computers

MDPI

Article

Cytological Image-Finding Generation Using Open-Source Large Language Models and a Vision Transformer

Atsushi Teramoto [1,*], Yuka Kiriyama [2], Tetsuya Tsukamoto [3], Natsuki Yazawa [4], Kazuyoshi Imaizumi [5] and Hiroshi Fujita [6]

[1] Faculty of Information Engineering, Meijo University, Nagoya 468-8502, Japan
[2] Narita Memorial Hospital, Toyohashi 441-8029, Japan
[3] Oncology Innovation Center, Fujita Health University, Toyoake 470-1192, Japan
[4] Fujita Health University Hospital, Toyoake 470-1192, Japan
[5] School of Medicine, Fujita Health University, Toyoake 470-1192, Japan
[6] Faculty of Engineering, Gifu University, Gifu 501-1193, Japan
* Correspondence: teramoto@meijo-u.ac.jp

Abstract

In lung cytology, screeners and pathologists examine many cells in cytological specimens and describe their corresponding imaging findings. To support this process, our previous study proposed an image-finding generation model based on convolutional neural networks and a transformer architecture. However, further improvements are required to enhance the accuracy of these findings. In this study, we developed a cytology-specific image-finding generation model using a vision transformer and open-source large language models. In the proposed method, a vision transformer pretrained on large-scale image datasets and multiple open-source large language models was introduced and connected through an original projection layer. Experimental validation using 1059 cytological images demonstrated that the proposed model achieved favorable scores on language-based evaluation metrics and good classification performance when cells were classified based on the generated findings. These results indicate that a task-specific model is an effective approach for generating imaging findings in lung cytology.

Keywords: cytology; finding; report generation; image captioning; large language model

1. Introduction

Lung cancer is a leading cause of death worldwide and has the highest mortality rate among men [1]. The prognosis of lung cancer depends strongly on the stage at diagnosis, making early detection and treatment critically important. Consequently, there is a growing demand for supportive technologies that improve both diagnostic accuracy and efficiency throughout the diagnostic process, including medical imaging and pathological examination.

Pathological examination is used for the definitive diagnosis of lung cancer, and cytological examination is often performed at an early stage because it imposes a relatively low burden on patients and allows rapid specimen preparation [2]. For cytology, specimens prepared from the collected cells are examined under a microscope by cytotechnologists who performed screening to assess the presence of abnormal findings. Subsequently, cytopathologists make a comprehensive diagnosis based on these findings and prepare a cytology report [3]. These reports describe the cytological findings observed in microscopic images, such as cellular morphology, arrangement, and nuclear characteristics, in textual form.

Academic Editor: M. Ali Akber Dewan

Received: 11 January 2026
Revised: 3 February 2026
Accepted: 6 February 2026
Published: 8 February 2026

Although cytological findings play a crucial role in diagnosis, the process of observing microscopic images and describing the findings in written form is labor-intensive and places a substantial burden on diagnosticians. In addition, there is a global shortage of pathologists and cytology specialists [4], further increasing the need for technologies that support the efficient generation of diagnostic reports.

Therefore, image-captioning techniques that generate textual descriptions from image content using deep learning have attracted increasing attention, and their application in the medical imaging domain has been actively explored. In studies focusing on X-ray images, Wang et al. proposed TieNet, which combines a convolutional neural network (CNN) [5] and a recurrent neural network (RNN) [6], and demonstrated a method for simultaneously performing disease classification and radiology report generation from chest X-ray images [7]. In addition, Hou et al. proposed RATCHET, a medical domain–specific model based on a transformer [8], and reported that it achieved superior diagnostic and image-finding generation performance compared with conventional methods [9].

In recent years, the application of large language models (LLMs) [10] to radiology report generation has been widely investigated. For example, studies have examined the feasibility of radiology report generation using GPT series models developed by OpenAI and have reported improvements in textual quality and clinical consistency through the introduction of LLMs [11]. Furthermore, approaches targeting structured report formats have been proposed for generating chest X-ray reports [12].

Report-generation methods for histopathological imaging have also been reported in the field of pathology. Zhou et al. proposed the GNNFormer, which represents the relationships between cells and tissue structures in pathological images as graphs and generates pathology reports by combining this representation with a transformer [13]. Recently, methods for generating pathology reports based on LLMs have been proposed. HistoGPT targets high-resolution histopathological images and integrates a vision encoder with an LLM to form a vision–language model (VLM), thereby enabling automatic generation of full pathology reports [14].

Previous studies have primarily investigated methods for generating textual descriptions from medical images, focusing mainly on X-ray and histopathological images. However, studies targeting cytological images and automatically generating image findings based on cellular morphology and spatial arrangement observed under a microscope are limited. In our previous work, we proposed a method that classified cytological images into benign and malignant categories using a CNN and generated image findings by switching between transformer decoders trained separately for benign and malignant cases according to the classification results [15]. The evaluation demonstrated that the agreement between the generated findings and the reference findings was higher than that achieved by conventional approaches, and favorable performance was also obtained for benign–malignant discrimination. However, the classification performance for histological subtypes of lung cancer, such as adenocarcinoma and squamous cell carcinoma, was insufficient, and additional limitations remained, including the absence of small cell carcinoma cases in the dataset.

In cytology, in addition to benign–malignant discrimination, a certain level of accuracy is required for the estimation of histological subtypes. Therefore, developing more accurate image-captioning models is necessary. In the field of image classification, vision transformers (ViT) have demonstrated high performance through pre-training on large-scale image datasets [16]. Similarly, in the field of language modeling, LLMs pretrained on massive text corpora have been developed. Recently, open-source (open-weight) LLMs such as GPT-OSS [17] and Qwen [18] have become available for research purposes. Although our previous study employed a relatively small transformer, utilizing an LLM as the net-

work responsible for text generation has the potential to further improve the accuracy of generated images.

Accordingly, in this study, we developed an image captioning model that combines ViT with open-source LLMs, aiming to improve the image-finding generation performance for cytological images. The proposed model was trained and evaluated using a visual language dataset composed of cytological images and corresponding captions that we constructed independently. Furthermore, the effectiveness of the proposed method is validated through a comparative evaluation using multiple language-based metrics.

