Enhancing Multimodal Large Language Models with Multi-instance Visual Prompt Generator for Visual Representation Enrichment

Wenliang Zhong
The University of Texas at Arlington
wxz9204@mavs.uta.edu

Wenyi Wu
Amazon
wenyiwu@amazon.com

Qi Li
Amazon
qlimz@amazon.com

Rob Barton
Amazon
rab@amazon.com

Boxin Du
Amazon
boxin@amazon.com

Shioulin Sam
Amazon
shioulin@amazon.com

Karim Bouyarmane
Amazon
bouykari@amazon.com

Ismail Tutar
Amazon
ismailt@amazon.com

Junzhou Huang
The University of Texas at Arlington
jzhuang@uta.edu

Abstract

Multimodal Large Language Models (MllMs) have achieved SOTA performance in various visual language tasks by fusing the visual representations with LLMs leveraging some visual adapters. In this paper, we first establish that adapters using query-based Transformers such as Q-former is a simplified Multi-instance Learning method without considering instance heterogeneity/correlation. We then propose a general component termed Multi-instance Visual Prompt Generator (MIVPG) to incorporate enriched visual representations into1 LLMs by taking advantage of instance correlation between images or patches for the same sample. Quantatitive evaluation on three public vision-language (VL) datasets from different scenarios shows that the proposed MIVPG improves Q-former in main VL tasks.

1. Introduction

In recent years, with the disruptive changes brought to the Machine Learning community by Large Language Models (LLMs)[4, 29–31], an increasing number of researchers have been exploring the application of LLMs in the realm of multimodality, giving rise to Multimodal Large Language Models (MLLMs)[2, 21, 22, 24, 48]. One of the most common forms of multimodality involves the combination of images and text. Just as humans excel in using both images and text to perform tasks, the fusion of images and text in multimodal applications finds wide real-world use, such as in Image Captioning[13, 32, 43, 44] and Visual Question Answering (VQA)[3, 11, 25, 38]. Leveraging the formidable generalization capabilities of large models, MLLMs have achieved state-of-the-art (SOTA) performance in various few-shot and fine-tuning tasks.

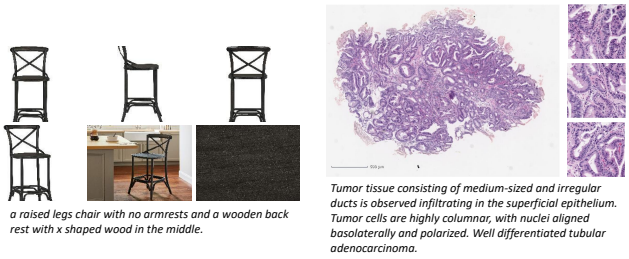


Figure 1. **Left**: Exemplary images from [7], portraying e-commerce products captured from various aspects. **Right**: Illustration of a Whole Slide Image (WSI) sourced from [36]. Each WSI is composed of multiple patches, exhibiting dimensions comparable to those of natural images.

Although MLLMs have achieved remarkable results in various multimodal tasks, the majority of existing open-source MLLMs are primarily pretrained on (image, text) pairs. However, in real-life scenarios, samples are usually represented by enriched visual representations. For example, e-commerce stores typically display products accompanied by several images and a textual description of their features [7]. These images may represent different angles of the product's appearance or various aspects of its overall and detailed characteristics. In medical image analysis, a Whole Slide Image (WSI)[36] is challenging to fit entirely into a network due to its gigapixel size (more than 10^8 pixels). Existing medical image analysis typically segments them into multiple patches as images[1, 16, 20, 46, 49],

yet these multiple patches still represent the same sample. Therefore, applying MLLMs to multimodal tasks with richer visual inputs holds much practical significance.

In contemporary MLLMs, the integration of images is achieved through a critical component for imparting visual understanding to LLMs through transforming images to visual tokens, which we termed **Visual Prompt Generators** (VPGs) in this paper. SOTA MLLMs, such as BLIP2[22], Flamingo[2], and MiniGPT-4[48], utilize attention-based VPGs with learnable query embeddings. These embeddings engage in cross-attention with visual embeddings, extracting visual information for LLM input. In this work, we introduce a novel approach, the Multi-instance Visual Prompt Generator (MIVPG), designed to handle diverse visual inputs. Drawing inspiration from Multiple Instance Learning (MIL), MIVPG treats images or patches of a sample as a set of instances, forming a "bag." Unlike traditional machine learning tasks, MIL performs predictions at the bag level rather than the instance level, employing permutation-invariant functions to aggregate instances. MIVPG extends this concept by considering correlations and relationships across visual representations, facilitating signal pooling from different dimensions. Additionally, we establish that the commonly used QFormer[22, 48] is a limited MIL module, prompting the introduction of MIVPG. We showcase MIVPG's enhanced performance across three distinct scenarios, including common natural images, gigapixel-sized pathological images, and e-commerce products with multiple images.

In summary, our contributions in this paper can be outlined as follows:

- We introduce a general and flexible component MIVPG to incorporate enriched visual representations and their relationship into the open source LLM.
- We establish that the commonly used QFormer is a simplified case of MIVPG with limited capability and conduct experiments to show the superiority of our component over the QFormer.
- We evaluate the MIVPG on three public datasets from distinct scenarios and showcase that the MIVPG supports visual representation aggregation from different dimensions: image dimension for e-commerce data and patch dimension for WSI. MIVPG outperforms the QFormer by a significant margin in all datasets, which demonstrates the effectiveness and generalizability of the proposed component.

2. Related Work

2.1. Multimodal Learning

Recently, various vision-language models (VLMs) have been proposed to enhance the fusion of text and images. For example, TCL [42] employed triplet contrastive learning to simultaneously learn from text and images. Many state-of-the-art MLLMs have also emerged, with one major distinction lying in the design of VPGs. For instance, FROMAGe [18] and LLaVA [24] employ a straightforward linear projection as their VPGs. On the other hand, Flamingo [2] introduces the novel use of the Perceiver Resampler, incorporating cross attention and learnable query embeddings. BLIP2 [22] innovatively employs the QFormer to improve image-text alignment. Meanwhile, MiniGPT-4 [48] integrates a frozen QFormer with additional learnable layers for enhanced performance.

While successful in diverse tasks, current multimodal models are primarily designed under the assumption of a one-to-one relationship between texts and image inputs. In reality, the relationship between text and images can be one-to-many or many-to-many. Effectively applying multimodal models in such scenarios poses an open challenge.

2.2. Multiple Instance Learning

Traditionally, Multiple Instance Learning [6, 28] can be broadly categorized into two main types: (1) **The instance-level approach** [5, 10, 14, 17] : In this approach, bag-level predictions are directly derived from the set of instance-level predictions. (2) **The embedding-level approach** [16, 20, 26, 34] : Here, bag-level predictions are generated from an bag-level embedding that represents multiple instances. For the former, hand-crafted pooling operators such as mean pooling or max pooling are often employed. However, in practical applications, these hand-crafted pooling operators often yield limited results. Hence, the majority of current research is grounded in the latter approach.

Aggregating instance features to form bag-level features typically leads to better outcomes but requires more complex pooling operations. Recent research has applied neural networks to the pooling process in MIL. For instance, MI-Net [40] utilizes a fully connected layer in MIL. Furthermore, AB-MIL [16] employs attention during the pooling process, allowing for better weighting of different instances. Another category of methods[34] attempts to consider the relationships between different instances using the self-attention mechanism. Moreover, DS-MIL [20] employs attention not only to consider instance-to-instance relationships but also instance-to-bag relationships; DTFD-MIL [46] incorporates the Grad-CAM[33] mechanism into MIL. While these approaches concentrate on single modality, the extension of MIL to multimodal applications is scarcely explored [39].

