MIVC: Multiple Instance Visual Component for Visual-Language Models

Wenyi Wu
Amazon
424 5th Ave, New York, NY
wenyiwu@amazon.com

Qi Li
Amazon
410 Terry Ave N, Seattle, WA
qlimz@amazon.com

Wenliang Zhong
The University of Texas at Arlington
500 UTA Boulevard, Arlington, TX
wxz9204@mavs.uta.edu

Junzhou Huang
The University of Texas at Arlington
500 UTA Boulevard, Arlington, TX
jzhuang@uta.edu

Abstract

Vision-language models have been widely explored across a wide range of tasks and achieve satisfactory performance. However, it's under-explored how to consolidate entity understanding through a varying number of images and to align it with the pre-trained language models for generative tasks. In this paper, we propose MIVC, a general multiple instance visual component to bridge the gap between various image inputs with off-the-shelf vision-language models by aggregating visual representations in a permutation-invariant fashion through a neural network. We show that MIVC could be plugged into the visual-language models to improve the model performance consistently on visual question answering, classification and captioning tasks on a public available e-commerce dataset with multiple images per product. Furthermore, we show that the component provides insight into the contribution of each image to the downstream tasks.

1. Introduction

In recent years, numerous efforts [11, 38] have been made to integrate images and text with multimodal models that typically utilizes distinct encoders for different modalities of data (e.g., CNNs as visual encoders and RNNs as text encoders). These encoders are subsequently fused in a shared embedding space. More recently, with the evolution of Transformer architectures, several studies [18,20,39] have sought to unify vision and text using text and vision Transformers. These methods commonly combine information from images and text tokens, enabling collaborative attention mechanisms within the Transformer for enhanced information fusion.

Despite the remarkable achievements of these methods in various multimodal tasks such as Visual Question Answering (VQA) [2] and Image Captioning [29, 36], there is an evident limitation. They often assume that input images and text are paired, meaning one image corresponds to one piece of text. However, in practical scenarios, this assumption can be challenged, as not all tasks involve a one-to-one relationship between image and text. For instance, when presented with two images, we may seek a textual description highlighting their differences. In this context, there exists a one-to-two relationship between text and images. Furthermore, multiple images may correspond to a single piece of text, especially when describing a complex object. For example, in e-commerce platform, each product is displayed with different background, from different angles or focusing on local details to provide enriched information. These images are correlated and essentially representing the same entity and therefore, it's crucial to learn a consolidated entity representation consolidating all images that could be aligned with the pre-trained language models for general generative tasks.

Given that state-of-the-art (SOTA) multimodal models [8, 18] are primarily pre-trained on the one-image-one-text paradigm, directly inputting multiple images with one piece of text is unfeasible. Consequently, existing approaches typically address this issue through two methods: (1) when processing image inputs, they concatenate multiple images into a single "concentrated" image, or (2) they employ multiple images' embeddings obtained via encoders as input, although this often requires fine-tuning to adapt the model to multiple visual embeddings. However, these methods are simplistic in their approach to fusing information from multiple images. How to more effectively integrate one-to-many or many-to-many of image-text data remains an open question.

In this paper, we tackle the challenge regarding how

to consolidate information from multiple images and text within a visual-language model, particularly when using multiple images to describe a single object. This problem is crucial in e-commerce [7], where a product is typically represented by multiple images along with a textual description to comprehensively convey its attributes. Notably, the scenario of multiple images describing an object differs from traditional multi-view problems, where multiple images possess information about relative positions. In our case, multiple images are simply used to describe the same object without requiring strong assumptions among them. Inspired by the Multiple-Instance Learning (MIL) problem [5, 24] where each input contains the varying number of entities, forming a set referred to as a bag, we consider the input images as a bag as well and aim to learn a consolidated representation per bag.

