X-Former: Unifying Contrastive and Reconstruction Learning for MLLMs

Sirnam Swetha[1], Jinyu Yang[2], Tal Neiman[2], Mamshad Nayeem Rizve[2], Son Tran[2], Benjamin Yao[2], Trishul Chilimbi[2], and Mubarak Shah[1,2]

[1] Center for Research in Computer Vision, University of Central Florida
swetha.sirnam@ucf.edu, shah@crcv.ucf.edu
[2] Amazon
{viyjy, taneiman, mnrizve, sontran, benjamy, trishulc}@amazon.com

Abstract. Recent advancements in Multimodal Large Language Models (MLLMs) have revolutionized the field of vision-language understanding by integrating visual perception capabilities into Large Language Models (LLMs). The prevailing trend in this field involves the utilization of a vision encoder derived from vision-language contrastive learning (CL), showing expertise in capturing overall representations while facing difficulties in capturing detailed local patterns. In this work, we focus on enhancing the visual representations for MLLMs by combining high-frequency and detailed visual representations, obtained through masked image modeling (MIM), with semantically-enriched low-frequency representations captured by CL. To achieve this goal, we introduce X-Former which is a lightweight transformer module designed to exploit the complementary strengths of CL and MIM through an innovative interaction mechanism. Specifically, X-Former first bootstraps vision-language representation learning and multimodal-to-multimodal generative learning from two frozen vision encoders, i.e., CLIP-ViT (CL-based) and MAE-ViT (MIM-based). It further bootstraps vision-to-language generative learning from a frozen LLM to ensure visual features from X-Former can be interpreted by the LLM. To demonstrate the effectiveness of our approach, we assess its performance on tasks demanding detailed visual understanding. Extensive evaluations indicate that X-Former excels in visual reasoning tasks involving both structural and semantic categories in the GQA dataset. Assessment on fine-grained visual perception benchmark further confirms its superior capabilities in visual understanding.

Keywords: Multi-Modal Learning · Masked Image Modeling · MLLMs

1 Introduction

Recently, Large Language Models (LLMs) have demonstrated remarkable success in diverse natural language tasks [3, 38], prompting researchers to explore the integration of visual understanding capabilities into these models, leading to multimodal LLMs (MLLMs). MLLMs aim to leverage the vast knowledge

contained within off-the-shelf LLMs and vision encoders to tackle complex visual understanding tasks, thereby opening up new possibilities in the domain of vision-language understanding. Flamingo [2] is one of the early MLLMs to align frozen visual encoders to LLMs, where it introduces a Perceiver Resampler module to extract a fixed set of features from image by optimizing image-to-text generation loss, in order to bridge the modality gap. Improving upon Flamingo, BLIP-2 [23] proposed a Querying Transformer (Q-Former) that performs vision-language alignment through cross modality fusion by employing both discriminative (contrastive & classification) and generative (image-to-text generation) losses to extract a fixed set of most useful visual features for LLM. Other concurrent works [26, 43] have explored different strategies to align visual representations with LLM input space for improving vision-language understanding.

It is noteworthy that all aforementioned MLLMs employ CLIP-ViT [32] as the vision encoder, hence, inherit its limitations including: (i) poor fine-grained vision-language alignment [28], and (ii) spatially-invariant global representations [30]. As a consequence, these models struggle to encode detailed visual nuances, including object orientation, structural intricacies, spatial relationships, and multiple object instances [34], thereby hindering the ability of LLMs to comprehend local visual patterns. To alleviate this issue, there has been growing interest to learn better visual representations for MLLMs. For instance, Shikra [6] proposes to learn visual grounding for objects by adding spatial coordinates in natural language for LLM. However, this requires high-quality curated data with bounding box annotations referring to the objects in the image.

GVT [36] on the other hand distills features from pre-trained CLIP [32] via L_1 loss and uses the distilled model as the image encoder for extracting visual tokens. However, this approach relies on instruction tuning utilizing LLaVA-150k [26] dataset. Most recently, MMVP [34] proposes to leverage self-supervised pre-trained vision encoder along with CLIP-ViT to learn Mixture of Features from multiple encoders in LLaVA framework with *LLM fine-tuning*. However, they do instruction tuning with LLaVA-150k [26] dataset. Therefore, its not clear whether such an approach can work on commonly available image-text data without relying on instruction tuning using curated datasets. An additional avenue of exploration involves constructing a self-supervised vision encoder capable of capturing both global, semantically enriched, and local, detailed visual features. The central concept involves linearly combining the training objectives of CL [32] and MIM [13]. This is motivated by the fact that MIM can effectively capture local and high-frequency representations, complementing the global and low-frequency representations captured by CL. However, this hasn't been explored for vision-laguage understanding and is also the focus of this work.

In this paper, we present X-Former, a lightweight transformer module designed to achieve effective vision-language alignment from both a global and local perspective. Particularly, X-Former adopts a two-stage training approach. The first stage involves vision-language representation learning and multimodal-to-multimodal generative learning by leveraging two frozen image encoders. Specifically, X-Former utilizes learnable query vectors to extract visual features by

utilizing both CLIP-ViT [32] and MAE-ViT [13] encoders as well as employ a dual cross-attention module to dynamically fuse the extracted features. Aimed at image reconstruction and text generation, X-Former is incentivized to extract visual features covering both low frequency and high frequency.

