Supplementary: 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

The supplementary material is organized as follows. First, we provide additional details in Section 1 along with computational cost, zero-shot retrieval and parameter comparison in Section 1.1, 1.2, 1.3 respectively.

In Section 2, we present the fine-tuning experiments followed by the Large-scale experimental results in Section 3. We report additional ablation analysis in Section 4, and present additional qualitative results in Section 5. Finally, we present fine-grained query analysis in Section 6 and additional fine-grained evaluations in Section 7.

1 Details

Implementation Details Our model undergoes pre-training for nine epochs in Stage-1 and one epoch in Stage-2, with OPT employed for Stage-2 alignment. Consistent with [7], we opt to utilize the output features from the second-to-last layer of CLIP-ViT and the parameters of the frozen ViTs and LLMs are converted into FP16. The AdamW optimizer with $[\beta_1 = 0.9, \beta_2 = 0.98]$ and a weight decay of 0.05 is used. We use cosine learning rate decay with a peak learning rate of 1e-4 and a linear warmup of 2k steps. Images are resized to 224×224 with random resized crop and horizontal flip augmentations applied. The masking ratio for MAE-ViT is set to 50%. During Stage-2 training, the minimum learning rate is maintained at 5e-5.

Note that BLIP-2 does not release the captions utilized for training. Therefore we use the WebCapFilt captions from BLIP [8] for the LAION115M, SBU, and Conceptual Captions datasets, each of which contains one synthetic caption per image. For BLIP-2 training, the authors perform further processing using CapFilt method to generate 10 synthetic captions per image. These captions, along with the original caption, are ranked using CLIP ViT-L/14 image-text similarity, and the top 2 captions per image are retained as training data. During training, one caption is randomly selected.

For fair comparison, we train BLIP-2 and our model on same training dataset and report results. Additionally, we present results with $FlanT5_{XL}$ [3] model, which is trained with a prefix language modeling loss as mentioned in BLIP-2. For further details, please refer to Section 3.

For a fair comparison with official BLIP-2 model, we take the official BLIP-2 checkpoint and use the provided evaluation script to report results as highlighted by * in *all* the tables.

Table 1: Computational Cost.

Method	Train	GPU Mem		#FLOPS		Inference
	time	S1	S2	S1	S2	time
BLIP-2	39 hrs	3.3G	21G	3.08T	17.12T	∼680 ms
X-Former (Ours)	43 hrs	4.6G	22G	3.16T	17.2T	∼890 ms

1.1 Computational Cost

Here, we discuss the training and GPU memory usage of our model compared to BLIP-2. Our method uses 4.7% more GPU memory than BLIP-2 with 10% higher train time for $OPT_{6.7B}$ model (Table 1). Note that vision encoders have much less params compared to LLMs, hence adding a vision encoder does not add much overhead. We also present detailed comparison between BLIP-2 and our model for both stage 1 and stage 2 in Table 1.

Table 2: Zero Shot Retrieval Flickr

Method	TR@1	TR@5	IR@1	IR@5
BLIP-2	89	98.3	**83.5**	96.2
X-Former (Ours)	**91.4**	**98.7**	83.3	**96.3**

1.2 Zero-Shot Retrieval Results

image-text retrieval on Flickr dataset. Note that we use the pre-trained stage one model without any fine-tuning and compare with BLIP-2. As shown, our method improves retrieval scores over BLIP-2.

1.3 Detailed Parameter Comparison

In Table 3, we provide detailed comparison with other variants as discussed (see Section 2.2 in main text) to fuse MAE and CLIP features. Specifically, we provide number of trainable parameters along with performance on three VQA benchmarks namely VQAv2, GQA and OKVQA. We show our approach achieves best performance with a gain of 2.8% (Concat), 2.2%(Early CA) on GQA. In terms of parameters, Early CA adds 75M more params than BLIP-2; 53M more params than ours (Table 3 row 3). Despite having more params the performance is inferior to ours by margins of 1.2% (VQA), 2.2%(GQA), 2.7% (OKVQA).

Table 3: Detailed Comparison. CA, C, M denote Cross-Attention, CLIP and MAE respectively

Method	Model Details	Input	# Trainable Params	VQAv2	GQA	OKVQA
BLIP-2	Q-Former	C	108M	52.4	33.1	31.5
Concat	Q-Former	M, C	110M	52.3	32.1	31.9
Early CA	Q-Former (M - CA)	M, C	183M	53.8	32.7	31.5
X-Former (Ours)	X-Former	M, C	130M	**55.0**	**34.9**	**34.2**

2 VQA Fine-tuning

We have implemented the fine-tuning code for VQA task, as the authors have not released the code for this task. For Visual Question Answering fine-tuning task, we utilize VQA train and val splits along with Visual Genome train dataset following [7, 8].