Specifically, we leverage off-the-shelf vision encoders to convert each image into a representation and employ attention mechanisms to effectively combine multiple images within a bag through multiple instance learning. These combined embeddings are then used as input, alongside text, to the off-the-shelf language models for generative tasks. This approach not only allows us to accept multiple images as input but also identifies the most relevant images for the task through attention, thereby enhancing the model's performance and providing interpretability. In summary, our primary contributions include: (1) we introduce a groundbreaking Multiple Instance Learning (MIL) component, MIVC 1, in the realm of multimodal representation learning. Our novel framework enables the adaptive integration of multiple images with textual data, a critical advancement in handling complex multimodal information; (2) through our innovative MIL framework, we achieve a significant improvement in multimodal representation learning. This enhancement contributes to more effective fusion of information from diverse modalities, promising substantial benefits in various applications. We validate the effectiveness of the proposed method through extensive experimentation on the publicly available Amazon Berkeley Objects Dataset (ABO) [7]. Our empirical results demonstrate its prowess in addressing real-world tasks, underlining its practical utility and robustness; and (3) providing insights into the contribution of each image to the generative tasks.

2. Related Work

2.1. SOTA Visual-Language Models

With the continuous evolution of deep learning, an increasing number of research endeavors have shifted their focus towards the fusion of different modalities to address more complex scenarios in the real world. For instance, in Visual Question Answering (VQA), users may pose questions to models based on a set of images, expecting answers. In image captioning, users provide a set of images, asking the model to generate descriptive text regarding the content of these images. In early works [2, 11, 38], images and text were separately transformed into features using ResNet and LSTM, followed by concatenation before being fed into a prediction layer for inference.

In recent years, with the emergence of Transformers [32], there has been a revolutionary shift in the universal network architecture in the field of Natural Language Processing (NLP). The utilization of the global attention mechanism within Transformers has become increasingly prevalent. Subsequently, exploration of Transformers in the field of Computer Vision (CV) has also taken flight. Vision Transformers (ViT) [9], for example, employ the same Transformer architecture as in NLP but divide images into several patches to be treated as vision tokens as input. With extensive pre-training on large scale data, ViT has demonstrated superior performance to ResNet. Through Transformers, CV and NLP have achieved structural unification which enables multimodal model development. For example, phrase grounding [12, 34] aligns the visual signals with arbitrary caption words semantically, which extends the object detection task beyond the fixed list of categories in the label set.

More recently, with the rise of generative models [3, 27, 31], the application of multimodal capabilities to generative tasks has become an open question. DALL-E [28], for instance, embarked on a pretraining task where images were tokenized, enabled text-to-image generation. Subsequently, various vision-language models (VLMs) have been proposed to enhance the fusion of text and images. For example, BLIP2 [18] introduced the use of a Q-Former to align images more effectively with the input space of text. TCL [35] employed triplet contrastive learning to simultaneously learn from text and images. FROMAGe [16] adopted a multitask approach to train a model for image captioning and image retrieval.

While these multimodal models have achieved substantial success across various tasks, they are predominantly built upon a crucial assumption - that a single piece of text pairs with a single image as input. However, in the real world, text and images may exhibit one-to-many or many-to-many relationships. How to effectively handle multimodal models in such scenarios remains an open question.

2.2. Multiple Instance Learning

Traditionally, Multiple Instance Learning (MIL) [5, 24] can be broadly categorized into two main types: (1) Bag-Level Prediction [4, 10, 13, 15]: In this approach, bag-level predictions are directly derived from instance-level predictions. (2) Bag-Level Prediction with Feature Aggregation [14, 17, 21, 30]: Here, bag-level predictions are generated