The main contributions of this paper are summarized as follows:

- Proposal of a cytology-specific VLM leveraging open-source LLMs:

We propose a vision–language model tailored for cytological analysis by combining image feature extraction using a vision transformer with text generation based on an LLM, enabling the generation of image findings suitable for the cytology domain.

- Improved image-finding generation and histological subtype estimation from cytological images:

By analyzing the generated image findings, we performed a histological subtype estimation of lung cancer and achieved a higher performance than that reported in previous studies. In particular, cytology-specific descriptions such as cellular morphology, spatial arrangement, and nuclear findings showed good agreement with the reference descriptions.

- Demonstration of superiority over general-purpose VLMs:

Through comparisons with existing general-purpose VLMs, we demonstrate that domain-specific training on cytological data leads to superior performance in both image-finding and histological subtype estimation.

2. Materials and Methods

2.1. Image Dataset

For this study, we constructed an original dataset consisting of paired cytological images and the corresponding image findings. An overview of the workflow from patient biopsy to dataset construction is shown in Figure 1.

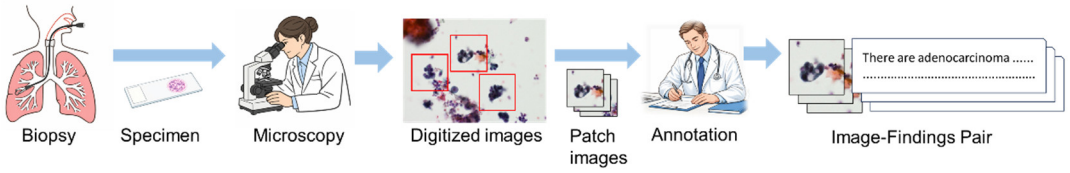


Figure 1. Workflow of dataset construction for cytological image–finding pairs. Red rectangles indicate the selected patch regions used for analysis.

Lung cytology specimens were obtained from 279 patients using interventional cytology procedures, including bronchoscopy and computed tomography-guided fine-needle aspiration cytology. The data collected included 82 benign and 197 malignant cases. The malignant cases included 93 adenocarcinomas, 54 squamous cell carcinomas, and 50 small cell carcinomas. The histological subtypes of the malignant cases were determined by integrating cytological findings with histological analyses of biopsy specimens. Biopsy tissues were collected concurrently with the cytological specimens, fixed in 10% neutral-buffered formalin, dehydrated, and embedded in paraffin. Cytological specimens were prepared using the BD SurePath liquid-based cytology (LBC) system (Becton Dickinson, Franklin Lakes, NJ, USA) and stained using the Papanicolaou method. Microscopic images were acquired using a microscope (BX53, Olympus Corporation, Tokyo, Japan) equipped with

a digital camera (DP20, Olympus Corporation) and stored in JPEG format at a resolution of 1280 × 960 pixels. In total, 261 benign cell images and 655 malignant cell images were obtained, including 211 adenocarcinoma, 281 squamous cell carcinoma, and 163 small cell carcinoma images.

To construct an image dataset suitable for cytological image finding, cytotechnologists and cytopathologists extracted patch images of 296 × 296 pixels from the original microscopic images, focusing on regions containing the target cells. The final dataset consisted of 1059 patch images, including 373 benign cell images and 686 malignant cell images (228 adenocarcinoma, 283 squamous cell carcinoma, and 175 small cell carcinoma images). For each image, the microscopic image findings were described in text, covering the cellular type, nuclear morphology, cellular arrangement, and background conditions unrelated to the target cells. The image findings were recorded by one cytotechnologist and one cytopathologist in accordance with the World Health Organization (WHO) reporting system for lung cytology [19]. Based on these procedures, a dataset of 1059 paired cytological images and the corresponding image findings was constructed. Representative images and image findings are shown in Figure 2.

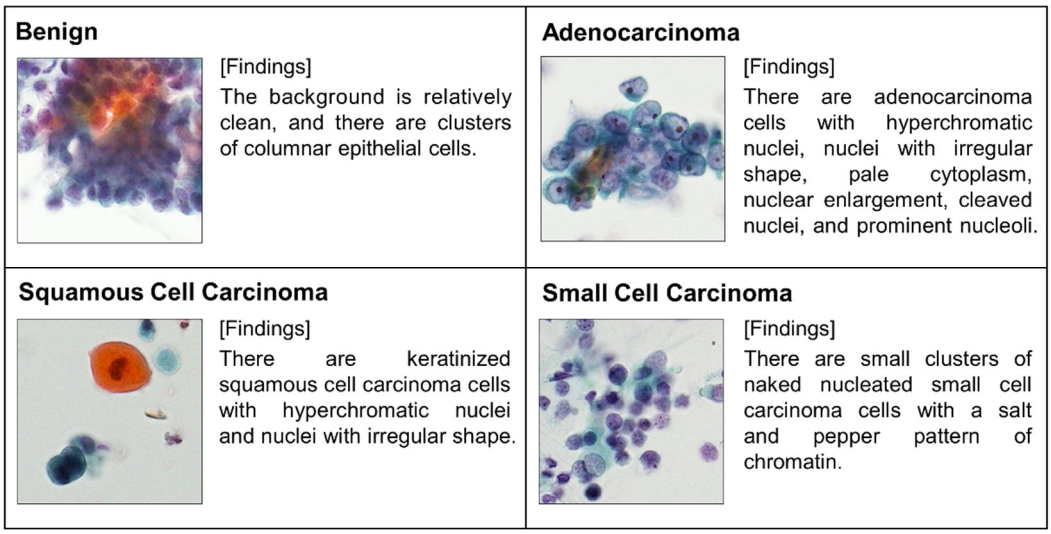


Figure 2. Sample of image-findings pairs.

The dataset was divided into 883 image–caption pairs for training and 176 pairs for evaluating the caption generation model. To prevent data leakage, images derived from the same patients were not shared between the training and evaluation datasets.

