninvasive imaging evaluation of tumor immune microenvironment to predict outcomes in gastric cancer. *Ann. Oncol.* **31**, 760–768 (2020).
- Zhang, W. et al. Development and validation of a ct-based radiomic nomogram for preoperative prediction of early recurrence in advanced gastric cancer. *Radiother. Oncol.* **145**, 13–20 (2020).
- Niazi, M. K. K., Parwani, A. V. & Gurcan, M. N. Digital pathology and artificial intelligence. *Lancet Oncol.* **20**, e253–e261 (2019).
- Ferreira, R. et al. The virtual microscope. *Proc AMIA Annu Fall Symp*, 449–453 (1997).
- Mukhopadhyay, S. et al. Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: A multicenter blinded randomized noninferiority study of 1992 cases (pivotal study). *Am. J. Surg. Pathol.* **42**, 39–52 (2018).
- Abels, E. et al. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the digital pathology association. *J. Pathol.* **249**, 286–294 (2019).

16. Pantanowitz, L. et al. Validating whole slide imaging for diagnostic purposes in pathology: Guideline from the college of American pathologists pathology and laboratory quality center. *Archives of Pathology. Lab. Med.* **137**, 1710–1722 (2013).
17. Zarella, M. D. et al. A practical guide to whole slide imaging: a white paper from the digital pathology association. *Arch. Pathol. Lab. Med.* **143**, 222–234 (2019).
18. Chen, S. et al. Applications of artificial intelligence in digital pathology for gastric cancer. *Front. Oncol.* **14**, 1437252 (2024).
19. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* **25** (2012).
20. Srinidhi, C. L., Ciga, O. & Martel, A. L. Deep neural network models for computational histopathology: A survey. *Med. Image Anal.* **67**, 101813 (2021).
21. Liang, Q. et al. Weakly supervised biomedical image segmentation by reiterative learning. *IEEE J. Biomed. Health Inform.* **23**, 1205–1214 (2019).
22. Qu, J. et al. Gastric pathology image classification using stepwise fine-tuning for deep neural networks. *J. Healthc. Eng.* **2018**, 8961781 (2018).
23. Abe, H. et al. Development and multi-institutional validation of an artificial intelligence-based diagnostic system for gastric biopsy. *Cancer Sci.* **113**, 3608–3617 (2022).
24. Sharma, H. et al. Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology. *Comput. Med. Imaging Graph.* **61**, 2–13 (2017).
25. Iizuka, O. et al. Deep learning models for histopathological classification of gastric and colonic epithelial tumours. *Sci. Rep.* **10**, 1504 (2020).
26. Park, J. et al. A prospective validation and observer performance study of a deep learning algorithm for pathologic diagnosis of gastric tumors in endoscopic biopsies. *Clin. Cancer Res.* **27**, 719–728 (2021).
27. Yoshida, H. et al. Automated histological classification of whole-slide images of gastric biopsy specimens. *Gastr. Cancer* **21**, 249–257 (2018).
28. Ba, W. et al. Assessment of deep learning assistance for the pathological diagnosis of gastric cancer. *Mod. Pathol.* **35**, 1262–1268 (2022).
29. Lan, J. et al. Using less annotation workload to establish a pathological auxiliary diagnosis system for gastric cancer. *Cell Rep. Med.* **4**, 101004 (2023).
30. Tung, C.-L. et al. Identifying pathological slices of gastric cancer via deep learning. *J. Formosan Med. Assoc.* **121**, 2457–2464 (2022).
31. Marletta, S., Treanor, D., Eccher, A. & Pantanowitz, L. Whole-slide imaging in cytopathology: state of the art and future directions. *Diagn. Histopathol.* **27**, 425–430 (2021).
32. Marletta, S. et al. Application of digital imaging and artificial intelligence to pathology of the placenta. *Pediatr. Dev. Pathol.* **26**, 5–12 (2023).
33. Fu, B. et al. Stohisnet: A hybrid multi-classification model with cnn and transformer for gastric pathology images. *Comput. Methods Programs Biomed.* **221**, 106924 (2022).
34. Kanavati, F. & Tsuneki, M. A deep learning model for gastric diffuse-type adenocarcinoma classification in whole slide images. *Sci. Rep.* **11**, 20486 (2021).
35. Tsuneki, M. & Kanavati, F. Inference of captions from histopathological patches. *Proc. Mach. Learn. Res.* **1235**, 1235–1250 (2022).
36. Qin, W. et al. What a whole slide image can tell? subtype-guided masked transformer for pathological image captioning. [arXiv:2310.20607](https://arxiv.org/abs/2310.20607) (2023).
37. Tsuneki, M. & Kanavati, F. Inference of captions from histopathological patches. In Konukoglu, E. et al. (eds.) *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning*, vol. 172 of *Proceedings of Machine Learning Research*, 1235–1250 (PMLR, 2022).
38. Cong, G. & Fung, V. Improving materials property predictions for graph neural networks with minimal feature engineering. *Mach. Learn. Sci. Technol.* **4**, 035030 (2023).
39. Macenko, M. et al. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE international symposium on biomedical imaging: from nano to macro*, 1107–1110 (IEEE, 2009).
40. Lee, H., Lee, K., Lee, K., Lee, H. & Shin, J. Improving transferability of representations via augmentation-aware self-supervision. *Adv. Neural Inf. Process. Syst.* **34**, 17710–17722 (2021).
41. Griffis, D., Shivade, C., Fosler-Lussier, E. & Lai, A. M. A quantitative and qualitative evaluation of sentence boundary detection for the clinical domain. *AMIA Summits Transl. Sci. Proc.* **2016**, 88 (2016).
42. Liang, M. et al. Caf-ahgcn: context-aware attention fusion adaptive hypergraph convolutional network for human-interpretable prediction of gigapixel whole-slide image. *Vis. Comput.* **40**(12), 8747–65 (2024).
43. Cai, C. et al. Pathologist-level diagnosis of ulcerative colitis inflammatory activity level using an automated histological grading method. *Int. J. Med. Inform.* **192**, 105648 (2024).
44. Wang, Z. & Nirjon, S. Characterizing disparity between edge models and high-accuracy base models for vision tasks. [arXiv:2407.10016](https://arxiv.org/abs/2407.10016) (2024).
45. Seo, M., Lee, H.-J. & Nguyen, X. T. Vit-p3de\*: Vision transformer based multi-camera instance. In *IJCAI*, 1340–1350 (2023).
46. Fu, L. et al. Rethinking patch dependence for masked autoencoders. [arXiv:2401.14391](https://arxiv.org/abs/2401.14391) (2024).
47. Guo, Z. et al. Histgen: Histopathology report generation via local-global feature encoding and cross-modal context interaction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 189–199 (Springer, 2024).
48. Das, K., Conjeti, S., Chatterjee, J. & Sheet, D. Detection of breast cancer from whole slide histopathological images using deep multiple instance cnn. *IEEE Access* **8**, 213502–213511 (2020).
49. He, T., Zhang, J., Zhou, Z. & Glass, J. Quantifying exposure bias for neural language generation (2019).
50. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318 (2002).
51. Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81 (2004).
52. Banerjee, S. & Lavie, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72 (2005).
53. Vedantam, R., Lawrence Zitnick, C. & Parikh, D. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575 (2015).
54. Japanese Gastric Cancer Association jgca@ koto. kpu-m. ac. jp. Japanese gastric cancer treatment guidelines 2018 Japanese gastric cancer association. *Gastr. Cancer* **24**, 1–21 (2021).
55. Google Cloud. Evaluating models – understanding the bleu score – interpretation. (accessed 06 Feb 2022) <https://cloud.google.com/translate/automl/docs/evaluate#interpretation> (2022).
56. WHO Classification of Tumours Editorial Board. *Digestive system tumours*, 5th edn (International Agency for Research on Cancer, Lyon, 2019).
57. Lauren, P. The two histological main types of gastric carcinoma: diffuse and so-called intestinal-type carcinoma: an attempt at a histo-clinical classification. *Acta Pathol. Microbiol. Scand.* **64**, 31–49 (1965).
58. Japanese Gastric Cancer Association jgca@ koto. kpu-m. ac. jp. Japanese classification of gastric carcinoma: 3rd English edition. *Gastr. Cancer* **14**, 101–112 (2011).