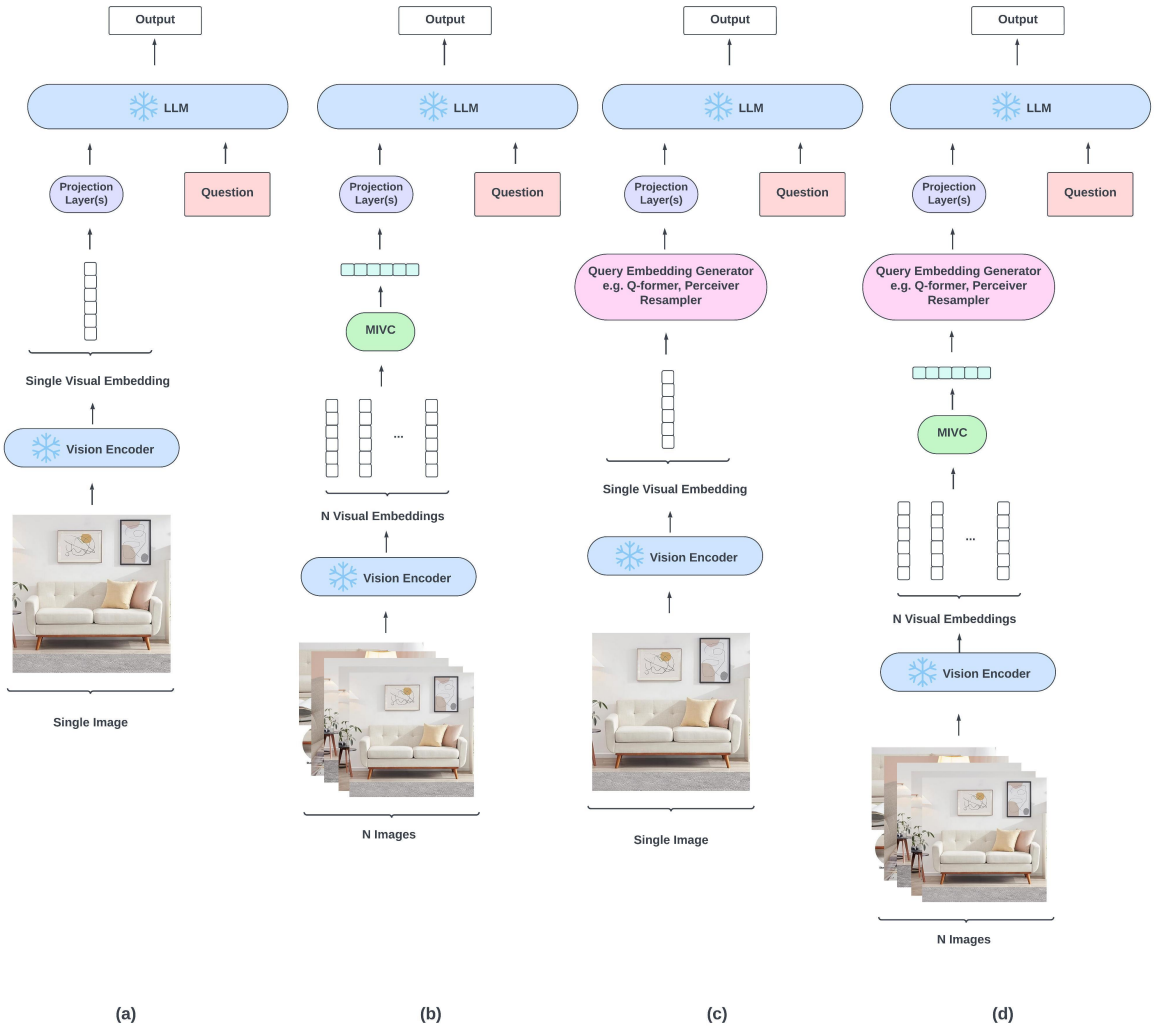

Figure 1. MIVC adaptation with off-the-shelf visual-language models. Both (a) and (b) utilize foundation vision encoder's output image patch embedding as visual representation. In contrast to (a), (b) is adapted with MIVC, which can take multiple images as input. Both (c) and (d) employ query embedding generator (e.g. Q-former [18] or Perceiver Resampler [1]) after vision encoder for visual representation generation. In contrast to (c), (d) is adapted with MIVC which can takes multiple images as input. In this paper, we use T5-XXL as the language model and ViT as the vision encoder.

by aggregating the features of all instances. For the former, often, hard-crafted pooling operators such as mean pooling or max pooling are employed. However, in practical applications, these hard-crafted pooling operators often yield limited results.