2.2. Outline of Findings Generation Model

An overview of the proposed image-finding generation model is presented in Figure 3. Cytological images are first input into a vision model, and the extracted image features are passed to a language model through a projection layer. Based on these features, the language model generates a sequence of tokens that are outputted as image-finding descriptions.

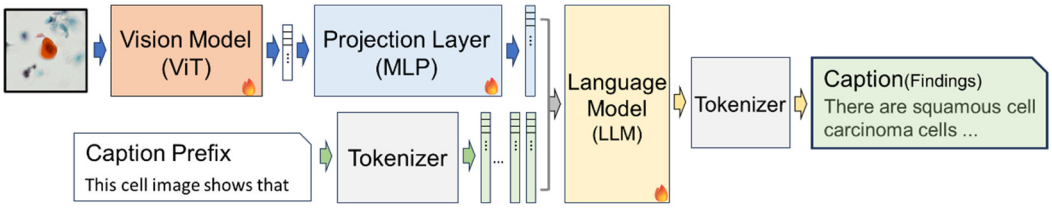


Figure 3. Schematic overview of the cytologic finding generation model.

2.3. Vision Model

The vision model analyzes input cytological images and extracts the visual features required for image-finding generation. In this study, feature extraction was performed using a ViT, which has demonstrated high performance in image recognition tasks. ViT has been widely adopted as an image encoder in contrastive language–image pretraining (CLIP) frameworks [20], in which visual–language correspondences are learned through contrastive learning, enabling the extraction of semantic visual features aligned with language representations. Unlike ViT models pretrained solely for image classification tasks, the ViT integrated into CLIP is trained to align visual representations with language representations through large-scale image–text pairs, which is particularly suitable for VLM construction.

For the vision model, we employed a ViT model (base architecture with a patch size of 16 × 16 pixels) integrated into OpenCLIP [21], which was pretrained using the LAION-2B dataset [22]. OpenCLIP is an open-source implementation of CLIP trained on large-scale image–text pairs, allowing ViT to acquire visual feature representations that capture correspondences between images and language.

Using this model, a 512-dimensional image feature vector is extracted from each input image and subsequently used as an input for the downstream language model.

2.4. Language Model

The language model generated image-finding descriptions corresponding to the cytological images based on the visual features extracted by the vision model. In general, LLMs generate text by taking an input prompt, such as a query or caption prefix, and predicting the subsequent text. In this study, the caption prefix "*This cell image shows that*" was used to initiate image-finding generation.

Specifically, the caption prefix was tokenized using a tokenizer, and the resulting token embeddings were combined with the image feature vector extracted using the vision model. The image features were mapped into the embedding space of the language model through a projection layer, as described in the following section. The concatenated embeddings were then provided as inputs to the LLM. Based on these inputs, the LLM sequentially generated a token sequence, which was subsequently converted into text using the tokenizer to obtain the final image-finding description.

Performance comparisons were conducted using open-source LLMs. Specifically, the GPT-OSS-20B and GPT-2 models developed and released by OpenAI [23] and the Qwen 2.5 14B model and Qwen 3 14 B models developed by Alibaba were evaluated in combination with a vision model. The token embedding dimensions of the LLMs were 2880, 1024, 5120, and 5120, respectively.

2.5. Projection Layer

In the projection layer, the image feature vector extracted by the vision model is treated as a single token and mapped onto the token embedding space of the LLM. As vision and language models differ in both feature dimensionality and distribution, a dimensional transformation is required to connect them.

In this study, a multilayer perceptron (MLP) was employed as the projection layer. The MLP uses a 512-dimensional image feature vector obtained from the vision model as input and outputs a feature vector with the same dimensionality as the token embedding layer of the LLM through a hidden layer consisting of 1024 units.

2.6. Training Strategy

Different training strategies were adopted for the vision model, language model, and projection layer. This design was motivated by the distinct roles of each network, availability of pretrained weights, and constraints related to the computational cost and amount of training data.

For the vision model, to adapt the pretrained image feature extraction capability to cytological images, only the first six layers of the ViT were fine-tuned, while the remaining layers were kept frozen. This strategy allows the model to preserve representations learned during pretraining, while enabling feature extraction tailored to cytological images.

For the language model, low-rank adaptation (LoRA) [24] was applied to reduce the computational cost and address limitations in the size of the training dataset. During training, only the parameters introduced by LoRA were updated, and the original weights of the language model were fixed. The LoRA configuration was set with a rank of 8, alpha value of 16, and a dropout rate of 0.05.

The projection layer, which was newly introduced to connect the vision and language models, was trained by updating all its parameters.

The batch size was set to 8. The learning rate was set to 1×10^{-5} for the language model and projection layer, and to 1×10^{-6} for the vision model. AdamW was used as the optimizer. The weight decay was set to 0 for the language model and 0.01 for the vision model and projection layer. Training was conducted for up to 50 epochs and 40 samples randomly selected from the training data were used as validation data. Training was automatically terminated when no further improvement was observed in the Consensus-based image description evaluation (CIDEr) score (described in the next section) [25] of the validation data.

All training and inference procedures were performed using an NVIDIA DGX H200 system equipped with eight H200 GPUs.

2.7. Evaluation Metrics

To quantitatively evaluate the validity of the generated image-finding descriptions, multiple language-based evaluation metrics were calculated on an evaluation dataset consisting of 176 cases based on the degree of agreement with the ground-truth texts (reference descriptions). These metrics assess the lexical and semantic similarities between generated descriptions and reference texts from different perspectives.

- Bilingual evaluation understudy (BLEU) [26]:

BLEU is a metric that evaluates text generation performance based on the n-gram overlap between generated descriptions and reference texts. In this study, BLEU-1 to BLEU-4, which consider 1 g to 4 g matches, were used to assess the extent to which the n-grams in the generated descriptions matched those in the reference texts.

- Metric for evaluation of translation with explicit ordering (METEOR) [27]:

METEOR evaluates text similarity by considering not only word-level matches, but also synonyms and morphological variations. As it can flexibly capture the correspondence between generated and reference texts, it is suitable for evaluating descriptions that include medical terminology.