Note that the VQAv2 dataset contains multiple answer annotations per question. For our fine-tuning experiments, we randomly select one of the answers as the output for the VQA dataset. As mentioned in BLIP-2 [7], we feed the question as input to the X-Former along with the image embeddings for our model.

It is important to highlight that BLIP-2 completely unfreezes the CLIP ViT, resulting in a total of 1.2 Billion trainable parameters during fine-tuning. However, due to resource constraints, we are unable to train such large models. Therefore, we only unfreeze the layer norms in CLIP ViT while keeping the MAE ViT completely frozen, resulting in a total of 216 Million trainable parameters, which is 6× lower compared to full fine-tuning. We set the image size to 490 and the learning rate to 1e-5. During generation, we utilize the prompt "Question: Short Answer:" and set the beam size to 5 for beam search. Upon fine-tuning on the VQA task, our method achieves a 4.4% improvement over the zero-shot performance for GQA dataset, which requires detailed visual understanding for accurate answers.

3 Large Scale Results

Here, we present the results for large-scale training to illustrate that our model performance scales with data size. We utilize, LAION115M dataset in addition to the 14M dataset used in the paper. Note that, the downloaded dataset using web urls has an approximate 20% miss rate, leading to overall dataset size of $105M$ for large-scale training.

We employ the synthetic captions released by BLIP [8] for LAION115M, SBU and Conceptual Caption datasets. We use the Stage-1 model pre-trained with 14M dataset as weight initialization. The LLM alignment stage is trained for an epoch on the large-scale dataset. We experiment with $OPT_{2.7B}$ LLM and $FlanT5_{XL}$ [3] LLM and present Zero-shot visual question answering results on GQA and OKVQA datasets. OPT is a decoder-only LLM while $FlanT5$ is enocder-decoder LLM. Following BLIP-2 [7], for OPT model we train with

Table 4: Zero-shot Visual Question Answering results on GQA and OKVQA datasets with Large scale training. * evaluated using official checkpoint

Method	Caption Pre-processing		Stage1 Data	Stage2 Data	GQA Acc.	OKVQA Acc.
	CLIP Caption Ranking	#Caps/Img				
Frozen [11]					-	5.9
VLKD [4]					-	13.3
FewVLM [6]					29.3	16.5
Flamingo3B [1]					-	41.2
Flamingo9B [1]					-	44.7
Flamingo80B [1]					-	50.6
PNP-VQA $T0_{3B}$ [10]					32.3	26.6
PNP-VQA $T0_{11B}$ [10]					33.4	30.5
PNP-VQA $UnifiedQAv2_{3B}$ [10]					42.3	34.1
PNP-VQA $UnifiedQAv2_{11B}$ [10]					41.9	35.9
BLIP-2 $OPT_{2.7B}$* [7]	✓	2	129M	129M	32.5	31.5
BLIP-2 $FlanT5_{XL}$* [7]	✓	2	129M	129M	43.9	41.2
BLIP-2 $OPT_{2.7B}$ [7]	✗	1	14M	105M	32.2	25
X-Former (Ours) $OPT_{2.7B}$	✗	1	14M	105M	**34.3**	**27.6**
BLIP-2 $FlanT5_{XL}$ [7]	✗	1	14M	105M	42.9	38.2
X-Former (Ours) $FlanT5_{XL}$	✗	1	14M	105M	**44.9**	**39.5**

language modeling loss while $FlanT5$ model is trained with prefix language modeling loss i.e., caption is split into two parts: prefix and suffix. The prefix text along with visual representation forms input to LLM encoder and the suffix text is used as generation target for LLM decoder. A random value from start to middle of sentence is picked to divide the caption into two parts.

We demonstrate that our model outperforms BLIP-2 [7] at scale in Table 4. Specifically, our model achieves a 2.1% gain on GQA dataset and 2.6% gain on OKVQA dataset respectively with $OPT_{2.7B}$ LLM. We show similar gains using $FlanT5_{XL}$ LLM as well; our approach improves by 2% on GQA dataset and 1.3% on OKVQA dataset respectively. Note that PNP-VQA [10] performance relies heavily on QA model specifically UnifiedQAv2 is a task-specific model pretrained for question answering, and OFA [13] trains visual encoders while we keep it frozen hence we do not compare with it.