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 [33] utilizes a fully connected layer in MIL. Furthermore, AB-MIL [14] employs attention during the pooling process, allowing for better weighting of different instances. Another category of methods attempts to consider the relationships between different instances using graph neural networks or capsule neural networks. More recently, DS-MIL [17] employs attention not only to consider instance-to-instance relationships but also instance-to-bag relationships; DTFD-MIL [37] incorporates the Grad-CAM mechanism into AB-MIL. While all these approaches focus on single modality, we adopt the effective attention mechanism proposed by AB-MIL to consolidate visual features in the visual-language models.

3. Method

3.1. Architecture Overview

By incorporating visual models with the capabilities of pre-trained Large Language Models (LLMs), multimodal LLMs have demonstrated dramatic improvements in various tasks, such as visual question answering (VQA), captioning, and etc. The majority of recent multimodal LLMs [8, 18, 20] share a similar framework by utilizing separate vision and text towers to independently encode the two modalities first. The encoded single modality representations are then fused together, e.g. by projecting image representation via a single or multiple projection layers, or by directly concatenating, and then fed into LLMs. Depending on how image representation embedding is generated, it can be further categorized into two types: first, image patch embedding based vision tower in Figure 1 (a), which is generally composed of a single visual foundation model encoder (e.g. ViT [9]) and utilizes the generated image patch embedding directly as visual representation; and second, image patch and query embedding based vision tower in Figure 1 (c), which sequentially combines a vision foundation model's encoder and a query embedding module (such as Q-former in BLIP-2 [18] or the Perceiver Resampler as in Flamingo [1]).

One constraint to such a framework is the lack of capability to process multiple image inputs per request, when all images contribute and correspond to a single label. These multiple images, also referred to as multiple instances, typically carry complementary information; therefore they are more informative than a single instance for the corresponding task and shouldn't be ignored. Such applications are not rare in industry and other scientific areas, including utilizing multiple product images corresponding to a single product for e-commerce-related classification, caption generation, product information inference, synthesizing multiple X-ray images for medical diagnosis [14], and geological simulation from multiple underground mapping [19], among others. To the best of our knowledge, all current visual-language models only consider a single image instance as input. Although it can be adapted to multiple image instances, this is largely achieved through customization at the input stage, either by taking only one single image as input or by concatenating multiple raw images into a single image. This results in either information loss, as multiple images' information is not efficiently synthesized, or a computational burden on LLM inference when dealing with one large-scaled concatenated image embedding.

To address the accuracy and efficiency challenges, we propose MIVC, a general multiple-instance visual component that bridges the gap between multiple image inputs and any off-the-shelf Vision Language Models (VLMs). The proposed component can robustly handle both multiple image instances learning and single image instance learning. In particular, as illustrated in Figure 1 (b, d) , compatible with any VLMs, we attach MIVC directly after the vision encoder. Multiple images are fed to the vision encoders to generate multiple visual representations via any VLMs' vision tower which retrieves the visual information from each individual image instance. The generated visual representations are then fed into MIVC to generate a single pooling image representation. This fused image representation not only retains essential information from multiple image instances but is also concise enough without introducing extra computational cost in the following LLM inference stage, where the pooling image representation is concatenated with text embedding and fed into the LLM for final inference. In this paper, to illustrate the effectiveness of MIVC, we use off-the-shelf pre-trained language model T5-XXL [6] as the large language component and ViT [9] as the vision encoder. We compare the performance and computational complexity of the aforementioned natural alternatives with MIVC in the following sections.