- Recall-oriented understudy for gisting evaluation (ROUGE) [28]:

ROUGE measures the overlap of n-grams between generated and reference texts from a recall-oriented perspective. This study used it to evaluate how important phrases contained in the reference texts were reflected in the generated descriptions.

- Consensus-based image description evaluation (CIDEr):

CIDEr emphasizes the importance of image-specific terms by applying TF–IDF (term frequency–inverse document frequency) weighting. Given its ability to highlight domain-specific terminology and characteristic findings in medical images, CIDEr was used as a primary evaluation metric in this study.

- Semantic propositional image caption evaluation (SPICE) [29]:

SPICE evaluates semantic similarity based on scene graphs constructed from generated and reference texts. Unlike surface-level word matching, it enables assessment focused on the overall semantic structure of descriptions.

In this study, the target cytological images were broadly categorized into benign and malignant cells, and the malignant cells were further classified into three types of lung cancer cells: adenocarcinoma, squamous cell carcinoma, and small cell carcinoma. Accordingly, we evaluated the accuracy of these categories by analyzing the generated image descriptions.

First, classification performance was evaluated for the two categories of benign and malignant cells. If the generated image finding description contained characteristic terms suggestive of malignancy, such as "*carcinoma*," "*cancer*," or "*abnormal*," the case was classified as malignant; otherwise, it was classified as benign. Based on these classification results, the sensitivity and specificity were calculated by considering malignant cases as positive. The definitions of the sensitivity and specificity are given in Equation (1), where *TP* (true positive) denotes the number of images correctly classified as malignant, true negative (*TN*) denotes the number of images correctly classified as benign, false positive (*FP*) denotes the number of benign images incorrectly classified as malignant, and false negative (*FN*) denotes the number of malignant images incorrectly classified as benign:

$$Sensitivity = \frac{TP}{TP + FN} \tag{1}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

Next, classification performance was evaluated for four categories: benign, adenocarcinoma, squamous cell carcinoma, and small cell carcinoma. The generated image-finding descriptions were analyzed, and cases judged as malignant were further classified as adenocarcinoma, squamous cell carcinoma, or small cell carcinoma based on textual descriptions. These results were combined with the benign cases to construct a confusion matrix from which the overall classification accuracy was calculated.

Because the number of images differed among the four categories in this study, accuracy was computed separately for each category to avoid bias arising from class imbalance. The final overall accuracy was obtained by averaging the accuracies of the four categories.

In addition to conventional language similarity metrics and classification-based evaluation, we introduced task-specific evaluation metrics based on the WHO reporting system for lung cytopathology. In cytological diagnosis, accurate reporting requires not only correct disease categorization but also appropriate descriptions of morphological findings that support the diagnosis. Accordingly, for each malignant category—adenocarcinoma, squamous cell carcinoma, and small cell carcinoma—we defined two key diagnostic features, as summarized in Table 1. We evaluated whether these features described in the reference captions were also preserved in the generated captions, which was quantified as the key-feature recall. Furthermore, from a clinical perspective, the erroneous inclusion of morphological features characteristic of other histological subtypes may lead to diagnostic confusion. Therefore, we additionally calculated the false mention rate, defined as the proportion of generated captions that incorrectly included key diagnostic features that should not appear in the corresponding histological category.

Table 1. Definition of key diagnostic features for task-specific evaluation based on the WHO reporting system for lung cytopathology.

Histological Subtype	Key Diagnostic Feature 1	Key Diagnostic Feature 2
Adenocarcinoma	Pale cytoplasm	Hyperchromatic nuclei
Squamous cell carcinoma	Hyperchromatic nuclei	Keratinization (Orange G)
Small cell carcinoma	Poor cytoplasm (High N/C ratio)	Salt-and-pepper chromatin

To analyze which image regions contributed to the generated captions and to improve the interpretability of the visual encoder, we conducted a patch ablation–based visualization analysis [30]. Specifically, the input image was divided into non-overlapping patches corresponding to the visual encoder input resolution. Each patch was sequentially masked, and the change in the model output was measured. Based on these output variations, a heatmap was generated to indicate image regions that strongly influenced the caption generation. This analysis was applied to representative test images, and the resulting heatmaps were used for qualitative interpretation of the model behavior. This visualization does not provide a complete explanation of the model decisions, but offers insights into spatial regions that are relevant to the generated diagnostic descriptions.

In this paper, we propose an original image-captioning model that integrates a vision model with an LLM. However, several VLMs capable of image captioning have recently been developed. To verify the effectiveness of the proposed method, comparative experiments were conducted using existing VLMs.

The BLIP2 OPT-6.7B model [31] and Qwen2.5 VL 32 B model [32] were selected as the baseline models. For these models, LoRA was applied, and fine-tuning was performed using the cytological image caption dataset constructed in this study. The LoRA configuration was set with a rank of 8 and an alpha value of 16. After fine-tuning, the image-finding descriptions generated from each model were evaluated using the aforementioned language-based evaluation metrics and image classification performance metrics, and their performances were compared with that of the proposed method.

3. Results

Table 2 presents the results of image-finding generation on the test images for the evaluated models, along with the corresponding BLEU, ROUGE, METEOR, CIDEr, and SPICE scores calculated using the reference descriptions. All metrics yield non-negative values; a higher CIDEr score indicates better performance, whereas values closer to 1 indicate better performance for the other metrics.

Table 2. Language evaluation metrics comparing proposed and existing models.

Model	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE_L	CIDEr	SPICE
ViT + GPT-oss 20B	0.859	0.811	0.768	0.735	0.506	0.855	5.990	0.760
ViT + GPT2 Medium	0.694	0.640	0.590	0.557	0.389	0.723	4.345	0.642
ViT + Qwen3 14B	0.863	0.812	0.767	0.734	0.497	0.849	5.796	0.746
ViT + Qwen2.5 14B	0.838	0.783	0.734	0.698	0.482	0.822	5.427	0.725
Our Previous Method (CNN + Transformer)	0.797	0.737	0.685	0.648	0.444	0.776	5.060	0.686
BLIP2 (OPT-6.7B)	0.763	0.694	0.633	0.589	0.417	0.742	4.391	0.624
Qwen2.5 VL 7B	0.764	0.701	0.647	0.606	0.422	0.737	4.513	0.630
Qwen2.5 VL 32B	0.816	0.761	0.713	0.678	0.469	0.792	5.160	0.707

Next, the image-finding descriptions generated by several models were analyzed to classify the benign cells and histological subtypes (adenocarcinoma, squamous cell

carcinoma, and small cell carcinoma). The classification performance metrics, including sensitivity and specificity with 95% confidence intervals and the corresponding confusion matrices, are presented in Tables 3 and 4, respectively.