4 Ablation Analysis

Ablation On CLIP Layers As mentioned in Section 3.3 of main text "Leveraging Early Layer CLIP features", we present additional results by experimenting with different layers from CLIP. We experiment with the following layers $\{22, 24, 26, 28, 30, 32, 34, 36\}$ as early layer features from CLIP ViT and report performance trend on GQA dataset. As shown in Figure 1, we observe that the best performance is achieved for layer 26 and layer 30, while utilizing features from layers below 26 leads to a drop in performance. Furthermore, using features from layers beyond 30 also results in a decline in performance.

Our findings demonstrate that the performance using early layer CLIP features is inferior to that of our model, with a 2% decrease in performance compared to the best layer.

More Ablations We perform comprehensive ablations studies to analyze the impact of different loss components, effect of Horizontal flip augmentation and effect of Self-Attention. Further, we analyze the impact of X-Former training for LLM alignment and importance of MAE to capture detailed visual information complementing the global semantic representation from CLIP-ViT. Note that for these ablations, we use 8-A100swith batch size of 320/272 for stage 1 and stage 2 respectively.

We find that ITG significantly affects retrieval more than ITM; without ITC, there is a slight drop in captioning performance. As shown in Table 5, horizontal flip augmentation does not effect overall performance. For comprehensiveness, we analyze the effect of Self-Attention (SA) layer in X-Former as shown in Table 5, row 5. There is a drop in captioning performance when we remove SA layer before the Cross-Attention with MAE.

Table 5: Ablations Analysis.

Method	**TR5**	**TR10**	**IR5**	**IR10**	**B@4**	**C**
w/o ITM	96.6	98.8	93.8	96.7	35.9	120.3
w/o ITG	84.9	93.1	88.2	92.9	-	-
w/o ITC	-	-	-	-	36.2	120.7
w/o HFlip	93.2	98.4	93.9	97.2	36.3	122.4
w/o SA	95.5	99	93.9	97.1	35.6	120
X-Former (Ours)	95.8	99	94	96.7	37	123.2

Table 6: Ablation Analysis. *: smaller batch size in both stages

Method	GQA	OKVQA
CLIP w Recon.	22.5	8.1
X-Former Frozen	25.5	15.9
X-Former (Ours)	**31.9**	**25.9**

For LLM alignment, we follow BLIP-2 protocol and train X-Former in stage-2 along with a Fully Connected layer. To analyze the impact of training X-Former for LLM alignment, we experiment with frozen X-Former in stage-2 and report results in Table 6 row 2. To demonstrate the importance of MAE further, we replace MAE encoder with CLIP-ViT and pass masked image to CLIP-ViT which is then optimized for image reconstruction with MAE decoder. As shown in Table 6 row 1, the performance drops significantly by replacing

MAE-ViT encoder with CLIP-ViT on both GQA and OKVQA dataset. Thus demonstrating MAE-ViT encoder plays crucial role in learning detailed visual features.

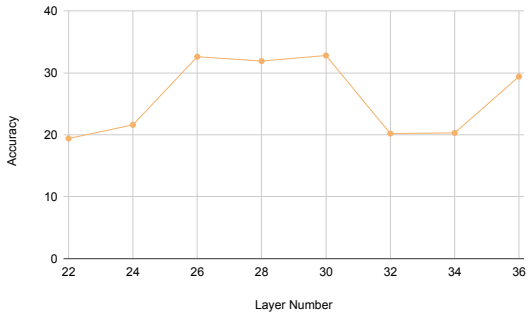


Fig. 1: Zero-shot visual question answering performance on GQA datasets for different layer features from CLIP.

5 Qualitative Results

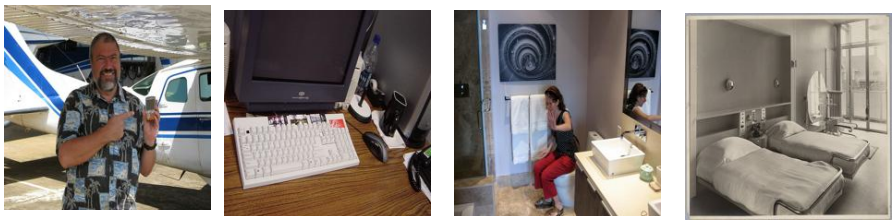


Fig. 2: Qualitative Comparison demonstrating ability to compare colors of specified objects.