3.2. MIVC Methods

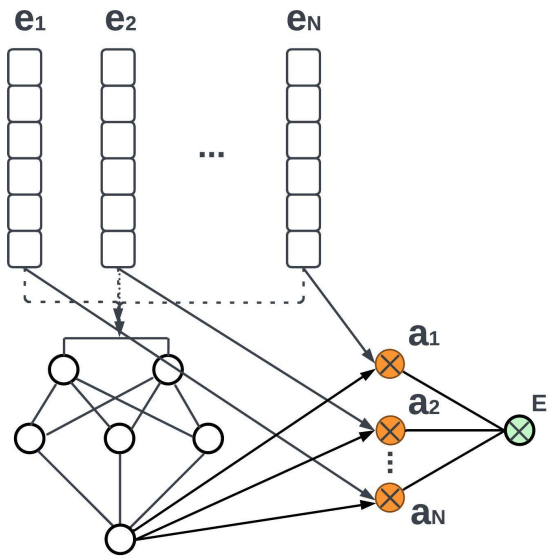

Figure 2. Detail structure of MIVC with attention pooling. The neural networks are trained to consolidate visual representations in a permutation-invariant fashion. The orange neurons represent the contribute of each image to the downstream generative tasks.

Problem Statement In the multiple instance learning problem, we have a bag of N instances, in any order, associated with a single label. In our problem setting, a single input data, $X = I_1, I_2, ..., I_N, T, L$, is composed of N images ($I_n \in \mathbb{R}^{L \times H \times C}$, where $n \in (1, N)$), along with a text prompt (T) and a text output (L). The value of N varies for different input data. For a vision encoder f in Vision

Language Models (VLMs), we can generate N vision representations:

$$\forall_{n=1,...,N} : e_n = f(I_n) \quad s.t. \ e_n \in \mathbb{R}^M \tag{1}$$

On top of that, we apply MIVC, our proposed multiple instance learning visual component, to generate a single pooling image representation:

$$E = MIVC(e_1, ..., e_N) \tag{2}$$

In case of 2-dimensional vision representation, we first flatten them and then convert back to the original dimensions after MIVC. We explore and evaluate four types of pooling strategies in the paper.

Multiple Instance Pooling Strategy

Following [14], we implemented four different embedding pooling strategies in MIVC :

- Average pooling. Average operator across multiple image instances representation embeddings.

$$E = \frac{1}{N} \sum_{n=1}^{N} e_n \tag{3}$$

- Max pooling. Maximum operator across each dimension of multiple image instances representation embeddings.

$$\forall_{n=1,...,N} : E_n = \max_{m=1,...,M} e_{nm} \tag{4}$$

- Attention pooling. We illustrate attention based pooling in the Figure 2. It's a weighted average of multiple image instances representation embeddings.

$$E = \sum_{n=1}^{N} \alpha_n e_n, \tag{5}$$

where:

$$\alpha_n = \frac{exp\{w^T tanh(Ze_n^T)\}}{\sum_{j=1}^{N} exp\{w^T tanh(Ze_j^T)\}} \tag{6}$$

$$s.t. \sum_{n=1}^{N} \alpha_n = 1 \tag{7}$$

in which $w \in \mathbb{R}^{K \times 1}$ and $Z \in \mathbb{R}^{K \times M}$

- Gated Attention pooling. The gated attention pooling introduces more non-linearity, and pooling more rich visual information across the multiple image instances.

$$E = \sum_{n=1}^{N} \alpha_n e_n, \tag{8}$$

where:

$$\alpha_n = \frac{exp\{w^T (tanh(Ze_n^T) \otimes sigm(Ge_n^T))\}}{\sum_{j=1}^{N} exp\{w^T (tanh(Ze_j^T) \otimes sigm(Ge_j^T))\}} \tag{9}$$

$$s.t. \sum_{n=1}^{N} \alpha_n = 1 \tag{10}$$

in which $w \in \mathbb{R}^{K \times 1}$, $Z \in \mathbb{R}^{K \times M}$, $G \in \mathbb{R}^{K \times M}$, and $\otimes$ represents element-wise mulplication operator.

4. Complexity

We measure the model's complexity by calculating the number of parameters associated with each pooling method, as detailed in Table 1. From the table, it can be observed that both the average and max pooling methods introduce no additional trainable parameters, resulting in the same overall parameter count as the original BLIP2 model. In contrast, the Attention and Gated Attention methods introduce new parameters due to the inclusion of attention modules. However, the proportion of these additional parameters is minimal, accounting for only 0.7% and 1.5% of the total model parameters, respectively. Consequently, their supplementary computational overhead is negligible.