Table 3. Image classification evaluation metrics of proposed and existing models.

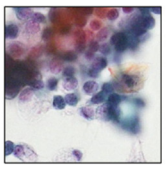
	Balanced Accuracy	**Sensitivity (95% CI)**	**Specificity (95% CI)**
ViT + GPT-oss 20B	0.858	0.992 (0.958–0.999)	1.000 (0.920–1.000)
ViT + GPT2 Medium	0.824	1.000 (0.972–1.000)	0.977 (0.882–0.996)
ViT + Qwen3 14B	0.835	1.000 (0.972–1.000)	1.000 (0.920–1.000)
ViT + Qwen2.5 14B	0.834	0.985 (0.946–0.996)	0.977 (0.882–0.996)
Our Previous method (CNN + Transformer)	0.711	0.909 (0.848–0.947)	0.932 (0.818–0.977)
BLIP2 (OPT-6.7B)	0.674	0.955 (0.904–0.979)	0.705 (0.558–0.818)
Qwen2.5 VL 32B	0.801	0.985 (0.946–0.996)	0.909 (0.788–0.964)

Table 4. Confusion matrices of image classification.

(a) ViT + GPT-oss 20B					
		Predicted			
		Benign	Adenocarcinoma	Squamous cell carcinoma	Small cell carcinoma
Actual	Benign	44	0	0	0
	Adenocarcinoma	0	30	5	0
	Squamous cell carcinoma	1	24	34	0
	Small cell carcinoma	0	0	0	38
(b) Our previous method					
		Predicted			
		Benign	Adenocarcinoma	Squamous cell carcinoma	Small cell carcinoma
Actual	Benign	41	2	1	0
	Adenocarcinoma	3	23	9	0
	Squamous cell carcinoma	5	25	29	0
	Small cell carcinoma	4	2	3	29
(c) Qwen2.5 VL 32B					
		Predicted			
		Benign	Adenocarcinoma	Squamous cell carcinoma	Small cell carcinoma
Actual	Benign	40	0	4	0
	Adenocarcinoma	0	24	10	1
	Squamous cell carcinoma	2	19	36	02
	Small cell carcinoma	0	0	0	38

Representative examples of image-finding descriptions generated for benign and malignant cells using the selected models are shown in Figure 4.

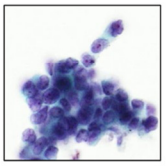
< Benign cells >



Input image

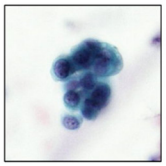
Gold Standard	The background is relatively clean, and there are clusters of columnar epithelial cells.
ViT+GPT-oss 20B	The background is relatively clean. There are small clusters of columnar epithelial cells.
Our previous	The background is relatively clean and there are clusters of columnar epithelial cells.
Qwen2.5 VL 32B	There are squamous cell carcinoma cells with thick cytoplasm, hyperchromatic nuclei, and nuclei with irregular shape.

< Adenocarcinoma cells >



Input image

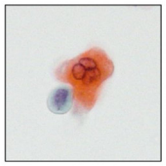
Gold Standard	There are adenocarcinoma cells with hyperchromatic nuclei, nuclei with irregular shape, pale cytoplasm, and eccentric nuclei.
ViT+GPT-oss 20B	There are adenocarcinoma cells in glandular structure or papillary clusters with hyperchromatic nuclei, nuclei with irregular shape, and pale cytoplasm.
Our previous	The background is relatively clean and there are clusters of columnar epithelial cells.
Qwen2.5 VL 32B	There are clusters of naked nucleated small cell carcinoma cells.



Input image

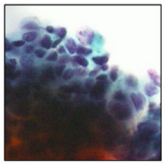
Gold Standard	There are adenocarcinoma cells with hyperchromatic nuclei, nuclei with irregular shape, pale cytoplasm, nuclear enlargement, and prominent nucleoli.
ViT+GPT-oss 20B	There are adenocarcinoma cells with hyperchromatic nuclei, nuclei with irregular shape, pale cytoplasm, nuclear enlargement and prominent nucleoli.
Our previous	There are adenocarcinoma cells with hyperchromatic nuclei, nuclei with irregular shape, pale cytoplasm, and nuclear enlargement.
Qwen2.5 VL 32B	There are squamous cell carcinoma cells with thick cytoplasm, hyperchromatic nuclei, and nuclei with irregular shape.

< Squamous cell carcinoma cells >



Input image

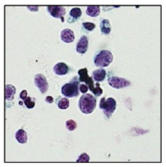
Gold Standard	There are keratinized squamous cell carcinoma cells with hyperchromatic nuclei and nuclei with irregular shape.
ViT+GPT-oss 20B	There are keratinized squamous cell carcinoma cells with hyperchromatic nuclei and nuclei with irregular shape.
Our previous	There are keratinized squamous cell carcinoma cells with hyperchromatic nuclei and nuclei with irregular shape.
Qwen2.5 VL 32B	There are keratinized squamous cell carcinoma cells with hyperchromatic nuclei and nuclei with irregular shape.



Input image

Gold Standard	There are squamous cell carcinoma cells with thick cytoplasm, dark nuclei, and nuclei with irregular shape.
ViT+GPT-oss 20B	The background is relatively clean, and there are clusters of columnar epithelial cells.
Our previous	The background is relatively clean and there are clusters of columnar epithelial cells.
Qwen2.5 VL 32B	There are squamous cell carcinoma cells in cohesive clusters of spindly epithelial cells with hyperchromatic nuclei and nuclei with irregular shape.