3. Methodology

3.1. Preliminaries and Notations

Existing MLLMs utilize VPGs, like Qformer and Perceiver Resampler, to encode images to visual tokens. To elabo-

rate, images are initially processed by a visual model, such as ViT[9], resulting in image embeddings denoted as $I \in \mathbb{R}^{P \times D_I}$, P represents the overall number of visual tokens, encompassing patches and the special token $[CLS]$, and D_I is the dimension of token embeddings. Subsequently, an adapter is employed to project these image embeddings into the LLM embedding space (Figure 2c), which we refer to as visual prompt embeddings. The adapter can be as simple as a linear projection. However, a popular choice for contemporary MLLMs is the cross-attention module, where learnable query embeddings interact with image embeddings. We denote query embeddings to have R tokens and D_q dimension $q \in \mathbb{R}^{R \times D_q}$. This mechanism compels the query embeddings to extract essential information from the image embeddings. Without loss of generality, we represent the attention[37] format as Equation 1, and we will use D to denote the dimension of hidden embeddings after projection throughout the rest of this paper.

$$Q_{\text{result}} = \text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{D}})V \quad \text{where } Q \in \mathbb{R}^{R_1 \times D}, K, V \in \mathbb{R}^{R_2 \times D} \tag{1}$$

$A = \text{softmax}(\frac{QK^T}{\sqrt{D}}) \in \mathbb{R}^{R_1 \times R_2}$ is termed as attention map indicating which entries are pivotal. The cross-attention between query embeddings and image embeddings can be represented as Attention$(Q = q, K = I, V = I)$. The resulting query embeddings are then forwarded to FeedForward and Residual layers. Specifically, Perceiver Resampler[2] comprises a single cross-attention layer, while the QFormer is a BERT[8] model in which query embeddings interact with image embeddings within each block. We denote $q^{(l)}$ as the query embeddings in the l^{th} block. Details of the QFormer architecture can be found in Appendix A. Furthermore, when dealing with samples containing multiple images, we use N to denote the number of images $I \in \mathbb{R}^{N \times P \times D_I}$.

MIL typically treats each sample as a bag containing multiple instances. Therefore, we describe MIL in a general form, where a bag is denoted as $B = \{x_1, x_2, \ldots, x_M\}$, with M representing the number of instances in the bag and x_i representing the latent embedding of an instance. In the subsequent discussions in this paper, we will explore MIL from different dimensions thus x and B may have different semantically meanings depending on circumstances. For instance, at the sample level, $M = N$, where instances represent images. At the image level, $M = P$, with instances representing patches.

3.2. Relations between Attention-based VPG and MIL

Embedding-level MIL, as a subset of set models, fundamentally relies on the essential property of permutation invariance[16] (Equation 2) within the aggregation function $g : \mathbb{R}^{M \times D} \to \mathbb{R}^D$.

$$g(x_1, x_2, \ldots, x_M) = g(x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(M)}) \tag{2}$$

$\{\pi(1), \cdots, \pi(M)\}$ represents a permuted order of the original set.

One popular choice of aggregation is the weighted pooling as shown in Figure 2b. In a bag $B = \{x_1, x_2, \ldots, x_M\}$, instance embeddings are transformed into their corresponding weights $\{\alpha_i\}_{i=1}^M$ through a nonlinear projection. All instance embeddings are subsequently pooled using the normalized weights to obtain the bag-level representation (Equation 3).

$$x_B = g(x_1, x_2, \ldots, x_M) = \sum_{i=1}^{M} \alpha_i x_i \tag{3}$$

In AB-MIL[16], weights are calculated as Equation 4.

$$\alpha_i = \frac{e^{w^T \tanh(ux_i^T)}}{\sum_{j=1}^{M} e^{w^T \tanh(ux_j^T)}}, w \in \mathbb{R}^{L \times 1}, u \in \mathbb{R}^{L \times D} \tag{4}$$

DeepSets[45] has demonstrated that when a model comprises a series of permutation-equivalence layers, it is possible to retain permutation invariance in the output. Subsequently, Set Transformer[19]'s findings confirm that Transformers equipped with self-attention inherently adhere to these principles. A permutation-equivalence layer can be formulated as Equation 5.

$$f_i(x, \{x_1, \cdots, x_M\}) = \sigma_i(\lambda x + \gamma pool(\{x_1, \cdots, x_M\})) \tag{5}$$

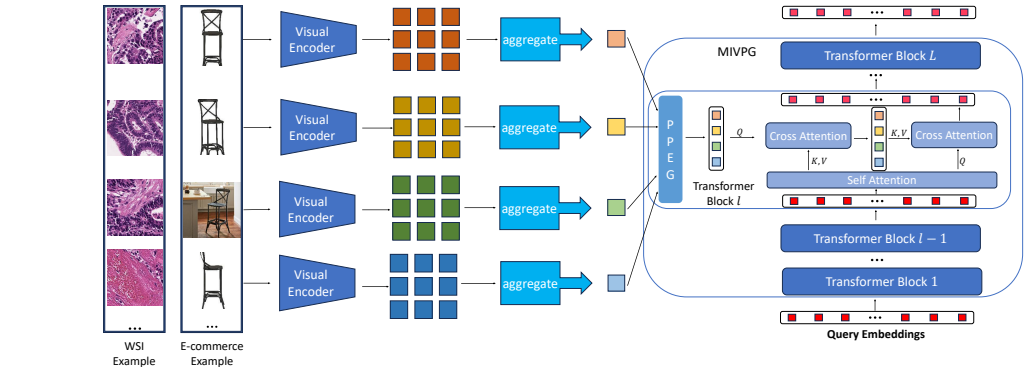
where *pool* is an aggregation function, λ, γ are learnable scalars, and σ_i is the activation function.

Proposition 1. *QFormer belongs to the category of Multiple Instance Learning modules.*

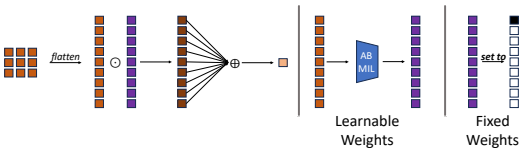
Within the cross-attention layer of QFormer, every query token computes weights for image embeddings. Query embeddings, being learnable parameters, can be seen as a linear transformation from an instance to its weight. To provide further clarification, each row in the attention map A signifies the weights assigned to instances for aggregation. Consequently, the cross-attention between the learnable query embeddings and the input is permutation invariance.

The result of cross-attention is combined with the original query embeddings using a residual connection. This process can be expressed as shown in Equation 5, by replacing *pool* with Equation 1, and setting $\lambda = \gamma = \mathbb{I}$, as illustrated in Equation 6, which is permutation equivalence.

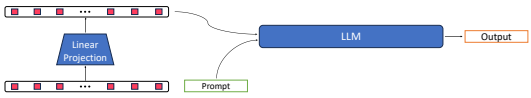
$$f_i(q, I) = q + \text{Attention}(Q = q, K = I, V = I) \tag{6}$$



(a) Overview of MIVPG



(b) Image-level Bag Aggregation



(c) Visual and Text Prompts as Inputs of LLMs

Figure 2. Overview of MIVPG. 2a: When handling multiple visual inputs, the initial step involves aggregating them at the image-level. QFormer can be treated as a Multiple Instance Learning module that takes multiple samples as instances. The MIVPG complements QFormer by introducing a correlated self-attention module and the pyramid positional encoding module, depending on specific scenarios. 2b: Image-level aggregation can employ various MIL strategies, either learnable, such as AB-MIL, or fixed, for example, always selecting a specific token. 2c: The visual prompt embeddings produced by Q-Former are combined with textual prompt embeddings and forwarded to the LLM for generating outputs.