Our main technical contributions can be summarized as:

– We propose to leverage vision encoders from CL [32] and MIM [13] to capture both global and local visual representations from frozen image encoders to improve vision-language understanding.
– We introduce X-Former with dual cross-attention to bootstrap multimodal-to-multimodal generative learning using image-text pairs, entirely without the need for curated or visual instruction data.

Empirical studies showcase the notable enhancement of our model in fine-grained visual perception tasks that demand a nuanced understanding of visual details. Specifically, in object counting tasks, X-Former demonstrates substantial improvement over BLIP-2 [23] (39.64 vs. 34.3 on COCO and 27.24 vs. 18.9 on VCR). Further, we perform fine-grained analysis comparing the image-text queries of our model and BLIP-2 to demonstrate our approach learns more diverse queries over BLIP-2 indicating the ability to capture detailed visual features. It's worth noting that BLIP-2 is pre-trained on a dataset of 129 Million image-text pairs, approximately $10\times$ larger than the dataset used for training X-Former (14 Million). This underscores the effectiveness and efficiency of our approach.

2   Method

In this section, we first briefly recapitulate the preliminaries of Q-Former [23]. Following this, we embark on early endeavors aimed at enhancing its visual learning capabilities by leveraging off-the-shelf vision encoders, namely CLIP-ViT and MAE-ViT. Specifically, CLIP-ViT is pre-trained through vision-language contrastive learning strategies, whereas MAE-ViT is trained through masked image modeling mechanisms. Our empirical studies reveal that naively combining these two encoders fails to yield significant performance improvements, especially in tasks necessitating detailed visual comprehension. To mitigate this limitation, we introduce a lightweight transformer module, dubbed X-Former, which extends Q-Former to encapsulate both global and local information.

2.1   Preliminaries of Q-Former

Q-Former is introduced in BLIP2 [23] as a solution designed to bridge the gap between a frozen CLIP-ViT and a frozen LLM (Figure 1 (a)). Given a collection of image-text pairs $\{(I_k, T_k)\}_{k=1}^{N}$, Q-Former operates by taking a predetermined number of learnable query embeddings z, T_k, and C as input, where C indicates CLIP image features of I_k. These queries engage in mutual interaction through self-attention layers and interact with frozen image features C through cross-attention layers in every alternate layer as shown in Figure 1(a) L1. The resulting
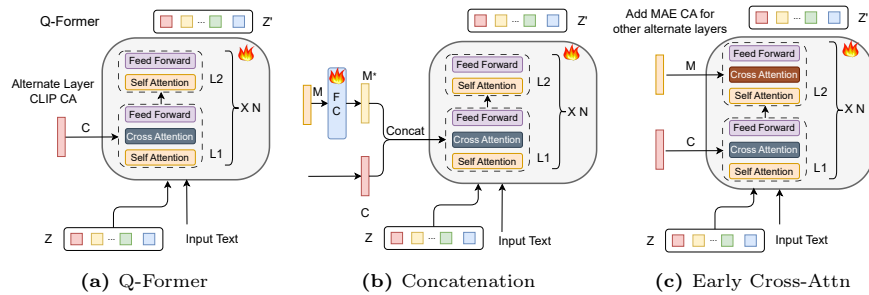
Fig. 1: **(a)** Vanilla Q-Former extracts a fixed number of output features Z' from the CLIP image encoder, where C and z denotes CLIP-ViT's image features and the query input, respectively; **(b)** Concatenated MAE-ViT (M^*) and CLIP-ViT (C) features are passed as input to Q-Former, **(c)** A Cross-Attention layer is added in L2 to enable MAE-ViT interaction in Q-Former.

query representation is denoted by Z', which is anticipated to encapsulate visual information derived from the frozen CLIP-ViT.

Though Q-Former has exhibited remarkable performance on various downstream tasks like VQA and image captioning, it encounters challenges in detailed visual feature comprehension. This limitation primarily stems from the training objective of CLIP, which incentivizes ViT to prioritize low-frequency signals and global visual patterns [30]. Fortunately, MAE-ViT [13], trained to reconstruct masked image patches, excels in understanding detailed visual features. However, the integration of CLIP-ViT and MAE-ViT in multimodal understanding remains unclear, given their inherently divergent perspectives when 'viewing' images. To address this inquiry, we embark on early attempts to combine CLIP-ViT and MAE-ViT in a straightforward manner as discussed below.

2.2   Simple Combinations of CLIP-ViT and MAE-ViT

Visual Feature Concatenation As shown in Figure 1 (b), our first attempt is to concatenate the frozen image features from CLIP-ViT and MAE-ViT, which are denoted by C and M, respectively. To accommodate the discrepancy between C and M, a linear layer is applied to align M with C, resulting in M^*, which is subsequently concatenated with C. This combined feature (C, M^*) serves as input to the Q-Former, which undergoes training in both stages following the methodology outlined in [23]. Our experiments show that the simple concatenation approach performs on par with BLIP-2, as illustrated in Figure 2. This observation highlights the non-trivial nature of integrating C and M to leverage their complementary strengths. The distinct information provided by MAE and CLIP presents challenges for the model in simultaneously learning both global and local information while preserving visual-text coherence. Moreover, it is crucial to note that introducing additional vision encoders does not necessarily guarantee improved performance.
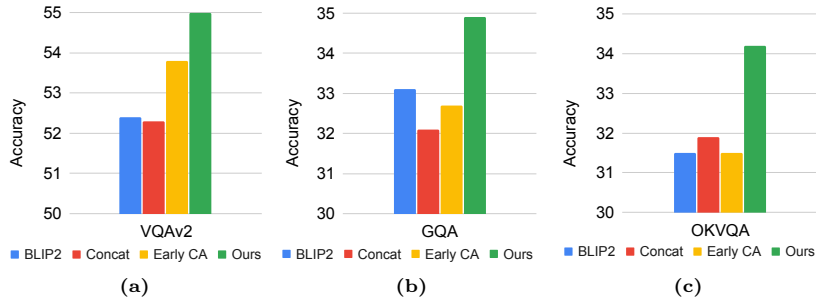
Fig. 2: Performance comparison of BLIP2, BLIP2+Concatenation, BLIP2+Early Cross-Attention, and our method on VQAv2 (a), GQA (b), and OKVQA (c) datasets.