In this section, we present qualitative results, including cases where our method did not perform as expected. In Figure 2, we present examples that involve comparing the colors of different objects within the image. As you can see in Figure 2 (a), (b), and (c), our method successfully understands the specified objects in the questions, regardless of their positions in the image, and

compares their colors accurately. However, Figure 2 (d) shows a more challenging scenario. Here, the pillow and the bed are not clearly distinguishable, which made it difficult for our model to identify the pillow in the image.

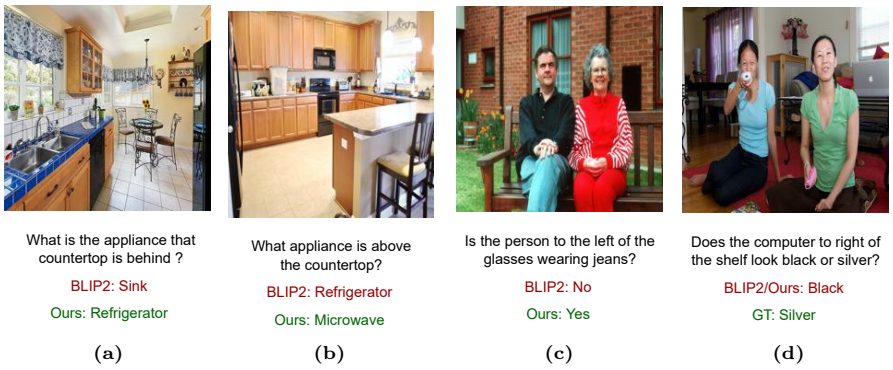


Fig. 3: Qualitative Comparison pertaining to question of spatial understanding.

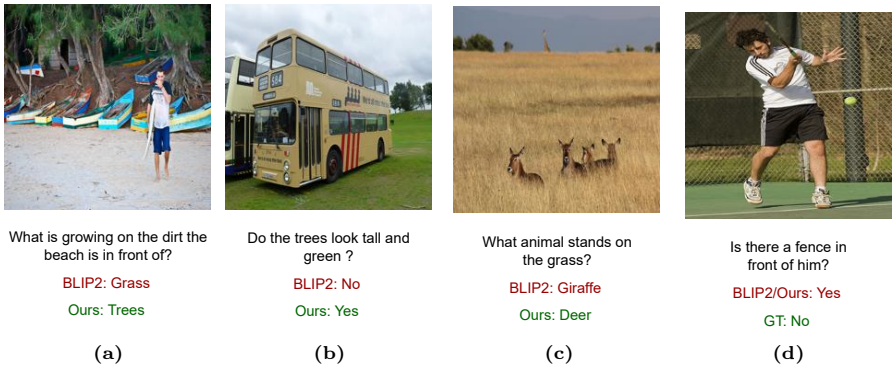


Fig. 4: Qualitative Comparison pertaining to question of relative object understanding in both background and foreground.

Figure 3 showcases the spatial understanding capabilities of our model in comparison to the BLIP-2. A kitchen scene depicted in images 3(a) and (b), our model accurately identifies the refrigerator behind the countertop and the microwave above it, respectively. In contrast, BLIP-2 erroneously predicts a sink and refrigerator for the same questions. Our model correctly discerns the attire of individuals in 3 (c), recognizing that the person to the left of the one wearing glasses is indeed wearing jeans—a detail that BLIP-2 overlooks. However, 3 (d) presents a more challenging scenario for both models. When asked about

the color of the computer to the right of the shelf, our model and BLIP-2 both incorrectly identify a silver laptop as black.

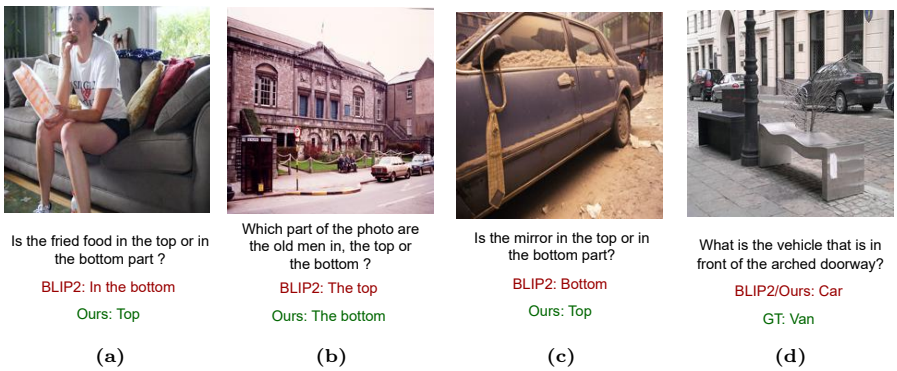


Fig. 5: Qualitative Comparison for questions relating to absolute image position understanding.