Considering that the self-attention layer within the QFormer block adheres to the principles of permutation equivalence, we can conceptualize the QFormer as a multi-head MIL mechanism.

From the standpoint of MIL, the weighted pooling in Equation 1 operates under the assumption that instances are independent and identically distributed (i.i.d)[34]. However, in practical scenarios, instances may exhibit correlations, and accounting for instance correlation can lead to improved performance. It's worth noting that when each sample contains only one image, the input to QFormer comprises patch embeddings that have already incorporated correlations through the self-attention layer in ViT. Moreover, performance enhancement is attainable through the integration of a Pyramid Positional Encoding Generator (PPEG)[34], which complements the proposed MIVPG when handling single-image inputs.

3.3. MIVPG for Multiple Visual Inputs

While previous approaches have touched upon the use of multiple images as inputs to MLLMs, they still exhibit certain limitations. For instance, while handling the scenario of using videos as inputs. Perceiver Resampler[2] simply concatenates patches from multiple images into a sequence to serve as the input, i.e., $I \in \mathbb{R}^{N\times P\times D_I} \rightarrow \mathbb{R}^{(N\cdot P)\times D_I}$. However, it is essential to treat each image as a distinct bag, with each patch considered within the context of its respective bag. Directly flattening the patches may lead to a misallocation of weights across instances from different bags.

When a sample comprises multiple images, it is imperative to consider MIL feature aggregation from different perspectives. In the context of individual images, each image can be treated as a 'bag,' and each patch within the image as an 'instance.' From the sample's perspective, each sample can also be regarded as a 'bag,' with each image within the sample as an 'instance.' When a sample contains only a single image, we can focus primarily on the former perspective since the latter perspective involves a single instance per bag. However, in a more general context, it is essential to adopt a hierarchical approach when considering the utilization of MIL for feature aggregation. Without loss of generality, we now consider the input of the MIVPG to be a bag B containing multiple instances. Hence, the cross-attention can be expressed as Attention$(Q = q, K = B, V = B)$.

3.4. Unveiling Instance Correlation in MIVPG for Enhanced Multi-instance Scenarios

When there are multiple images as inputs and each image is regarded as an instance, their correlation should also be considered. TransMIL[34] has demonstrated that, when multiple instances are as the input, applying self-attention mechanisms can effectively learn the correlations between them. However, in certain applications, such as Whole Slide Image analysis, each bag may contain a large number of in-

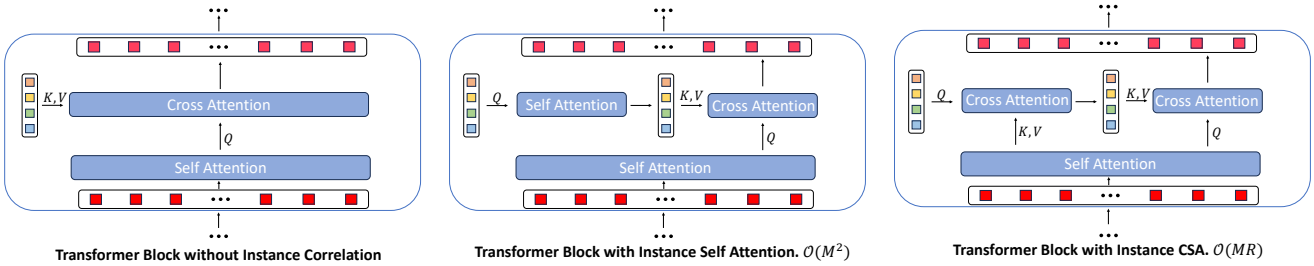


Figure 3. **Left**: The original transformer block without considering instance correlation. **Middle**: Instance correlation is computed through a self-attention layer among input instances, incurring a time complexity of $\mathcal{O}(M^2)$. **Right**: Instance correlation is calculated using query embeddings from the previous layer. This approach reduces the time complexity in computing correlation to $\mathcal{O}(MR)$.

stances ($M > 1000$). Directly computing self-attention between instances Attention($Q = B, K = B, V = B$) results in a time complexity of $\mathcal{O}(M^2)$, making such calculations computationally intensive. To address this issue, it[34] employs the Nystrom approximation of attention[41]. In this paper, following QFormer, we propose a method for computing the correlations between instances using a low-rank projection. This approach only requires the incorporation of a Correlated Self Attention (CSA) module in each Transformer block to achieve the desired results.

Considering the input to a Self Attention module is a bag of instance embeddings $B = [x_1, x_2, \ldots, x_M]$ with shape $\mathbb{R}^{M \times D}$, instead of directly computing self-attention, one can adopt a more efficient approach[19] by projecting the original matrices into a lower-rank space using a learnable matrix $L \in \mathbb{R}^{M' \times D}$, where $M \gg M'$, as illustrated in Equation 7.

$$L' = \text{Attention}(Q = L, K = B, V = B) \tag{7}$$

$$B_{result} = \text{Attention}(Q = B, K = L', V = L') \tag{8}$$

Subsequently, the aggregated low-rank matrix can be reintegrated with the original embeddings, as shown in Equation 8. This low-rank projection effectively reduces the time complexity to $\mathcal{O}(MM')$.

Recall that QFormer is a stack of Transformer blocks, with each block consisting of a self-attention layer followed by a cross-attention layer. In the l^{th} block, the input consists of query embeddings from the previous block $q^{(l-1)}$ have already aggregated instance information in the prior block. These query embeddings can be directly utilized as L' in the cross-attention layer (as shown in Equation 9) since the self-attention layer retains their permutation equivalence. Therefore, we can efficiently harness the query embeddings from the previous block to learn instance correlations.

$$(B_{result})^{(l)} = \text{Attention}(Q = B, K = q^{(l-1)}, V = q^{(l-1)}) \tag{9}$$

$(B_{result})^{(l)}$ is the updated instance embeddings that have aggregated instance correlation at layer l.

Proposition 2. *MIVPG, when equipped with the CSA (Correlated Self-Attention) module, continues to fulfill the essential properties of MIL*

We prove the proposition 2 in the supplementary B.

In summary, as depicted in Figure 2a, we establish that QFormer falls under the MIL category and is a specialized instance of our proposed MIVPG. The latter extends to visual inputs with multiple dimensions, accounting for instance correlation.

4. Experiments

To assess the effectiveness of our proposed approach, we conduct evaluations across various scenarios:

1. where each sample comprises a single image, and patches are naturally considered as instances;
2. where each sample includes multiple instances, but we use a general embedding for each image;
3. where each sample contains multiple images, with each image containing multiple patches.

4.1. General Setup

We initialize our model using BLIP2 [22] with FLAN-T5-XL. MIVPG is initialized with weights from QFormer. The model consists of a frozen language model and a frozen visual model. During training, we only update the MIVPG. The visual encoder, ViT-G, is employed to encode images into patches of embeddings, and the images are resized to dimensions of 224×224. In our experiments, we observed that unfreezing the visual encoder does not lead to additional improvements in datasets with small sizes. Further details can be found in the supplementary C.1.