< Small cell carcinoma cells >



Input image

Gold Standard	There are scattered individual naked nucleated small cell carcinoma cells.
ViT+GPT-oss 20B	There are scattered individual naked nucleated small cell carcinoma cells.
Our previous	There are scattered naked nucleated small cell carcinoma cells with a salt and pepper pattern of chromatin.
Qwen2.5 VL 32B	There are scattered individual naked nucleated small cell carcinoma cells.

Figure 4. Images and their findings.

Table 5 shows the comparison of key diagnostic feature recall and false mention rate for histology-specific cytological findings.

Figure 5 shows the visualization results of the ViT used as the vision model in the proposed method (ViT + GPT-oss). The results indicate that the model primarily focuses on cellular regions during feature extraction.

Table 5. Comparison of key diagnostic feature recall and false mention rate for histology-specific cytological findings. Values in parentheses (X/Y) indicate the number of cases in which the feature was correctly identified (X) over the number of reference cases containing that feature (Y).

Model	Histological Subtype	Key-Feature 1 Recall		Key-Feature 2 Recall		False Mention Rate
ViT + GPT-oss 20B	Adenocarcinoma	0.857	(30/35)	0.886	(31/35)	0.143
	Squamous cell carcinoma	0.983	(58/59)	1.000	(11/11)	0.441
	Small cell carcinoma	1.000	(38/38)	0.333	(5/15)	0.000
Our previous method	Adenocarcinoma	0.657	(23/35)	0.914	(32/35)	0.343
	Squamous cell carcinoma	0.915	(54/59)	1.000	(12/12)	0.525
	Small cell carcinoma	0.763	(29/38)	0.533	(8/15)	0.237
Qwen2.5 VL 32B	Adenocarcinoma	0.686	(24/35)	0.971	(34/35)	0.343
	Squamous cell carcinoma	0.932	(55/59)	1.000	(11/11)	0.390
	Small cell carcinoma	1.000	(38/38)	0.533	(8/15)	0.000

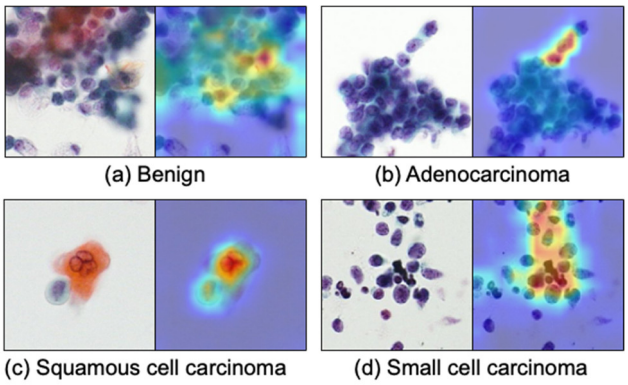


Figure 5. Visualization results of vision model. (ViT + GPT-oss).

4. Discussion

In this study, we developed an image-captioning model for cytological images by combining a vision transformer with open-source LLMs. Unlike general-purpose VLMs, the proposed approach focuses on a specific medical task, namely, image-finding generation for cytological images, and aims to develop a task-specialized model tailored to this domain. While several medical VLMs have been proposed mainly for VQA-style interaction, this study instead explores the applicability of a general-purpose vision–language framework to image-finding generation for cytological images, which are not included in existing medical VLM training domains.

As shown in Figure 4, the descriptions generated by the proposed method exhibited minor differences in wording compared to the reference descriptions; however, the overall semantic consistency was high, and the generated findings were confirmed to be clinically plausible. In addition, the proposed model achieved favorable scores on language similarity metrics, such as BLEU and CIDEr, indicating that it has a certain level of performance as an image-finding generation model for cytological images.

When comparing the performance across different open-source LLMs, the model using GPT-2 showed lower performance than the more recent models, likely owing to its older architecture and the limited amount of data used for pretraining. In contrast, the models employing Qwen2.5, Qwen3, and GPT-OSS demonstrated consistently favorable evaluation metrics, with the GPT-OSS-based model achieving the highest overall performance. Notably, GPT-OSS, which adopts a mixture-of-experts architecture, is the only LLM evaluated in this study that adopts a mixture-of-experts (MoE) architecture. Cytological images exhibit substantial variability in cellular morphology and staining patterns, and contain diverse visual features. Therefore, the MoE architecture, which dynamically acti-

vates different expert submodels depending on the input, may be particularly effective in handling such heterogeneous data.

Analysis of the generated descriptions for clinically relevant category classification revealed that all the evaluated models generally achieved good discrimination performance. In particular, the model combining the vision transformer and GPT-OSS achieved very high sensitivity (0.992) and specificity (1.000) for benign–malignant classification. Furthermore, for the four-category classification, this model achieved an accuracy of 0.858, indicating strong overall performance. Taken together, these results suggest that the proposed model achieves the best performance among the evaluated models. In contrast, the model combining the vision transformer and Qwen3 achieved a sensitivity and specificity of 1.000, indicating that it may be especially useful for screening purposes.

Although high sensitivity and specificity were obtained for benign–malignant discrimination, the accuracy of the histological subtype classification was relatively low. One possible reason is that in lung cancer cytology using Papanicolaou staining, the discrimination between adenocarcinoma and squamous cell carcinoma, both classified as non-small cell carcinoma, is clinically challenging, and even experienced physicians have been reported to achieve an accuracy of only approximately 70% [33]. The proposed method achieved a classification accuracy of 68.8% for these two subtypes based on confusion matrix analysis, suggesting that the model can perform histological subtype estimation at a level comparable to that of specialists.

In our previous work, we proposed an image-finding generation model that switches transformer decoders based on the benign–malignant classification results obtained using a CNN. Compared to this approach, the proposed method using a vision transformer and open-source LLMs achieved higher sensitivity and specificity. This improvement is likely attributable to the difference in discriminative capability between CNN-based architectures and the vision transformer, which was pretrained on approximately two billion images. In addition, the proposed models achieved higher language similarity scores than the previous model, suggesting that the combination of a high-performance vision model and LLM plays an important role in image-finding generation for cytological images.