Figure 4 provides insights into the ability of our model to comprehend the relative positioning of objects within an image, both in the background and foreground. Figure 4 (a) and (b), probe the understanding of background elements, our model demonstrates a clear capacity to correctly identify objects, distinguishing trees on a beach and recognizing the tall, green trees beside a double-decker bus. This is in contrast to BLIP-2, which incorrectly identifies grass instead of trees and fails to acknowledge the verdancy and height of the trees. Further, in Figure 4 (c), which shifts the focus to foreground objects, our model accurately discerns the presence of deer in a grassy field. However, in Figure 4 (d), both our model and BLIP-2 inaccurately detect a fence in front of a tennis player, when, in fact, it is behind the player as shown. Overall, our model shows enhanced understanding of object contexts and positioning.

Figure 5 exemplifies the absolute position reasoning capabilities of our model by assessing its ability to identify objects and their locations within an image, whether they are situated at the top or bottom parts of the image. In Figure 5 (a), (b) and (c), our model accurately determines the position of the fried food, the location of the old men, and the placement of the mirror, respectively, demonstrating better understanding of absolute positions within various contexts. However, Figure 5 (d) introduces a more complex situation involving multiple vehicles parked. Our model encounters difficulty here, incorrectly identifying a van as a car due to its close resemblance to car in this image.

Figure 6 demonstrates the capacity of our model to discern fine details of objects in close proximity within an image. In the living room scene depicted in Figure 6 (a), when questioned about the presence of a girl to the right of a pillow, our model accurately confirms the absence of a girl, whereas BLIP-2 incorrectly asserts a presence. The image of a food plate, shown in Figure 6(b)

and (c), further probe the model's ability to understand foreground and background distinctions. Our model correctly identifies the fruit next to the cake as a strawberry, where as BLIP-2 incorrectly categorizes it as raspberries. Additionally, our model successfully distinguishes the white color of the plate amidst the various food items placed upon it, indicating a enhanced visual perception capabilities.

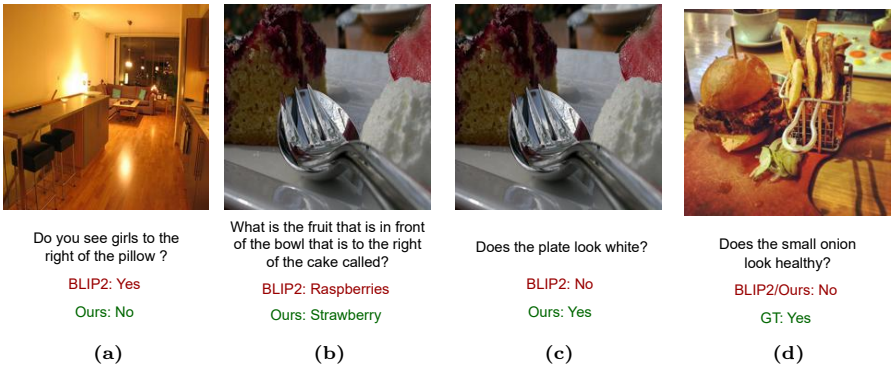


Fig. 6: Qualitative Comparison for samples with objects in close proximity in the scene.

However Figure 6 (d), presents a complex scenario involving an assessment of an onion's quality, both our model and BLIP-2 fail to correctly evaluate its healthiness. This highlights the challenge in assessing condition/quality of food which is subject to interpretation.