4.2. Scenario 1: Samples with Single Image

We start by assessing the performance of our method on common single-image datasets to validate the effectiveness of considering Multiple Instance Learning through the addition of Pyramid Positional Encoding Generator for each

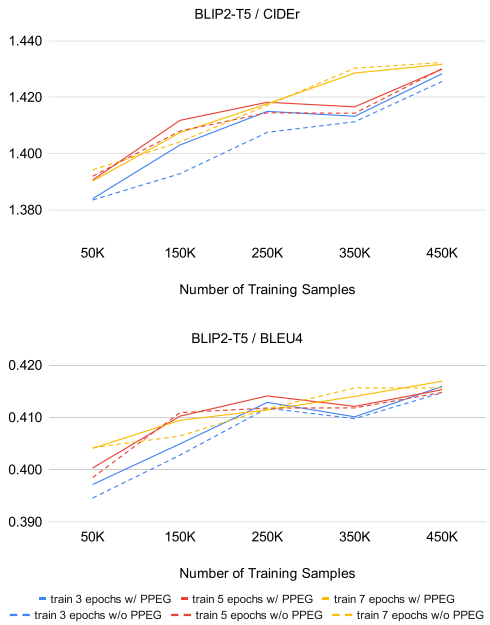


Figure 4. Experiment Results on **MSCOCO**. We adopt the metrics used in [22]. It is evident that the incorporation of MIL modules enhances the QFormer in the majority of cases.

layer containing MIVPG. Following the fine-tuning baseline in BLIP2, we choose **MSCOCO**[23] as the evaluation dataset and employ the Karpathy validation and testing set split. The original training set contains approximately 560K image-text pairs. Given that most existing MIL methods are tailored for small datasets, we evaluate performance across various sizes of training subsets obtained through random sampling. In this dataset, we treat patches as individual instances, and each sample comprises only one image, indicating that $N = 1$.

The result from the **MSCOCO** dataset is shown in Figure 4. It reveals that the enhancements achieved through the use of PPEG are more noticeable when working with smaller datasets. As the dataset size increases, the difference in performance becomes less significant. This can be attributed to the fact that in cases of limited data, models often struggle to discern latent and implicit patterns. Therefore, more sophisticated modules are required to uncover deeper relationships within the data. Conversely, existing MLLMs are typically pretrained on extensive datasets, which tend to mitigate the impact of data scarcity. In practical applications, we demonstrate that one can draw upon MIL techniques to enhance MLLMs performance in scenarios where there is insufficient data for the downstream task.

Table 1. Experiments on the PatchGastricADC22 dataset [36], we evaluate our proposed method against baselines from [36], considering four widely-adopted metrics [1]. Augmented baselines, denoted as **aug**, which signifies a model trained with data augmentation.

	BLEU@4	CIDEr	METEOR	ROUGE
DenseNet121 x20 p3x3	0.336±0.023	2.03±0.245	0.284±0.009	0.481±0.016
EfficientNetB3 x20 p3x3	0.364±0.019	2.154±0.200	0.302±0.014	0.510±0.026
DenseNet121 x20 p3x3 aug	0.347±0.017	2.024±0.198	0.292±0.007	0.485±0.012
EfficientNetB3 x20 p3x3 aug	0.414±0.024	2.820±0.326	0.327±0.012	0.540±0.021
BLIP2-MIVPG w/o CSA	0.441±0.009	2.902±0.233	0.359±0.004	0.583±0.011
BLIP2-MIVPG (Ours)	**0.447**±0.012	**2.930**±0.173	**0.363**±0.005	**0.590**±0.004

4.3. Scenario 2: Samples with Multiple Images, with Each Image as a General Embedding

Next, we evaluate our method in scenarios involving multiple images, where each image contributes only one embedding as its representation. Specifically, we utilize the **PatchGastricADC22**[36] dataset, which is a Whole Slide Image (WSI) dataset. This dataset includes 991 WSIs of $H\&E$-stained gastric adenocarcinoma specimens, accompanied by diagnostic captions extracted directly from existing medical reports. The dataset encompasses a total of 262,777 medical patches, with each WSI containing up to 1860 patches. Each medical patch has a size of 300×300, which will be encoded by the visual encoder after resizing. The dataset is partitioned into training, validation, and test subsets using the methodology outlined in [36], with a split ratio of 0.7, 0.1, and 0.2, respectively. We compare the proposed method against baselines in [36], which are a combination of a visual model (**DenseNet121**[15] or **EfficientNetB3**[35]) and an LSTM[12] as the language model. To ensure a fair comparison, we conduct three experiments with different random seeds and follow the same data augmentation in [36]. In a medical patch, the focus is typically on global information rather than local details. Additionally, given that a WSI can comprise a large number of patches, we aim to reduce computational overhead. Therefore, we choose to use only the $[CLS]$ token output by ViT as the representation for the entire medical patch. In this case, $P = 1$.

As demonstrated in Table 1, our method outperforms the baselines significantly. This result highlights the effectiveness of employing large-scale models in downstream tasks. Moreover, the experiments indicate that the model performs even better when considering correlations among instances, underscoring the effectiveness of our CSA module. Furthermore, we are interested in observing how captions generated by the LLM evolve as the number of training epochs increases. Given the substantial domain gap between medical images and natural images, we believe that existing MLLMs have rarely been trained on medical im-

[1]For consistency, we opted for metrics implemented in https://github.com/salaniz/pycocoevalcap.

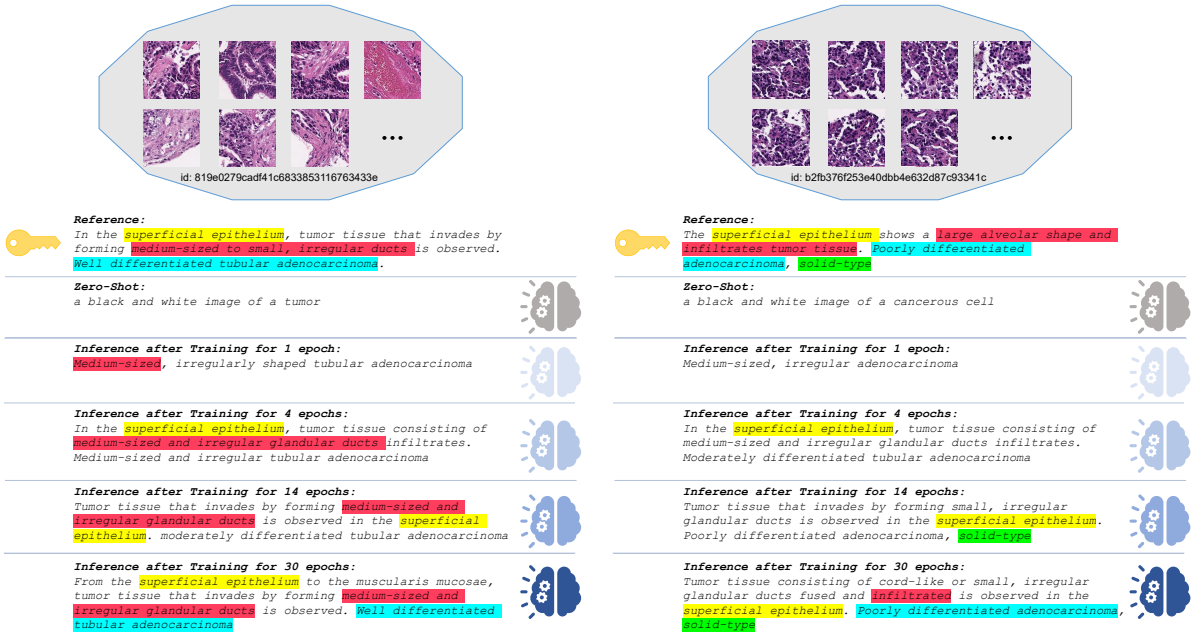


Figure 5. Visualization of Inference Results on PatchGastricADC22. We highlight details that should be focused on the reference. Zero-shot inference is performed using the pretrained BLIP2[22]. As the number of epochs increases, the model acquires more domain knowledge.

ages, rendering them less domain-specific in medical analysis. As depicted in Figure 5, under the zero-shot setting, BLIP2 struggles to generate detailed captions for the provided WSIs. However, with an increasing number of training epochs, the model acquires domain-specific knowledge and produces more relevant captions. Similar to the process of human learning, a discernible trend is observed, where the model initially generates very general captions and gradually incorporates more and more details as the number of epochs increases.