Models w/ pooling	# params
T5-XXL+ViT	12.23B
T5-XXL+ViT w/ avg or max	12.23B
T5-XXL+ViT w/ attn (extra params)	12.32B (92.23M)
T5-XXL+ViT w/ gated attn (extra params)	12.42B (185.63M)

Table 1. Model complexity in terms of the number of parameters. This directly impacts the inference efficiency.

5. Training

For attention and gated attention based MIVC, it learns the neural network w,Z and G which determine how each image contributes to the downstream generative tasks. To make fair comparison between different pooling methods and vanilla alternatives, besides the zero-shot evaluation of the off-the-shelf models, we further fine-tune all models to report performance. We use product images and textual metadata to train the image-textual alignment layers and MIVC simultaneously using generative tasks. Both dataset and task details are presented in 6.1 and 6.2. We freeze the visual encoder and the language model during our training procedure.

6. Experiment and Analysis

6.1. Datasets

The products in the e-commerce website are presented by one main image and multiple images from different

views or with zoomed in images to display details, such as patterns, flavours and etc. We leverage ABO dataset [7] with 147,702 products which contains multiple images and textual metadata as shown in the e-commerce website. The number of images per product ranges widely from 2 to 21 and images could be with different background, angles and focal lengths. The main image is commonly attractive by putting the product into a live scene which makes it hard to focus on the correct entity or detail region for generative tasks requiring detailed vision signals. On the contrary, the subsequent images with white background or of details benefit generative models with fine-grained visual representations. The textual metadata mainly contains one short descriptive sentence to be displayed as the title in the e-commerce product page and seven product attributes, e.g., color, pattern, style and etc. We illustrate one product example and the generated caption using the MIVC-BLIP2 model in Figure 3.

6.2. Tasks

In total, we have 3 types of tasks. We split each of them into 80% training set and 20% evaluation set. All training sets are mixed and the model are trained with the unified generative tasks [26] despite they are evaluated differently.

Categorization Product categorization is an important task for e-commerce which benefits search and recommendation experience. In ABO, there exist hundreds of categories with a long tail distribution. We keep 10 high frequency and representative categories to test our MICV including furniture, shoes and etc. We intent to keep categories that are not trivial to distinguish like chair and sofa. We form this as a multi-choice visual question answering task with 10 options and feed the MIVC-BLIP2 model with multiple product images, the categorization question and 10 candidate categories. We compare the performance using accuracy, macro average precision, and recall across 10 categories.

Product Information Inference Besides the product categorization, we further look into detail product metadata from seven attributes: style, color, finish type, pattern, fabric type, material and shape. Following the science QA [22] format, we frame them into multi-choice questions and generate answer based on images and the prompt. Because the metadata is not as clean as scientific question answering dataset, for example, 100% cotton and pure cotton both exist, we clean the dataset and apply regex to only keep alphabetic characters. We further keep top frequent values per attribute to be used as multi-choice answers. The number of values ranges from 3 to 5 across different attributes. For example, we have solid, textual, floral, geometric and striped in pattern. We compare the performance using accuracy, macro average precision, and recall across choices for each attribute and report the aggregated performance across all attributes.

Image Captioning We use the general image captioning prompts [8] to ask the model to generate a short descriptive sentence of the product using multiple images. Instead of using original product titles, which are often less descriptive, we leverage the manually annotated captions [23] on the same dataset as the reference captions to measure the quality of our generated captions. It's not trivial to evaluate the quality of the generated captions because each product could be described from very different perspectives. We illustrate how the generated title could be different from the reference title but still very relevant to the image content in Figure 3. Therefore, we feed the generated captions, annotated captions and the main product image to the pre-trained CLIP model [25] to retrieve text given image. We report the text retrieval top 1 recall as our metrics, which is the proportion of samples whose generated caption has the highest similarity with the image compared to all annotated captions.