Furthermore, compared with existing general-purpose VLMs, the proposed method achieved approximately 10% higher language evaluation scores and 20% higher classification accuracy. Although existing models pretrained on large-scale datasets exhibited strong generalization performance, they were outperformed by a task-specific model developed from scratch for cytological image-finding generation. This performance gap may be attributed to the substantial domain gap between natural images and general text used for pre-training, cytological images, and cytological findings. Additionally, cytological reports are characterized by a limited vocabulary of specialized terms and relatively standardized expressions. As a result, general-purpose VLMs designed for freeform text generation may not sufficiently capture the precise expressions required for cytological image finding. These results highlight the importance of task-specific model design in medical applications.

In Table 5, which evaluates morphological features described in the generated findings, the proposed method (ViT + GPT-oss) achieved the highest recall for key-feature 1. This indicates that the proposed method effectively captures the primary diagnostic morphological features, which contributes to its overall performance. In contrast, for key-feature 2, other models showed slightly higher recall for some histological subtypes. Regarding the false mention rate, the proposed method exhibited low values for adenocarcinoma and small cell carcinoma (0.143 and 0.000, respectively), indicating that diagnostically unnecessary or potentially misleading features were rarely included in the generated descriptions. In contrast, for squamous cell carcinoma, the false mention rate was slightly higher than

that of Qwen 2.5 VL. This tendency is consistent with the confusion matrix shown in Table 4, which indicates that the proposed method includes a relatively higher number of misclassifications for squamous cell carcinoma. Overall, the proposed method demonstrates a favorable balance between key-feature recall and false mention rate, suggesting its potential clinical usefulness for cytological image-finding generation.

An analysis of the visualization results, which highlight the regions attended to by the ViT during feature extraction, revealed that attention was concentrated on characteristic cellular regions, as shown in Figure 5. Taken together with the fact that the generated findings successfully captured the primary diagnostic features, these results suggest that the proposed method effectively focuses on relevant cellular regions and accurately extracts meaningful morphological features.

This study primarily aims to support pathologists in writing image-finding descriptions after diagnosis. In diagnostic reports, image-findings are routinely described; in the proposed workflow, candidate descriptions are generated by the proposed method and presented to pathologists, who then review and revise the content to finalize the report in a human-in-the-loop manner. This approach is expected to reduce the workload associated with report writing, mitigate inter-observer variability in descriptive styles, and decrease the risk of omitting diagnostically important findings.

Despite these promising results, this study had several limitations. First, this study represents a preliminary investigation and the dataset was limited to cytological specimens collected from a single institution with a relatively small number of images. In addition, this study focused on LBC specimens with well-preserved cellular morphology and high cell density. However, conventional smear specimens, which are cost-effective, are widely used in many clinical settings, and differences in specimen preparation methods may introduce a domain shift in cytological image appearance. Compared with LBC, conventional smear specimens often exhibit variations in cellular distribution, nuclear contrast, staining artifacts, and background components. As a result, models trained exclusively on LBC specimens may suffer from feature misalignment and reduced morphological fidelity when applied to conventional smear samples, potentially affecting the reliability of generated cytological findings. Future work should, therefore, involve larger datasets incorporating specimens prepared using different methods and data collected from multiple institutions, as well as external validation.

In this study, the effectiveness of the proposed method was evaluated through its overall performance on the image-finding generation task using a VLM, rather than through an analysis of the contributions of individual components. Accordingly, the visual encoder and the language model were updated simultaneously, and ablation experiments designed to independently assess the contribution of each component were considered beyond the scope of this study. Future work will require a more detailed analysis of how design differences in individual components, including the visual encoder, projection layer, and language model, affect the performance of image-finding generation.

Another consideration is that we employed a CLIP-based ViT pretrained on the large-scale LAION-2B dataset as the vision model. By fixing this high-performance visual encoder, we systematically compared multiple large language models to evaluate their impact on cytological image-finding generation. Future work will explore optimized combinations of visual encoders, including those pretrained on medical images, and language models. In addition, further comparison between ViT models pretrained for image classification and those trained as part of CLIP will be required.

Regarding the projection layer, an MLP-based projection layer was used to map visual features into a single-token representation for conditioning the language model. While Transformer-based projection methods such as Q-Former aim to model multiple

visual tokens, this study focused on compact single-token summarization. Although the MLP introduces minimal nonlinearity compared with linear projection, a systematic comparison with alternative projection strategies was not conducted and will be addressed in future work.

5. Conclusions

In this study, we proposed a task-specific VLM for generating image findings from lung cytological images by combining a vision transformer and open-source LLMs. Visual features were captured using a vision transformer pretrained on large-scale image datasets, and image-finding descriptions were generated using recent LLMs such as GPT-OSS and Qwen. The experimental results demonstrated that the proposed method generated higher-quality image findings than our previous approach and existing general-purpose VLMs and that histological subtypes inferred from the generated findings were more accurate. These results indicate that the proposed approach is effective for medical task-oriented applications such as image finding.

Author Contributions: Conceptualization, A.T., T.T. and H.F.; data curation, K.I., Y.K. and N.Y.; methodology, A.T.; software, A.T.; validation, T.T., Y.K., N.Y. and A.T.; investigation, A.T., Y.K., T.T. and H.F.; writing—original draft preparation, A.T.; writing—review and editing, T.T., K.I. and H.F.; visualization, A.T.; project administration, A.T.; funding acquisition A.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by a Grant-in-Aid for Scientific Research (No. 23K07117) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

Institutional Review Board Statement: This study was approved by the Ethical Review Committee of Fujita Health University (HM23-390) and was conducted in accordance with the World Medical Association's Declaration of Helsinki.

Informed Consent Statement: Informed consent was obtained in the form of an opt-out at Fujita Health University Hospital, and all data were anonymized.