Early Cross-Attention Inspired by the observations from the concatenation strategy outlined earlier, we delve into early interactions akin to CLIP-style cross-attention within Q-Former. To pursue this, we introduce early cross-attention by integrating new cross-attention layers, alternating with non-CLIP interaction layers, as depicted in Figure 1 (c). While this approach modestly improves performance compared to the concatenation strategy (see Figure 2), it notably escalates the number of parameters in Q-Former, resulting in a total of 183M trainable parameters (approximately 75M more than BLIP-2). Importantly, increasing parameters doesn't inherently enhance performance. While enhancements are observed for the VQAv2 dataset, there's a decline in performance for the GQA dataset and comparable results for the OKVQA dataset against BLIP-2. To mitigate this and facilitate the extraction of local information from MAE, we advocate for incorporating late-interaction for the Masked Image Modeling (MIM) objective during training.

2.3   X-Former Overview

In Figure 3, we present an overview of our method, comprising two frozen image encoders (CLIP-ViT and MAE-ViT), a frozen image decoder, and a trainable X-Former aimed at bridging the modality gap and extracting interpretable visual features for the LLM. For MAE-ViT, random masking of patches in the input image is performed. X-Former processes a set of learnable queries Z along with the input text T_k and the image features (C, M) as input. Our model extends the framework of BLIP2 by incorporating Image-Text Matching (ITM), Image-Text Contrastive (ITC), and Image-Text Generation (ITG) losses, while also introducing a reconstruction loss for the image decoder.

X-Former To address the limitations of Q-Former, primarily its lack of fine-grained alignment and its focus on capturing global information, we propose integrating MAE features (M) into our X-Former module, depicted as an orange block in Figure 3. This addition facilitates the extraction of both local and global information, optimizing image reconstruction alongside the ITC, ITM,
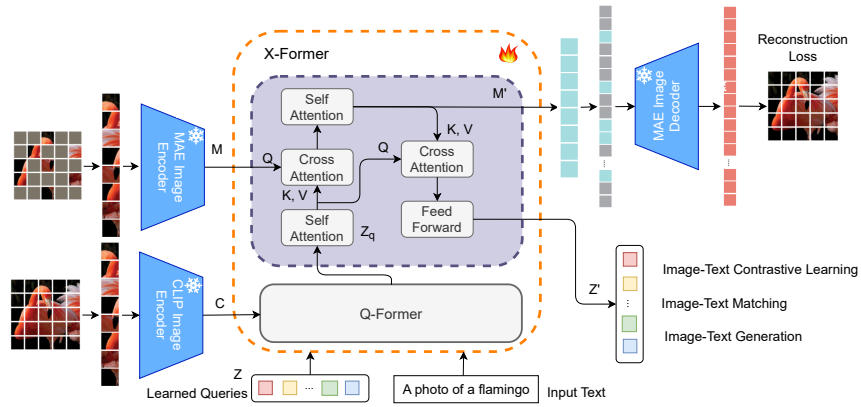
Fig. 3: An overview of X-Former which extends Q-Former by introducing a dual cross-attention module to capture both local and global visual features. First it computes CLIP visual features (C) and MAE features (M) (with random masking) from the input image-text pair. Q-Former employs C, Z, Text to generate output queries optimized for three objectives - ITC, ITM and ITG. The proposed block (purple) enriches Q-Former global representation (Z_q) with local information from MAE features (M). Initially, M is aligned and enriched by Z_q resulting in enriched MAE representation (M'), optimized for image reconstruction. Then, M' enhances Z_q with local representations through cross-attentions, optimized using VL objectives. Jointly optimizing these four objectives facilitates the learning of both global and local representations.

and ITG objectives, represented by the purple block in Figure 3. The first cross-attention block employs MAE features (M) as queries and Q-Former output (Z_q) as keys and values to align and enhance M by integrating global semantic information from Q-Former, resulting in enriched MAE features (M'). Subsequently, these enriched MAE features enhance the Q-Former output (Z_q) to Z' by integrating both global and local information through cross-attention, as depicted. The enhanced queries (Z') are optimized for ITC, ITM, and ITG, along with a reconstruction objective applied to M'. Finally, M' is passed to the frozen MAE decoder to reconstruct the masked patches.