4.4. Scenario 3: Samples with Multiple Images, with Each Image Having Multiple Patches to be Considered

Thirdly, we assess the method in a comprehensive manner where each sample contains multiple images and considers multiple patches within an image. Specifically, we utilize the **Amazon Berkeley Objects (ABO)**[7] dataset, consisting of samples of e-commerce products. Each product is accompanied by multiple images illustrating various characteristics from different perspectives. Typically, one of these images provides an overview of the product, commonly displayed first on an e-commerce website. We refrain from utilizing the product title as the caption due to its limited descriptiveness. Instead, we rely on manually annotated captions [27] from the same dataset as reference captions to assess the quality of our generated captions. Specifically, there are 6410 product samples, and we randomly partition them into train, validation and test subsets with a ratio of 70%, 10%, 20%. Each product comprises images ranging from 2 to 21. In this dataset, we must consider both image details and multiple images simultaneously. Consequently, we apply MIL in both image and patch dimensions. To be specific, we employ AB-MIL (Equation 4) to generate image-level embeddings for images. Each image is then treated as an instance and passed to the QFormer as a sample-level MIL. Since the number of patches per sample is much smaller than that of WSIs, we do not apply PPEG in this setting.

We primarily compare the proposed method against BLIP2 with different settings: 1) **Zero-Shot**: Directly feed the overview image of a sample to query the BLIP2 without fine-tuning; 2) **Single-Image**: Fine-tune the original BLIP2 with the overview image of each product; 3) **Patch-Concatenation**: Fine-tune the original BLIP2 with multiple images, with patches concatenated in one sequence. We repeat the experiments three times with different random seeds and report the mean and standard deviation. As shown in Table 2, results from the ABO dataset demonstrate that

Table 2. Experiment results on the ABO dataset[7]. We compare BLIP2 with different settings as our baselines.

	BLEU@4	CIDEr	METEOR	ROUGE
BLIP2 Zero-Shot	0.024±0.002	0.144±0.005	0.087±0.002	0.264±0.003
BLIP2 Single Image	0.413±0.010	1.515±0.031	**0.304**±0.006	0.584±0.006
BLIP2 Patch-Concatenation	0.412±0.011	1.516±0.032	0.301±0.006	**0.589**±0.007
BLIP2-MIVPG w/o CSA	0.412±0.010	1.528±0.033	0.301±0.005	0.585±0.008
BLIP2-MIVPG	**0.415**±0.011	**1.549**±0.030	**0.304**±0.005	0.586±0.008

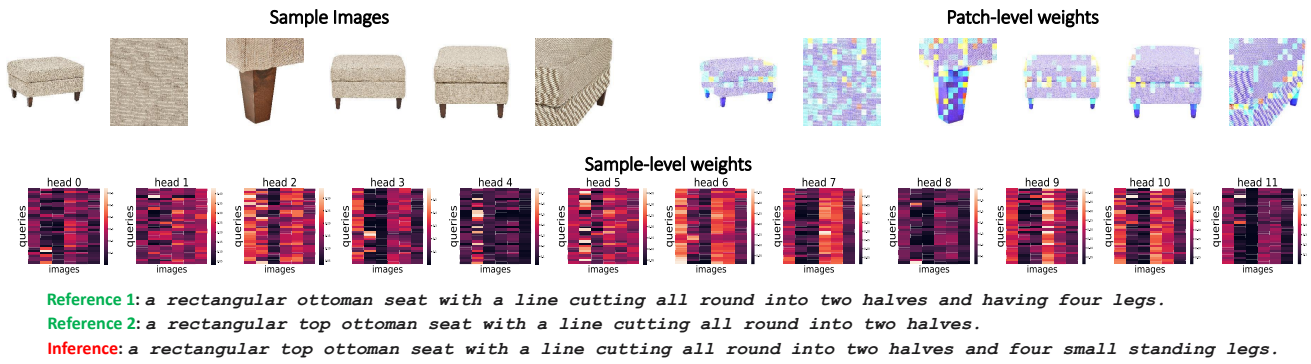


Figure 6. Example from ABO. **Top Left**: a sample consisting of six images. **Top Right**: Attention weights of patches among different images. We use the *COLORMAP JET* to represent the weights, where lighter colors indicate higher weights. **Bottom**: Attention weights among different images. There are 12-head attention maps. In an attention map, each row indicates the weights of images for one query.

our method outperforms the use of a single image or images with concatenated patches mostly, underscoring the efficacy of considering MIL from different dimensions. It is worth noting that fine-tuning BLIP2 with a single image has already achieved respectable performance, indicating that the overview image contains general information about a sample. Additionally, while fine-tuning BLIP2 with multiple concatenated image patches shows good results in terms of *ROUGE*, it should be emphasized that the concatenation results in a complexity of $\mathcal{O}(RNP)$. In contrast, the proposed method applied on different dimensions will only have a complexity of $\mathcal{O}(P + RN)$, ensuring computational efficiency.

Unlike the groundtruth captions that describe the details of a product, we find the zero-shot BLIP2 tends to provide less detailed information. This discrepancy can be attributed to the model's pretraining, where it is predominantly tasked with describing an overview of an image, with less emphasis on details. Nonetheless, when we input multiple images for a single item, the model showcases its capacity to discern what aspects should be emphasized. This capability arises from the presence of common patterns across different images that collectively describe the item, thus affirming the effectiveness of utilizing multiple visual inputs.

A visualization example can be seen in Figure 6, featuring an ottoman seat composed of six images. We present both patch-level attention weights and image-level attention weights. In the patch-level attention weights, the model emphasizes edges and legs of the seat, leading to an output that recognizes the rectangular shape and four legs. The image-level attention weights show maps for all twelve heads. Each row in a map represents a query, and each column represents an image. Notably, different heads and queries exhibit varying attention patterns towards images. Generally, the first, fourth, and fifth images attract the most attention.

4.5. Case Study

	ABO		PatchGastricADC22	
	MIVPG w/ SA	MIVPG w/ CSA	MIVPG w/ SA	MIVPG w/ CSA
BLEU@4	0.409±0.012	**0.415**±0.011	0.444 ±0.020	**0.447** ±0.012
CIDEr	1.532±0.024	**1.549**±0.030	**2.961**±0.242	2.930±0.173
METEOR	0.299±0.006	**0.303**±0.006	0.362±0.008	**0.363**±0.005
ROUGE	0.586±0.008	**0.586**±0.008	0.586±0.012	**0.590**±0.004

Table 3. Ablation results of effectiveness of CSA

To assess the impact of instance correlation, we conduct additional ablation studies involving self-attention (SA) and correlated self-attention (CSA). Please refer to Table 3 for the results. The results in PatchGastricADC22 indicate that self-attention and correlated self-attention among instances yield similar performance. However, in the case of ABO, correlated self-attention outperforms self-attention. We posit that this discrepancy arises from the fact that images of e-commerce products typically do not exhibit explicit correlations. In the correlated self-attention mechanism, images are initially aggregated into query embeddings, which may help reduce the impact of irrelevant information. Due to space constraints, we have postponed additional case studies to supplementary C.2 and more visualization results to supplementary C.3.