### Author contributions

Youngseop Lee conceptualized the study, developed the methodology, conducted the experiments, and analyzed the results. Kyungah Bai contributed to data collection, validation of experimental results, and manuscript editing. Youngjae Kim provided technical support for model implementation and contributed to the preprocessing pipeline. Jisup Kim and Kwanggi Kim supervised the study, served as corresponding authors, and provided pathological and experimental guidance. All authors reviewed and approved the final manuscript.

### Funding

This work was supported by the Digital Medical Products Development Based on Medical Data Synthesis and AI Technologies Program(RS-2025-02305698, Development of On-Device AI Digital Medical Products Utilizing Synthetic Technology and Synthetic Data for Atypical Medical Data) funded by the Ministry of Trade, Industry&Energy(MOTIE) of Korea. This work was supported by the Gachon University research fund of 2024 (GCU-2024-202410530001).

### Declarations

### Competing interests

The authors declare no competing interests.

### Ethical approval

This study was approved by the Institutional Review Board (IRB) of Gil Hospital of Gachon University (Approval Number: GBIRB2024-121). All experimental protocols were conducted in accordance with relevant guidelines and regulations, strictly adhering to the ethical principles outlined in the Declaration of Helsinki.

### Informed consent

The requirement for informed consent was waived due to the retrospective nature of the study design.

### Additional information

**Correspondence** and requests for materials should be addressed to J.K. or K.G.K.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026