Stage 1: Pre-Training During the pre-training stage, the X-Former learns to extract both local and global representation by optimizing Reconstruction, ITC, ITM and ITG losses. The reconstruction loss together with the image-text alignment objectives enforces to align and capture local representation, while the VL objectives align it with text representation. The incorporation of MAE and CLIP features ensures that the queries extract a enhanced visual representation that is aligned with the accompanying text. We follow BLIP-2 [23] for computing ITC, ITM and ITG losses. For ITC, we compute similarity between [CLS] token of the text-embedding and each of the final output query embeddings Z', selecting the *highest* as the image-text similarity. For this objective, to prevent data leak a unimodal self-attention mask is employed, ensuring that the queries
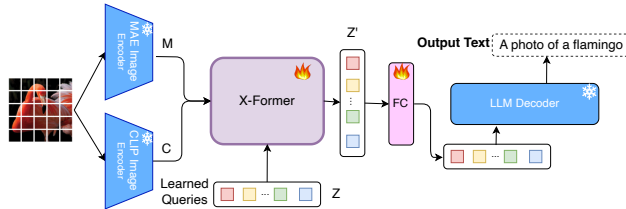
Fig. 4: LLM Alignment. X-Former queries are aligned with a *frozen* decoder-based LLM. FC layer adapts the query output(Z') to LLM embedding space.

and text do not interact with each other. It maximizes the image-text similarity of positive pairs by contrasting with in-batch negatives.

For ITM, the model is asked to predict whether image-text pair match (positive) or not (negative). Here, a bi-directional self-attention mask is employed, allowing all queries and texts to attend to each other. Consequently, the output query embeddings capture multimodal information, which is then fed to a two-class linear classifier to obtain logits. These logits are averaged across all the queries to compute the final matching score. To generate negative pairs, a hard negative mining strategy [24] is employed. In the context of ITG, X-Former utilizes an input image as a condition to generate text. A multimodal causal self-attention mask is used, allowing queries to attend to each other while excluding text tokens, and enabling text tokens to attend to all queries and previous text tokens. The [CLS] token is substituted with the [DEC] token as the first text token, serving as an indicator for the decoding task.

Stage 2: LLM Alignment During pre-training, the X-Former acquires the ability to extract information from both MAE and CLIP, resulting in queries that capture a blend of global and local information. Subsequently, we align the features of the X-Former with the frozen LLM, aiming to harness the comprehensive visual representations acquired by the X-Former module and integrate them with the robust language generation capabilities of the LLM. This integration involves connecting the pre-trained X-Former output (Z') to the LLM via a single fully-connected layer, aligning it with the LLM representation space, as depicted in Figure 4. Specifically, we experiment with the OPT model, which is a decoder-based LLM, and train it using a language modeling loss keeping *both* image encoders and LLM *frozen*.

3 Experiments

Pre-Trained Models We employ pre-trained ViT-G model from EVA-CLIP [10] as CLIP-ViT. For MAE, we utilize the pre-trained ViT-H model [13]. Our choice

Table 1: Zero-shot Visual Question Answering results on the VQAv2 dataset. Note that * indicates the result is obtained using the official checkpoint.

Method	#Trainable Params	Data	Overall	Other	Yes/No	Number
				VQAv2 Accuracy		
			Open-ended generation models			
FewVLM [19]	740M	9.1M	47.7	-	-	-
Frozen [35]	40M		29.5	-	-	-
VLKD [7]	406M	3.7M	42.6	-	-	-
BLIP-2 $OPT_{6.7B}$* [23]	108M	129M	55.1	47.3	72.6	34.6
BLIP-2 $OPT_{2.7B}$ [23]	107M	14M	49.9	39.3	71.5	27.3
X-Former (Ours) OPT$_{2.7B}$	129M	14M	**51.3**	**41.5**	71.2	**30.9**
BLIP-2 $OPT_{6.7B}$ [23]	108M	14M	52.4	43.6	71.5	30.8
X-Former (Ours) OPT$_{6.7B}$	130M	14M	**55.0**	**45.6**	**73.3**	**37.8**

for the LLM involves the OPT model [42]. Our model undergoes pre-training for nine epochs in Stage-1 and one epoch in Stage-2, with OPT employed for Stage-2 alignment. See Supplementary Section 1 for implementation details.

Datasets and Tasks To demonstrate the effectiveness of our approach, we leverage a standard dataset comprising 14M Image-Text pairs sourced from COCO [25], Visual Genome [21], SBU [29], CC3M [33], and CC12M [4] for model pre-training. Our evaluation spans across various benchmarks, including COCO [25], NoCaps [1], VQAv2 [12], GQA [15], OK-VQA [27], Flickr30k [31], and VCR [41]. Furthermore, we employ a fine-grained visual perception benchmark [36], featuring Object Counting (OC) and Multi-Class Identification (MCI) tasks, to assess the model's fine-grained visual understanding capabilities.
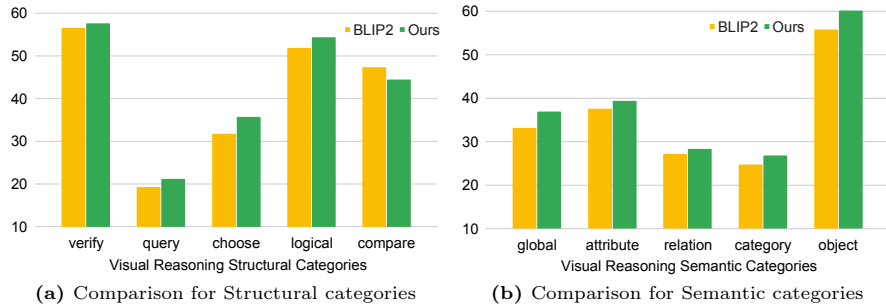
3.1   Experimental Results