6.3. Benchmark Models

Single Image We evaluate the T5-XXL and ViT on the aforementioned data and tasks using only the first image in zero-shot fashion.

Concatenated Image Another vanilla approach to infer visual signals from multiple images is to concatenate all images together, as being illustrated in [8, 20]. Because the number of images could be as many as 21, horizontal concatenation will lead to aspect ratio challenge. Therefore, we concatenate images in a square grid such that 4 images are concatenated in a 2 by 2 grid, 5 to 9 images are concatenated in a 3 by 3 grid with blank image fill-in and so on and so forth. We evaluate the BLIP2 on the aforementioned data and tasks using the concatenated image in zero-shot fashion.

Pooling	Accuracy	Precision	Recall
Single (zs)	97.1%	97.1%	97.1%
Concat (zs)	97.5%	97.6%	97.5%
Single	97.2%	97.3%	97.2%
Concat	97.8%	97.8%	97.8%
MIVC-Avg	96.9%	97.0%	96.9%
MIVC-Max	94.7%	94.7%	94.7%
MIVC-Attn	97.9%	97.9%	97.9%
MIVC-gated	97.4%	97.4%	97.4%

Table 2. Categorization performance comparison. We illustrate the effectiveness of MIVC with T5-XXL and ViT as language model and the vision encoder, respectively. We first evaluate them in zero-shot (zs) fashion and then fine-tune the model with and without various MIVC pooling method. The reported precision, recall and f1-score are macro average across 10 categories.

Metadata: {item name: Amazon Brand – Stone & Beam Elise Upholstered Barstool, 42.5\"H, Midnight Blue, Material: Wood, ...}
Annotated Caption: A chair with raised legs and having no armrests.
Image Captioning: A bar stool with a blue upholstered seat

Figure 3. Data illustration. The e-commerce product contains one attractive main images (the leftmost image), several detailed images and textual metadata. Instead of item name displayed in e-commerce website, we use manually annotated image captions to measure captioning performance.

6.4. Results and Analysis

We summarize the performance of three tasks in the following tables. From the table 4, the results show that the proposed MIVC with attention pooling outperforms all benchmarks on the 10 categories classification task. From the table 3, we observe that the MIVC with attention pooling outperforms the benchmark methods by selecting the most accurate options from candidates in the product inference task. It improves the performance the most by 9% accuracy, compared to the single image benchmark. It aligns with our conjecture that including additional images would benefit generative tasks with fine-grained visual information. From the table 4, we show that with MIVC, the general visual-language can generate high quality comprehensive titles.

Pooling	Accuracy	Precision	Recall
Single (zs)	64.5%	65.3%	63.0%
Concat (zs)	65.8%	68.9%	65.5%
Single	62.9%	62.8%	62.9%
Concat	66.0%	68.0%	65.0%
MIVC-Avg	64.7%	65.7%	63.7%
MIVC-Max	65.8%	66.7%	64.7%
MIVC-Attn	67.4%	72.7%	70.1%
MIVC-gated	66.9%	68.5%	65.1%

Table 3. Product attribute inference performance comparison. We illustrate the effectiveness of MIVC with T5-XXL and ViT as language model and the vision encoder, respectively. We first evaluate them in zero-shot (zs) fashion and then fine-tune the model with and without various MIVC pooling. The reported precision, recall and f1-score are first macro average across the number of options per attribute and then simple averaged across tasks.

6.5. Ablation: Concatenation

Besides concatenating raw images, another natural alternative is to concatenate the image representations generated from the vision encoder and then project the concatenated image embedding to lower dimension space. For the BLIP series model using Q-former, the image representation is 257×1408 dimension. In order to project N concatenated image embeddings back to the same dimension, it results $N \times 97B$ parameters where N is the maximum number of images per product , i.e. 21. To make the number of parameters under control, we first limit the maximum number of input images to 6 and then map image representations to a lower dimensional 2048 before mapping to the proper dimension to the Q-former. We ends up with 4B parameters. To understand the performance loss caused by input image limitation and dimension reduction, we compare the performance of this approach against the above mentioned models on the pattern attribute recognition task, which is one of the above mentioned VQA questions that require fine-grained image signals. The results in Table 5 shows that after training the projection layers, the embedding concatenation performs worse than the rest models.