5. Conclusion

In this paper, we introduce the MIVPG, a flexible, general and powerful component to fuse enriched visual representations with MLLMs, which achieves superior performance in diverse use cases. We demonstrate that QFormer is a limited variant of MIVPG, providing theoretical support for its efficacy.We believe the enriched visual signals and advanced MIL-based techniques will contribute to the future development of MLLMs.

References

[1] Famke Aeffner, Mark D Zarella, Nathan Buchbinder, Marilyn M Bui, Matthew R Goodman, Douglas J Hartman, Giovanni M Lujan, Mariam A Molani, Anil V Parwani, Kate Lillard, et al. Introduction to digital image analysis in whole-slide imaging: a white paper from the digital pathology association. *Journal of pathology informatics*, 10(1):9, 2019. 1

[2] JB Alayrac, J Donahue, P Luc, A Miech, I Barr, Y Hasson, K Lenc, A Mensch, K Millican, M Reynolds, and R Ring. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1, 2, 3, 4

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[5] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019. 2

[6] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018. 2

[7] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21126–21136, 2022. 1, 7, 2

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3, 1

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[10] Ji Feng and Zhi-Hua Zhou. Deep miml network. In *Proceedings of the AAAI conference on artificial intelligence*, 2017. 2

[11] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 6

[13] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019. 1

[14] Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2424–2433, 2016. 2

[15] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 6

[16] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 1, 2, 3

[17] Fahdi Kanavati, Gouji Toyokawa, Seiya Momosaki, Michael Rambeau, Yuka Kozuma, Fumihiro Shoji, Koji Yamazaki, Sadanori Takeo, Osamu Iizuka, and Masayuki Tsuneki. Weakly-supervised learning for lung carcinoma classification using deep learning. *Scientific reports*, 10(1):9297, 2020. 2

[18] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823*, 2023. 2

[19] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pages 3744–3753. PMLR, 2019. 3, 5, 1

[20] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021. 1, 2

[21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 2, 5, 6, 7

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1, 2

[25] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29, 2016. 1

[26] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021. 2

[27] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models, 2023. 7

[28] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997. 2

[29] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. 1

[30] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[31] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1

[32] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017. 1

[33] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2

[34] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021. 2, 4, 5

[35] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 6

[36] Masayuki Tsuneki and Fahdi Kanavati. Inference of captions from histopathological patches. In *International Conference on Medical Imaging with Deep Learning*, pages 1235–1250. PMLR, 2022. 1, 6, 2

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[38] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427, 2017. 1

[39] Peiqi Wang, William M Wells, Seth Berkowitz, Steven Horng, and Polina Golland. Using multiple instance learning to build multimodal representations. In *International Conference on Information Processing in Medical Imaging*, pages 457–470. Springer, 2023. 2

[40] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, 2018. 2

[41] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14138–14148, 2021. 5

[42] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022. 2

[43] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *Proceedings of the IEEE international conference on computer vision*, pages 4894–4902, 2017. 1

[44] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016. 1

[45] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017. 3

[46] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18802–18812, 2022. 1, 2

[47] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 2

[48] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 2

[49] Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang. Wsisa: Making survival prediction from whole slide histopathological images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7234–7242, 2017. 1

Enhancing Multimodal Large Language Models with Multi-instance Visual Prompt Generator for Visual Representation Enrichment

Supplementary Material

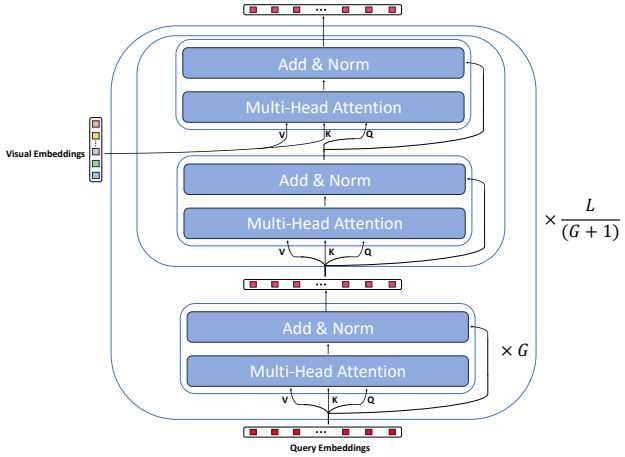


Figure 7. Overview of QFormer

A. Detailed Architecture of QFormer

The architecture overview is depicted in Figure 7. Specifically, QFormer is initialized as a BERT-based model[8] comprising a total of $L = 12$ layers. In contrast to typical BERT models that process textual inputs, QFormer takes $R = 32$ learnable query embeddings as inputs. These embeddings are utilized to extract visual information from the input visual data during Stage-1 pretraining in BLIP2[22]. Subsequently, they serve as visual prompt embeddings for the LLM inputs after projection.

Inside the QFormer, each layer includes a self-attention module composed of a Multi-Head Attention component and a Forward module (consisting of Linear, LayerNorm, and Residual Connection). The cross-attention module, initialized with random values, is inserted every G layers, where learnable query embeddings interact with visual embeddings. *In the main paper, for the sake of conciseness, we condensed the representation of the multi-head attention and forward modules into self(cross) attention modules. Furthermore, we exclusively illustrated the modifications made to the cross-attention module in MIVPG, as the self-attention modules remain unchanged.* The final QFormer output is represented by the last layer's query embeddings.

For a more comprehensive understanding, readers are encouraged to refer to [22].

B. Proof of Proposition

In Proposition 2, we illustrate that MIVPG, when augmented with the CSA (Correlated Self-Attention) module, maintains the crucial permutation invariance property of MIL. In this section, we provide a theoretical demonstration of this property.

Proof. Recall that both the original cross-attention and self-attention mechanisms have already demonstrated permutation equivalence for the visual inputs (Property 1 in [19] and Proposition 1 in the main paper). Our objective is to establish that the CSA module also maintains this permutation equivalence, ensuring that the final query embeddings exhibit permutation invariance.