Zero-Shot Visual Question Answering First, we present the results for zero-shot visual question answering on the VQAv2-val dataset, which encompasses three question types: open-ended (other), Yes/No, and Number questions, as illustrated in Table 1. We utilize the prompt "Question: Short Answer:" for the generation process, employing beam search with a beam width of 5. We set the length-penalty to 0 to encourage short answers. Our results indicate that our approach surpasses BLIP-2 for both $OPT_{2.7B}$ and $OPT_{6.7B}$ LLMs by 1.4% and 2.6% respectively, highlighting superior visual comprehension. Particularly noteworthy are the significant enhancements observed for the Number task, which demands precise local understanding for object counting or identification. Fine-tuning results are reported in Supplementary Section 2, while large-scale experimental findings are detailed in Supplementary Section 3.

In Table 2, we report zero-shot visual question answering results for the GQA test-dev dataset. The results demonstrate the superior performance of our method over BLIP-2. Furthermore, we conducted a comprehensive comparison to demonstrate the effectiveness of our approach across both structural and semantic categories in the GQA dataset. The structural category encompasses five question types (verify, open-ended query questions, choose from options, logical

Table 2: Zero-shot Visual Question Answering Results on GQA and OKVQA datasets. Note that * indicates the result is obtained using the official checkpoint.

Method	Data	GQA	OKVQA
FewVLM [19]	9.1M	29.3	16.5
Frozen [35]		-	5.9
VLKD [7]	3.7M	-	13.3
Flamingo3B [2]	>2B	-	41.2
Flamingo9B [2]	>2B	-	44.7
Flamingo80B [2]	>2B	-	50.6
BLIP-2 $OPT_{6.7B}$* [23]	129M	34.2	35.3
BLIP-2 $OPT_{2.7B}$ [23]	14M	33.6	24.2
X-Former (Ours) OPT$_{2.7B}$	14M	**34.1**	**27.7**
BLIP-2 $OPT_{6.7B}$ [23]	14M	33.1	31.5
X-Former (Ours) OPT$_{6.7B}$	14M	**34.9**	**34.2**

(a) Comparison for Structural categories   **(b)** Comparison for Semantic categories

Fig. 5: Detailed Comparison for both Structural and Semantic categories in GQA.

inference, and object comparison) as depicted in Figure 5. Our results indicate that we outperform in the majority of these categories. In Figure 5, we provide a comparison for the semantic categories, which include questions related to object existence, object attributes, object category, global scene, and object relationships. Across all these categories which includes both global and local reasoning, our approach consistently demonstrates better performance, highlighting its detailed visual understanding capabilities.

We report zero-shot visual question answering performance on OKVQA test dataset in Table 2. This dataset poses a significant challenge as it requires methods to draw upon external knowledge to answer questions effectively. Our method demonstrates a significant improvement in accuracy over BLIP-2, achieving a 2.7% and 3.5% gain with $OPT_{6.7B}$ and $OPT_{2.7B}$ LLM respectively. This signifies the robustness of our approach in accurately aligning visual information with LLM and effectively leveraging external knowledge to answer the questions.

Fine-Grained Visual Perception Evaluation To demonstrate that our approach has better visual understanding, we evaluate perception abilities at fine-

Table 3: Zero-shot Fine-Grained Visual Perception evaluation of MLLMs on Object Counting (OC) & Multi-class Identification (MCI) tasks. For fair comparison, we compare with models trained only on image-text data. *evaluated using official checkpoint.

		OC		MCI	
Method	Data	COCO	VCR	COCO	VCR
BLIP-2* [23]	129M	34.3	18.9	69.44	74.16
BLIP-2 [23]	14M	25.88	21.12	61.5	65.3
X-Former (Ours)	14M	**39.64**	**27.24**	**69.44**	**69.28**

Table 4: Zero-shot Image Captioning Results on COCO & NoCaps without fine-tuning for captioning task. B:BLEU, C: CIDEr, S: SPICE. *evaluated using official checkpoint

		COCO			NoCaps	
Method	Data	B@4	C	S	C	S
BLIP-2* [23]	129M	39.9	134.3	24.3	113.4	15.2
BLIP-2 [23]	14M	39.2	131.0	23.7	113.1	14.9
X-Former (Ours)	14M	**39.3**	**131.1**	23.6	**113.2**	14.9

grained scale [36], we evaluate our approach for fine-grained visual perception capabilities OC and MCI task. We use the prompt "Question: {} Short Answer:". to evaluate for this task. For generation, we use beam search with a beam width of 5. We also set the length-penalty to 0 to encourage shorter answers. The questions for object counting tasks is of the form "How many {$objects$} are there in the image?" and for multi-class identification task it is "Does {$objects$} exist in the image?". For fair comparison, we compare with methods that only employ image-text datasets for training. In Table 3, we show that our model outperforms BLIP-2 on both datasets i.e., COCO and VCR. It can be seen that for Object Counting task our approach improves BLIP-2 by 13% on COCO and 6.1% on VCR datasets respectively. This indicates that X-Former is able to extract detailed visual features. Please refer to Supplementary Sections 6, 7 for more fine-grained evaluations.