Data Availability Statement: Owing to ethical and privacy considerations regarding the patient data, the dataset cannot be made publicly available. However, researchers with a legitimate scientific purpose may request access to both the code and data, subject to appropriate ethical review and approval from the relevant institutional review board and the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bray, F.; Laversanne, M.; Sung, H.; Ferlay, J.; Siegel, R.L.; Soerjomataram, I.; Jemal, A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2024**, *74*, 229–263. [CrossRef] [PubMed]
2. Frankel, D.; Nanni, I.; Ouafik, L.; Greillier, L.; Dutau, H.; Astoul, P.; Daniel, L.; Kaspi, E.; Roll, P. Cytological samples: An asset for the diagnosis and therapeutic management of patients with lung cancer. *Cells* **2023**, *12*, 754. [CrossRef] [PubMed]
3. Al-Abbadi, M.A. Basics of cytology. *Avicenna J. Med.* **2011**, *1*, 18–28. [CrossRef] [PubMed]
4. Walsh, E.; Orsi, N.M. The current troubled state of the global pathology workforce: A concise review. *Diagn. Pathol.* **2024**, *19*, 163. [CrossRef] [PubMed]
5. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; Volume 25, pp. 1097–1105.
6. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
7. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Summers, R.M. TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2018; pp. 9049–9058. [CrossRef]

8. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 5998–6008.
9. Hou, B.; Kaissis, G.; Summers, R.M.; Kainz, B. RATCHET: Medical transformer for chest X-ray diagnosis and reporting. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021*; De Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C., Eds.; Springer International Publishing: Cham, Switzerland, 2021; Volume 12907, pp. 293–303. [CrossRef]
10. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. In Proceedings of the Advances in Neural Information Processing Systems, Online, 6–12 December 2020; Volume 33.
11. Nakaura, T.; Yoshida, N.; Kobayashi, N.; Shiraishi, K.; Nagayama, Y.; Uetani, H.; Kidoh, M.; Hokamura, M.; Funama, Y.; Hirai, T. Preliminary assessment of automated radiology report generation with generative pre-trained transformers: Comparing results to radiologist-generated reports. *Jpn. J. Radiol.* **2024**, *42*, 190–200. [CrossRef] [PubMed]
12. Delbrouck, J.-B.; Xu, J.; Moll, J.; Thomas, A.; Chen, Z.; Ostmeier, S.; Azhar, A.; Li, K.Z.; Johnston, A.; Bluethgen, C.; et al. Automated structured radiology report generation. In *Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Vienna, Austria, 2025; pp. 26813–26829.
13. Zhou, Y.-F.; Yao, K.-L.; Li, W.-J. GNNFormer: A Graph-Based Framework for Cytopathology Report Generation. *arXiv* **2023**, arXiv:2303.09956. [CrossRef]
14. Tran, M.; Schmidle, P.; Guo, R.R.; Wagner, S.J.; Koch, V.; Lupperger, V.; Novotny, B.; Murphree, D.H.; Hardway, H.D.; D'Amato, M.; et al. Generating dermatopathology reports from gigapixel whole slide images with HistoGPT. *Nat. Commun.* **2025**, *16*, 4886. [CrossRef] [PubMed]
15. Teramoto, A.; Michiba, A.; Kiriyama, Y.; Tsukamoto, T.; Imaizumi, K.; Fujita, H. Automated description generation of cytologic findings for lung cytological images using a pretrained vision model and dual text decoders: Preliminary study [Preliminary study]. *Cytopathology* **2025**, *36*, 240–249. [CrossRef] [PubMed]
16. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
17. OpenAI; Agarwal, S.; Ahmad, L.; Ai, J.; Altman, S.; Applebaum, A.; Arbus, E.; Arora, R.K.; Bai, Y.; Baker, B.; et al. Gpt-Oss-120b & Gpt-Oss-20b Model Card. *arXiv* **2025**, arXiv:2508.10925.
18. Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. Qwen3 Technical Report. *arXiv* **2025**, arXiv:2505.09388. [CrossRef]
19. Schmitt, F.C.; Bubendorf, L.; Canberk, S.; Chandra, A.; Cree, I.A.; Engels, M.; Hiroshima, K.; Jain, D.; Kholová, I.; Layfield, L.; et al. The World Health Organization reporting system for lung cytopathology. *Acta Cytol.* **2023**, *67*, 80–91. [CrossRef] [PubMed]
20. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the International Conference on Machine Learning, Online, 18–24 July 2021.
21. Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; Jitsev, J. Reproducible Scaling Laws for Contrastive Language–Image Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 2818–2829.
22. Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Wightman, R.; Coombes, L.; Jitsev, J.; Komatsuzaki, A. LAION-5B: An open large-scale dataset for training next generation image-text models. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; Volume 35.
23. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models Are Unsupervised Multitask Learners. *OpenAI Blog* **2019**, *1*, 9.
24. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *ICLR* **2021**, *1*, 3.
25. Vedantam, R.; Zitnick, C.L.; Parikh, D. CIDEr: Consensus-based image description evaluation. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Boston, MA, USA, 2015; pp. 4566–4575. [CrossRef]
26. 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 on Association for Computational Linguistics—ACL '02*; Association for Computational Linguistics: Philadelphia, PA, USA, 2001; p. 311. [CrossRef]
27. Denkowski, M.; Lavie, A. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*; Association for Computational Linguistics: Baltimore, MD, USA, 2014; pp. 376–380. [CrossRef]
28. Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Text Summarization Branches Out*; Association for Computational Linguistics: Barcelona, Spain, 2004; pp. 74–81.

29. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. SPICE: Semantic propositional image caption evaluation. In *Computer Vision—ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; Volume 9909, pp. 382–398. [CrossRef]
30. Chefer, H.; Gur, S.; Wolf, L. Transformer Interpretability Beyond Attention Visualization. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Online, 19–25 June, 2021; pp. 782–791.
31. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023.
32. Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. Qwen2.5-VL Technical Report. *arXiv* **2025**, arXiv:2502.13923. [CrossRef]
33. Sigel, C.S.; Moreira, A.L.; Travis, W.D.; Zakowski, M.F.; Thornton, R.H.; Riely, G.J.; Rekhtman, N. Subtyping of non-small cell lung carcinoma: A comparison of small biopsy and cytology specimens. *J. Thorac. Oncol.* **2011**, *6*, 1849–1856. [CrossRef] [PubMed]