A permutation-equivalence function f satisfies the property $\pi(f(x)) = f(\pi(x))$, where π is a permutation of the element order in the set x. In the context of this paper, we have $f(x = B) = \text{Attention}(Q = B, K = q, V = q)$. Here, we use B_π to denote the bag after permutation, and Q_B, K_q, and V_q to denote the Q, K, V matrices in the attention calculation after projection.

$$
\begin{aligned}
f(\pi(x)) &= \text{Attention}(Q = B_\pi, K = q, V = q) \\
&= \text{softmax}(\frac{Q_{B_\pi} K_q^T}{\sqrt{D}}) V_q \\
&= \pi(\text{softmax}(\frac{Q_B K_q^T}{\sqrt{D}})) V_q \\
&= \pi(\text{softmax}(\frac{Q_B K_q^T}{\sqrt{D}}) V_q) \\
&= \pi(\text{Attention}(Q = B, K = q, V = q)) \\
&= \pi(f(x))
\end{aligned}
\tag{10}
$$

In Equation 10, recall that $\text{softmax}(\frac{Q_{B_\pi} K_q^T}{\sqrt{D}})$ is the attention map with shape $\mathbb{R}^{M \times R}$ where M is the number of elements in the bag and R is the number of query tokens. The permutation of the bag is row-wise, and therefore $\text{softmax}(\frac{Q_{B_\pi} K_q^T}{\sqrt{D}}) = \pi(\text{softmax}(\frac{Q_B K_q^T}{\sqrt{D}}))$. Hence, the CSA module is permutation equivalence. Given that other properties of the original self-attention and cross-attention mentioned above, the MIVPG enhanced by the CSA module remains permutation invariant. □

C. More Experiments

We implemented the proposed method on NVIDIA A100 GPUs with BFloat16. Except for the number of train-

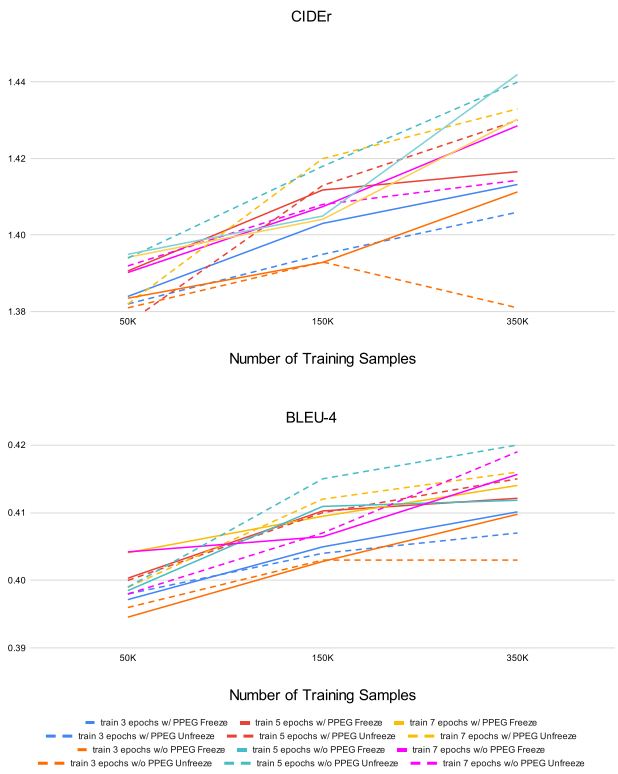


Figure 8. Experiment results on MSCOCO with or without freezing the visual encoder. We adopt the metrics used in [22].

ing epochs mentioned in the main paper, we kept all other hyperparameters the same as in BLIP2[22]. For PatchGastricADC22[36] and ABO[7], we trained the model for 40 epochs.

C.1. Frozen Visual Models

In the original BLIP2[22], image sizes are upscaled to 364×364, and consequently, the ViT is unfrozen during the fine-tuning process. This approach yields slightly better performance, albeit at a higher computational cost while training on the entire COCO training set.

In this section, we validate the performance of fine-tuning while keeping the ViT frozen and image sizes unchanged. Experiment results can be seen as Figure 8. We observed that when working with limited data, such as 50K samples, models exhibit comparable performance whether or not the visual encoder (ViT) is frozen. However, as the number of training epochs increases, the performance gap varies. In some cases, unfreezing the ViT leads to improved performance, while in others, the opposite holds true. Considering that many real-world applications may not have access to massive training data, freezing the ViT can be a more efficient approach while still maintaining similar performance levels.

C.2. Case Study

In the main paper, we employ the FLAN-T5-XL as the language model. Existing large language models can be broadly categorized into two types: encoder-decoder based and decoder-only based models. The FLAN-T5-XL falls into the former category. The decoder-only based models are more computationally efficient and the encoder-decoder based models can handle more sophisticated tasks. In this section, we assess the performance of MIVPG on models from the decoder-only category. Specifically, we use the BLIP2[22] with OPT-2.7b[47] as the base LLM. We validate the performance on the PatchGastricADC22 dataset. In the experiments, we only replace the LLM while keeping other hyperparameters unchanged.

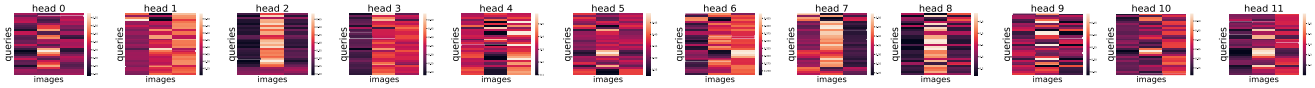
Table 4. Experiments on the PatchGastricADC22 dataset [36] with OPT-2.7b as the language model

	BLEU@4	CIDEr	METEOR	ROUGE
BLIP2-MIVPG w/o CSA	0.427±0.012	3.12±0.118	**0.349**±0.003	0.494±0.123
BLIP2-MIVPG	**0.432**±0.025	**3.21**±0.105	0.347±0.016	**0.569** ±0.019

The experiment results on PatchGastricADC22 using OPT-2.7b as the language model are presented in Table 4. Overall, the model continues to outperform the baselines shown in Table 1, emphasizing the advantages of integrating MLLMs into the WSI captioning task. Notably, the model with CSA performs better than the one without it, reaffirming the effectiveness of CSA. It's also worth noting that the performance of using OPT-2.7b is not superior to using Flan-T5-XL. This could be attributed, in part, to the insufficiency of training data. Since OPT-2.7b is relatively less sophisticated, more training data may be required to train a more powerful model.

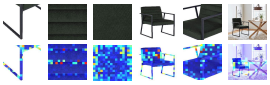
C.3. More Visualization

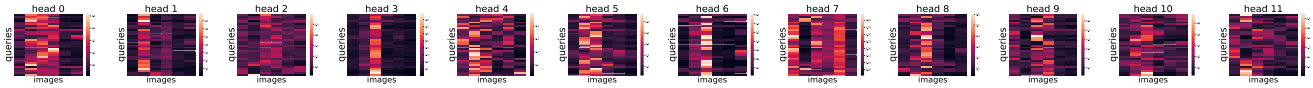
This section provides additional visualization results on the ABO dataset, including both patch-level attention weights and image-level attention weights. In the patch-level attention weights, it is evident that the model excels in detecting the shapes of objects, as a significant portion of the patch-level weights is assigned to edges and contours. The image-level attention weights display maps for all twelve heads. Each row in a map represents a query, while each column represents an image. It's important to note that different heads and queries exhibit varying attention patterns towards the images, demonstrating the diversity in how the model processes and attends to the input images.