Zero-shot Image Captioning. In addition to the visual reasoning tasks, we report results for image captioning without fine-tuning in Table 4 for COCO and NoCaps dataset. Captioning task requires image-level semantic understanding as the annotated captions briefly describe the image. We show that our approach improves on fine-grained visual reasoning tasks without impacting the captioning performance.

3.2   Qualitative Results

To effectively demonstrate the capabilities of our model, we present qualitative results that highlight its performance in the object counting task. Accurate object counting requires a deep understanding of local contexts within an image.
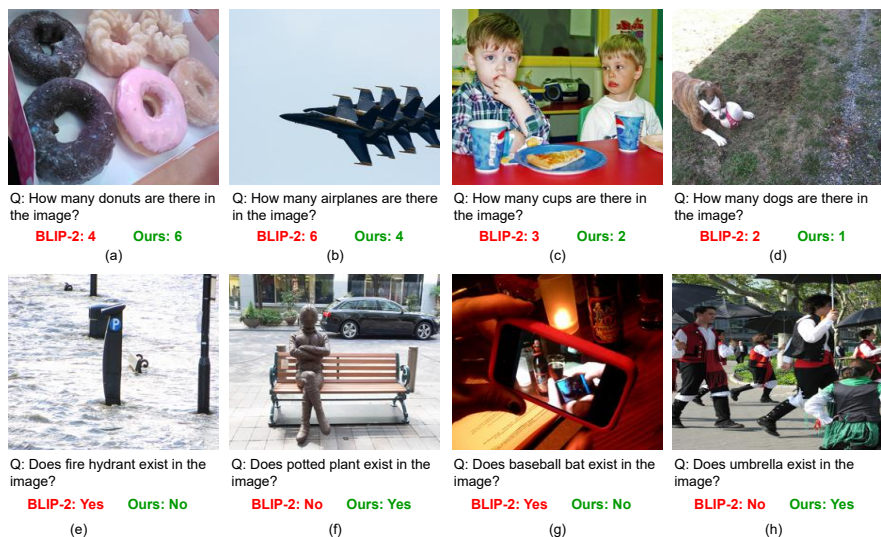
Q: How many donuts are there in the image?	Q: How many airplanes are there in the image?	Q: How many cups are there in the image?	Q: How many dogs are there in the image?
BLIP-2: 4 **Ours: 6**	**BLIP-2: 6** **Ours: 4**	**BLIP-2: 3** **Ours: 2**	**BLIP-2: 2** **Ours: 1**
(a)	(b)	(c)	(d)
Q: Does fire hydrant exist in the image?	Q: Does potted plant exist in the image?	Q: Does baseball bat exist in the image?	Q: Does umbrella exist in the image?
BLIP-2: Yes **Ours: No**	**BLIP-2: No** **Ours: Yes**	**BLIP-2: Yes** **Ours: No**	**BLIP-2: No** **Ours: Yes**
(e)	(f)	(g)	(h)

Fig. 6: Qualitative Comparison demonstrating Fine-Grained Visual Understanding in Object Counting and Multi-class Identification Tasks. Our model showcases better visual understanding by accurately counting objects and effectively identifying them without confusion based on shape or color.

As shown in Figure 6(a), our method correctly counts six donuts in an image, while BLIP-2 incorrectly predicts four. This demonstrates the model's ability to distinguish individual objects even when they are closely clustered together. Figure 6(b) presents a more challenging scenario where four airplanes are flying in close proximity. BLIP-2 struggles with this task, erroneously predicting six airplanes instead of the correct number of four. Our method, on the other hand, accurately identifies the four airplanes, showcasing its robustness in handling dense object arrangements. In Figure 6(c), we encounter two cups and a plate that share a similar color, for which BLIP-2 incorrectly predicts three cups. Our method, however, correctly identifies the two cups, demonstrating its ability to handle objects with similar visual properties. Figure 6(d) depicts a dog that blends into the background due to its similar coloring. BLIP-2 makes an incorrect prediction. Our method, in contrast, correctly identifies the dog. These qualitative results collectively demonstrate the effectiveness of our model in the object counting task, outperforming BLIP-2 in various scenarios that demand robust local understanding and the ability to handle challenging object arrangements and color similarities.

In the Multi-Class Identification task, BLIP-2's object recognition capabilities exhibit limitations when presented with Figure 6(e). BLIP-2 mistakenly interprets the shape of a parking pole as a fire hydrant. Figure 6(f) presents a challenge where a potted plant is positioned in the background, occupying a relatively small portion of the image. BLIP-2 fails to detect the presence of the

Table 5: Ablation study with early layer features from CLIP-ViT. L_i indicates i^{th} layer of CLIP-ViT.

Table 6: Ablation for Effect of Reconstruction (Recon.) Loss. Reconstruction during pre-training plays an important role in aligning the MAE to extract meaningful information.

Method	VQAv2	GQA	OKVQA
Ours	**55.0**	**34.9**	**34.2**
L_{26}	53.7	32.6	31.2
L_{28}	52.5	31.9	28.0
L_{30}	52.4	32.8	30.9

	Stage 1 Recon.	Stage 2 Recon.	VQAv2	GQA	OKVQA
			33.1	25.4	12.1
	✓	✓	52.4	32.2	29.2
Ours	✓		**55.0**	**34.9**	**34.2**

potted plant, whereas our method successfully identifies it. Figure 6(g) showcases image of a bottle whose shape closely resembles that of a baseball bat, leading BLIP-2 to identify it as a baseball bat. Figure 6(h) presents a challenge due to the subtle color of an umbrella, making it difficult task. BLIP-2 fails to recognize the object, while our method accurately identifies it as an umbrella. A comprehensive qualitative analysis is provided in Supplementary Section 5. We also present query diversity analysis for fine-grained qualitative comparison and present them in Supplementary Section 6.