6.6. Interpretability

The attention pooling in the MIVC generates a weighted average of visual representations for a bag of input images where the weights are parameterized by the neural network

Figure 4. Interpretability: four images of the same product are fed to the VQA task to identify the pattern of the product. The weights from left to right are [0.24, 0.05, 0.65, 0.06] indicating that the region image contributes more to the pattern recognition.

Image Pooling	Text Retrieval R@1
Single (zs)	79.1%
Concat (zs)	77.4%
Single	76.0%
Concat	76.8%
MIVC-Avg	76.7%
MIVC-Max	77.6%
MIVC-Attn	81.7%
MIVC-gated	80.2%

Table 4. Image captioning performance evaluation. We illustrate the effectiveness of MIVC with T5-XXL and ViT as language model and the vision encoder, respectively. We first evaluate them in zero-shot (zs) fashion and then fine-tune the model with and without various MIVC pooling. We report the top 1 recall of retrieving the generated captions among manual annotated captions given the image.

Pooling	Accuracy	Precision	Recall
single image	51.8%	58.5%	54.6%
concate image	53.4%	59.5%	55.4%
concate embed	49.6%	51.3%	49.8%
MIVC-Avg	51.2%	53.1%	51.6%
MIVC-Max	53.5%	54.8%	53.9%
MIVC-Attn	68.4%	71.0%	69.9%
MIVC-gated	64.3%	66.4%	63.7%

Table 5. Ablation regarding the embedding concatenation. After training, the model with concatenated image embeddings perform worse than the rest model on one of the VQA tasks: product pattern recognition.

that are learned during training. These weights provide insights on which image contributes the most to the downstream tasks, as illustrated in 4. In the example, the rug contains 4 images, where the first one is the rug in the live scene background and the third is enlarged local pattern details. The attention based pooling method learns to mainly focus on pattern details in the third image to infer the generative task. The lower weights of the second the the last images may cause by the fact that the second image is distorted and vague while the last image is too detail to contain useful information. The first image may provide additional context or usage information of the product that could be learned from the live scene background.

7. Conclusion and Future Work

In this paper, we propose MIVC, a multiple instance visual component to address the challenge where the visual representation of one entity should be inferred from multiple images. We show that MIVC outperforms the vanilla alternatives on the e-comemrce dataset where each product is presented by multiple images. We also explore various approaches to pool the image representations. This component is compatible with a wide range of vision-language models besides Flan-T5 model and Qformer used in this paper, which could be explored in the future. The attention-based pooling could be further improved by cross modality attention which could be explored in the future.

References

[1] 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, Dec 2022. 3, 4

[2] 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, 2

[3] 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. 2

[4] 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

[5] 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

[6] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 4

[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. 2, 6

[8] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1, 4, 6

[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. 2, 4

[10] Ji Feng and Zhi-Hua Zhou. Deep miml network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 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, 2

[12] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020. 2

[13] 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

[14] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 2, 3, 4, 5

[15] 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

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

[17] 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. 2, 3

[18] 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, 3, 4

[19] Qi Li and Roberto Aguilera. Unsupervised statistical learning with integrated pattern-based geostatistical simulation. In *SPE Western Regional Meeting*, page D041S008R005, 04 2018. 4

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1, 4, 6

[21] 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

[22] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 6

[23] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models, 2023. 6

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

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[26] 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. 6

[27] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 2

[28] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2

[29] 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

[30] 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

[31] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[32] 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. 2

[33] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, 2018. 3

[34] Wenyi Wu, Karim Bouyarmane, and Ismail Tutar. Catalog phrase grounding (cpg): Grounding of product textual attributes in product images for e-commerce vision-language applications. 2022. 2

[35] 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

[36] 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

[37] 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. 3

[38] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5014–5022, 2016. 1, 2

[39] 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