References: *['a short bar with weights at each end that is used usually in pairs for exercise.', 'gymnastic weight for dogs.', 'a six sided barbell.', 'a hexagonal shaped dumbbell', 'pesa de mano con exterior hexagonal']*

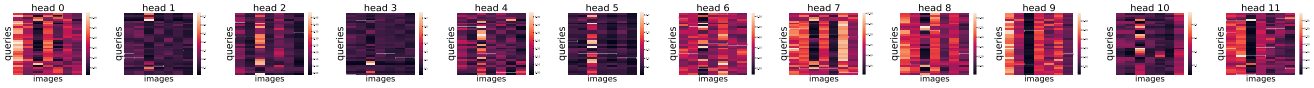
Inference: *a six sided barbell.*





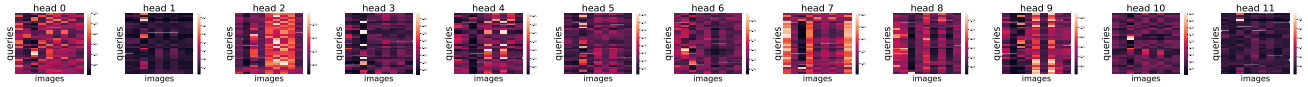
References: *['a chair with a metal square like right forming the armrests and the legs and also having lines on the backrest and the seat.', 'the chair is composed of a seat and a square backrest with two armrests and two square legs on each side', 'a one seater chair with flat metal armrests extended to form the legs.', 'a one seater chair with flat metal armrests extended to form the legs and having rows patterns on the backrest and seat pillow.']*

Inference: *a one seater chair with flat armrests and having four thin standing legs.*



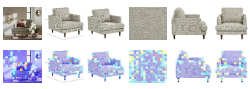
References: *['a stool with raised legs and a curved seat.', 'a rectangular top stool with four standing legs.', 'gray chair without arms or back with four legs that support it.', 'a stool with a rectangular cushion and has four long legs', 'a rectangular top stool .']*

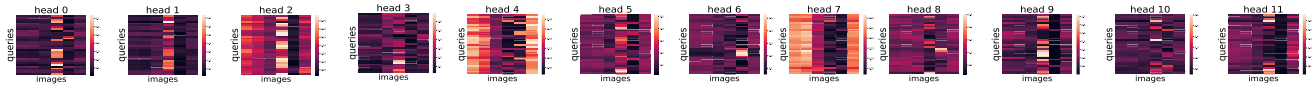
Inference: *a rectangular top stool with four standing legs.*



References: *['single seater tufted sofa with seat attached to long back and side arms having for wooden legs.', 'a one seater chair with hole pattern on the backrest and having low armrests.']*

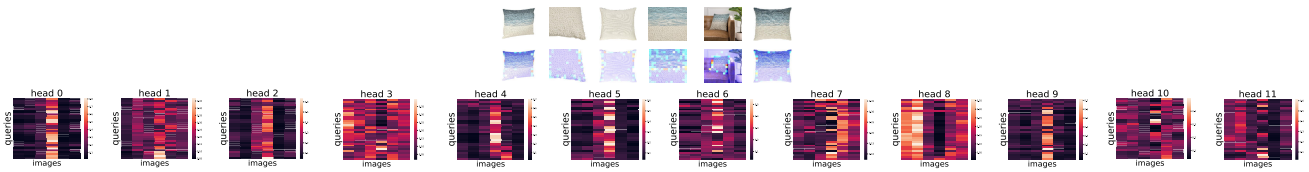
Inference: *a one seater chair with hole pattern on the backrest and having four thin standing legs.*





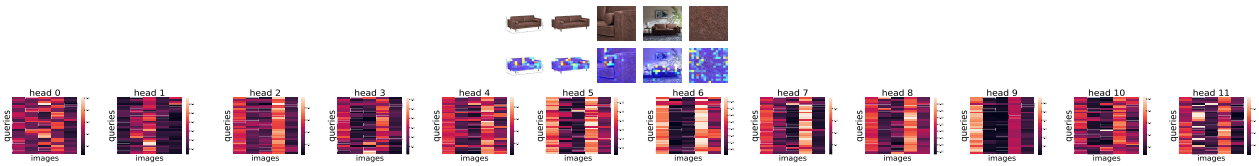
References: *['a one seater sofa with huge back pillow and broad seat pillow and having four thin standing legs.', 'a broad one seater sofa with four thin standing legs.', 'one seater sofa with armrests on each side and four short legs. the seat and backrest have rounded rectangular cushions', 'one seater sofas wide with medium legs']*

Inference: *a one seater sofa with huge back pillow and having four thin standing legs.*



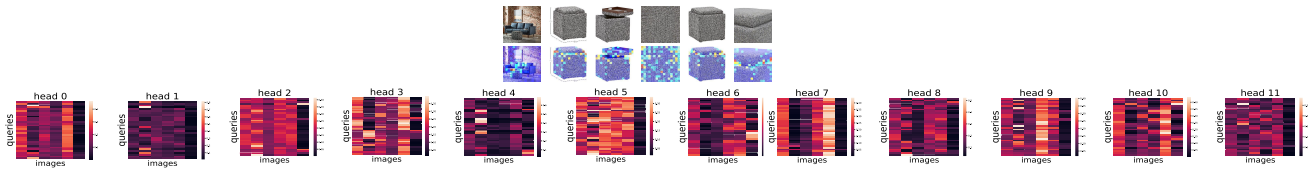
References: *['durable spongy waterproof small size pillow.', 'a structure, observed in certain extrusive igneous rocks, that is characterized by discontinuous pillow-shaped masses ranging in size from a few centimeters to a meter or more in greatest dimension (commonly between 30 cm and 60 cm).', 'square pillow with cushions', 'sofa,bed pillow, square in shape']*

Inference: *a square shaped pillow*



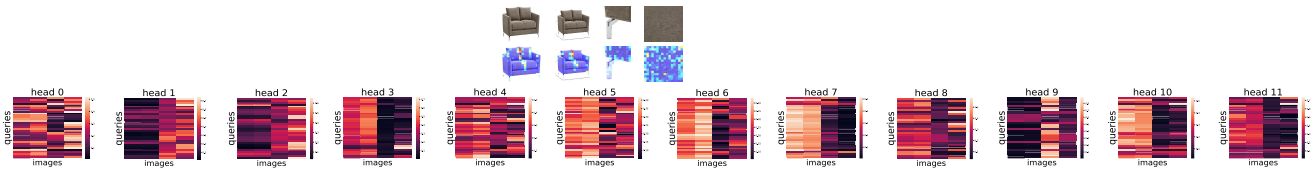
References: *['a two seater sofa with four thin standing legs, cylindrical pillow at each armrest and square pattern on the seat.']*

Inference: *a two seater sofa with square pillow at each armrest and having four thin standing legs.*



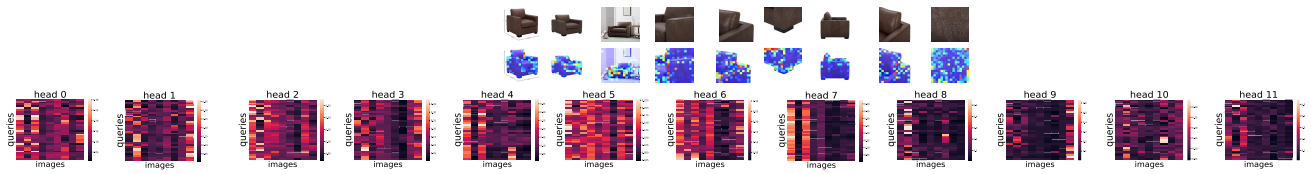
References: *['a cubical shaped ottoman seat.', 'a seat with a square shaped cushion', 'a cube shaped ottoman seat.']*

Inference: *a cube shaped ottoman seat with square shaped legs.*



References: ['the sofa consists of two huge pieces with armrests. it has two big size pillows and four little legs.', 'the couch has square set and back consists of square pillow held up by four curved legs.', 'a two seater sofa with huge back pillows and four thin standing legs.', 'two seater sofa with armrests on each side. it has rounded rectangular cushions on the seats and backrest']

Inference: *a two seater sofa with huge back pillows and having four thin standing legs.*



References: ['a one seater sofa with broad seat pillow.', 'padded gray sofa.', 'a one seater sofa with broad seat and four small standing legs.']

Inference: *a one seater sofa with huge back pillow and seat pillow and having four short standing legs.*