3.3   Ablation Analysis

We perform the following ablations to analyze the various components of our approach. Please refer to Supplementary Section 4 for more ablation analysis.

Leveraging Early Layer CLIP features To illustrate the efficacy of MAE embeddings in learning better local representations, we conduct experiments using features from intermediate layers of CLIP. Specifically, we explore the utilization of features from layers $26, 28, 30$ as substitutes for MAE features in our proposed approach. It is important to note that for this training, there is no reconstruction loss since we are employing CLIP features. From Table 5, it can be seen that employing MAE features leads to best results. We present more results in Supplementary Section 4.

Impact of Image Reconstruction Loss in Pre-training. We examine the influence of the reconstruction loss during the pre-training stage. In this pre-training phase, we employ four objectives: image-text contrastive and matching loss (for discriminative vision-language alignment), reconstruction loss, and text generation loss. Our findings demonstrate that combining alignment and reconstruction objectives during pre-training, the image reconstruction loss becomes effective in extracting aligned and meaningful representations. To illustrate this, we conducted an experiment without the MAE decoder and reconstruction objective during pre-trainingand computed MAE features for the entire image without any masking. As shown in Table 6, row 1, the significant performance drop highlights the crucial role of the reconstruction loss in aligning and extracting useful

information from MAE. This suggests that network cannot find a shortcut thus leading to drop in performance while MIM enables extracting meaningful and aligned representations leading to best performance.

Impact of Image Reconstruction Loss in LLM Alignment We investigate the influence of reconstruction loss in stage-2. In the context of LLM alignment, the model undergoes training with language-modeling loss exclusively, without the inclusion of image-text contrastive and matching losses, and Q-Former does not receive any text-input. It is evident that combining only language modeling loss and reconstruction loss yields suboptimal results, as indicated by the performance drop shown in Table 6, row 2.

4   Related Works

Multimodal Large Language Models (MLLMs) The success of Large Language Models (LLMs) has prompted researchers to delve into the exploration of integrating visual components into these models, culminating in the development of Multimodal LLMs (MLLMs) [22, 40]. MLLMs have garnered significant traction in both academic and industrial spheres due to their remarkable proficiency in comprehension and generation. The key idea is to leverage off-the-shelf pre-trained vision encoders and LLMs and keep them frozen during the training. However, the most critical challenge in utilizing a frozen LLM lies in narrowing down the gap between visual features and the text space. Existing MLLMs can be broadly divided into three categories according to the modules/components they used for bridging the modality gap: (i) Perceiver-based [16, 17], (ii) Q-Former-based [23], and (iii) linear projection layer-based. In Perceiver-based methods such as Flamingo [2], they employ a Perceiver Resampler to produce a small fixed number of visual tokens per image, subsequently amalgamating them with text tokens as input for LLMs. In other words, the Perceiver Resampler relies on the image-to-text generative learning to bridge the modality gap. Q-Former shares the similar spirits with the Perceiver Resampler, except that Q-Former relies on an extra vision-language representation learning stage. Owing to its simplicity and efficiency, Q-Former is widely used in such as BLIP-2 [23], SEED [11], MiniGPT-4 [43], and InstructBLIP [8]. In linear projection layer-based methods, the common practice is to align visual features with text features through a singular linear layer before incorporating them into LLMs. The effectiveness of this simple strategy is evidenced by recent studies such as LLaVa [26] and FROMAGe [20]. Our work is inspird by Q-Former but with the following differences: (i) we extend Q-Former to handle two off-the-shelf vision encoders, i.e., CLIP-ViT and MAE-ViT, and (ii) we introduce multimodality-to-multimodality generative learning to further bridge the modality gap.

Self-Supervised Vision Encoders Self-supervised vision encoders (VEs) play a crucial role in MLLMs by providing visual features that are understandable by LLMs. Among them, VEs that are pre-trained by vision-language-based contrastive learning (CL) has been the most popular one, where the VE is trained

to bring representations of matched image-text pair close together and push representations of unmatched pairs apart [5, 9, 14, 18, 32, 39]. This encourages the VE to capture semantic similarities and differences in visual content. However, recent work reveals that CL mainly focuses on low frequency signals and longer-range global patterns inheriting from its training objective [30]. In consequence, CL-based MLLMs suffer from understanding detailed perceptions which are essential for tasks that require fine-grained visual understanding such as object counting. As a counterpart, masked image modeling (MIM) involves masking parts of an image and tasking the vision encoder with predicting the masked image patches [13]. This enhances the VE's ability to understand detailed visual features by promoting contextual understanding, encouraging the learning of spatial relationships, and facilitating the development of transferable representations. Inspired by these observations, recent work attempt to build VEs that is able to understand both global semantic and detailed local patterns [30,37]. The key idea is to leverage the strength of CL and MIM by linearly combining two training objectives with a shared VE. While simple and effective, these models are not readily applicable to MLLMs due to their sole pre-training on limited datasets and modest model sizes, significantly lagging behind their CL and MLM counterparts. Although scaling up data and model size is possible, it introduces substantial carbon emissions and fails to capitalize on the advantages offered by off-the-shelf VEs from both CL and MIM. In contrast, our approach incorporates a lightweight transformer that harnesses the benefits of pre-trained CL and MIM models, showcasing superior performance in fine-grained perception understanding without imposing a significant computational burden.

5   Conclusion

In this paper, we introduce X-Former, a novel architecture designed to enhance visual representations for Multimodal Language Models (MLLMs) by integrating pre-trained MAE and CLIP vision encoders. Our motivation stems from several observations: (i) existing MLLMs primarily rely on CLIP-ViT, which often fails to capture fine-grained visual signals; (ii) our empirical studies reveal that simply combining CLIP-ViT and MAE-ViT does not necessarily yield performance improvements; and (iii) the efficacy of MLLMs heavily depends on large-scale image-text pairs for pre-training and meticulously curated instruction tuning datasets for fine-tuning. X-Former effectively tackles these limitations by integrating CLIP-ViT and MAE-ViT through a dual cross-attention mechanism, all while keeping computational demands manageable. Our approach is plug-and-play and can be applied to other models. Our experimental results unequivocally show that X-Former surpasses BLIP-2 in a variety of visual reasoning tasks requiring robust visual comprehension. Remarkably, these superior results are achieved using only one-tenth of the image-text pair dataset, without the need for any instruction tuning datasets.

References

1. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: Nocaps: Novel object captioning at scale. In: Proceedings of the IEEE/CVF international conference on computer vision (2019)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems (2022)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems (2020)
4. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
5. Chen, C., Zhang, J., Xu, Y., Chen, L., Duan, J., Chen, Y., Tran, S., Zeng, B., Chilimbi, T.: Why do we need large batchsizes in contrastive learning? a gradient-bias perspective. Advances in Neural Information Processing Systems (2022)
6. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic (2023)
7. Dai, W., Hou, L., Shang, L., Jiang, X., Liu, Q., Fung, P.: Enabling multimodal generation on CLIP via vision-language knowledge distillation. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Findings of the Association for Computational Linguistics: ACL 2022 (2022)
8. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023)
9. Duan, J., Chen, L., Tran, S., Yang, J., Xu, Y., Zeng, B., Chilimbi, T.: Multimodal alignment using representation codebook. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
10. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva: Exploring the limits of masked visual representation learning at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
11. Ge, Y., Ge, Y., Zeng, Z., Wang, X., Shan, Y.: Planting a seed of vision in large language model (2023)
12. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: 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 (2017)
13. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2022)
14. Hu, Z., Zhu, X., Tran, S., Vidal, R., Dhua, A.: Provla: Compositional image search with progressive vision-language alignment and multimodal fusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
15. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2019)
16. Jaegle, A., Borgeaud, S., Alayrac, J.B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., et al.: Perceiver io: A general architecture for structured inputs & outputs. arXiv preprint arXiv:2107.14795 (2021)

17. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: International conference on machine learning. PMLR (2021)
18. Jiang, Q., Chen, C., Zhao, H., Chen, L., Ping, Q., Tran, S.D., Xu, Y., Zeng, B., Chilimbi, T.: Understanding and constructing latent modality structures in multimodal representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
19. Jin, W., Cheng, Y., Shen, Y., Chen, W., Ren, X.: A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (2022)
20. Koh, J.Y., Salakhutdinov, R., Fried, D.: Grounding language models to images for multimodal inputs and outputs. International Conference on Machine Learning (2023)
21. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision (2017)
22. Li, C.: Large multimodal models: Notes on cvpr 2023 tutorial. arXiv preprint arXiv:2306.14895 (2023)
23. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: bootstrapping language-image pretraining with frozen image encoders and large language models. In: ICML (2023)
24. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. PMLR (2022)
25. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer (2014)
26. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
27. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition (2019)
28. Mukhoti, J., Lin, T.Y., Poursaeed, O., Wang, R., Shah, A., Torr, P.H., Lim, S.N.: Open vocabulary semantic segmentation with patch aligned contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
29. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. Advances in neural information processing systems (2011)
30. Park, N., Kim, W., Heo, B., Kim, T., Yun, S.: What do self-supervised vision transformers learn? arXiv preprint arXiv:2305.00729 (2023)
31. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision (2015)
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR (2021)

33. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2018)
34. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms (2024)
35. Tsimpoukelli, M., Menick, J.L., Cabi, S., Eslami, S.M.A., Vinyals, O., Hill, F.: Multimodal few-shot learning with frozen language models. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems (2021)
36. Wang, G., Ge, Y., Ding, X., Kankanhalli, M., Shan, Y.: What makes for good visual tokenizers for large language models? (2023)
37. Weers, F., Shankar, V., Katharopoulos, A., Yang, Y., Gunter, T.: Masked autoencoding does not help natural language supervision at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2023)
38. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al.: Emergent abilities of large language models. arXiv preprint arXiv:2206.07682 (2022)
39. Yang, J., Duan, J., Tran, S., Xu, Y., Chanda, S., Chen, L., Zeng, B., Chilimbi, T., Huang, J.: Vision-language pre-training with triple contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
40. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. arXiv preprint arXiv:2306.13549 (2023)
41. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
42. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)
43. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)