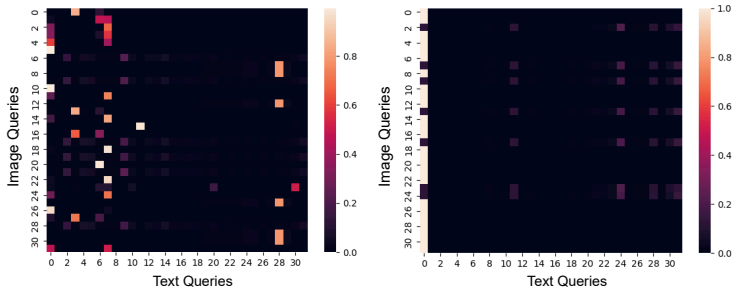


Fig. 7: Query Diversity: Ours (left), BLIP-2 (right)

6 Query Diversity Comparison

In addition to the fine-grained qualitative comparison, we further perform a fine-grained analysis of image-text queries from BLIP-2 and Ours. Particularly, we analyze the diversity of image-text query similarities as a proxy to investigate the fine-grained interaction between the image and text queries. For this, we first compute similarity between output queries for images and text using both Ours and BLIP-2 model to get 32×32 matrix for each image-text pair as shown in Figure 7. We then aggregate the similarities by marginalizing them over the text to get scores for each image query. To compute the overall diversity, we repeat this for all the samples in COCO karpathy test split and average it across samples and image queries. Our findings indicate that the queries learnt by our model are 7% more diverse than those of BLIP-2, demonstrating the enhanced capability of our approach to capture a broader range of nuances in image-query representations. We present more fine-grained qualitative examples to show query diversity of our model compared to BLIP-2. As shown in Figure 8, our model learns diverse queries than BLIP-2 for few samples from COCO karpathy test split.

Table 7: Performance on SugarCrepe.

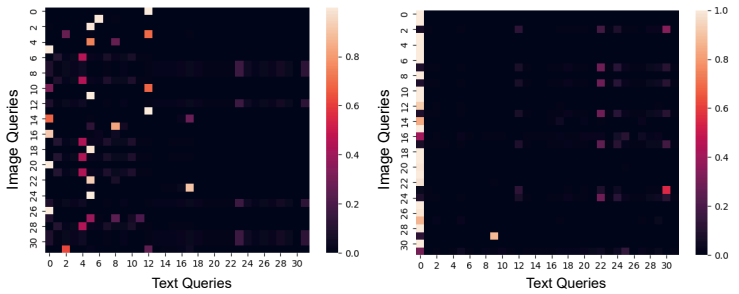
Method	**Object**			**Attribute**			**Relation**
	Replace	Swap	Add	Replace	Swap	Add	Replace
BLIP-2	93.4	56.7	89.1	81.9	66.9	83	72.1
X-Former (Ours)	**95.7**	**64.1**	**92.1**	**84.2**	**68.6**	83	**75.8**

Table 8: Results on Flowers-102/Food-101

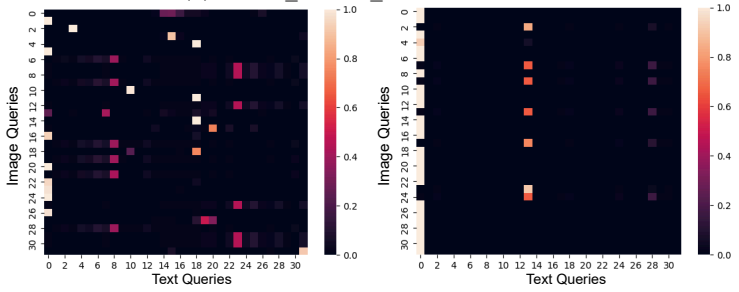
	Flower102	**Food101**
BLIP-2	53.3	82.7
X-Former (Ours)	**58.2**	**83.7**

7 Additional Fine-Grained Results

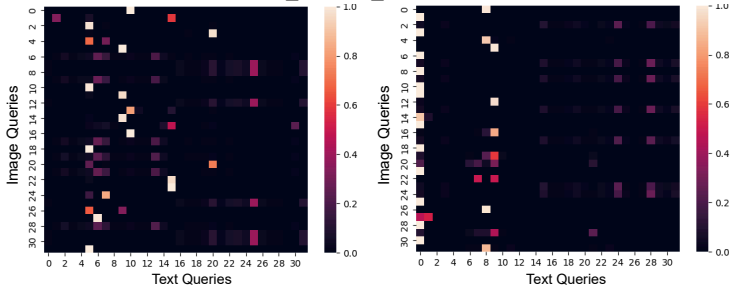
OC/MCI [12] have been proposed in 2023 as benchmarks for measuring the capability of MLLM in comprehending and reasoning about fine-grained visual features. Note that our usage of 'fine-grained' refers to high-frequency and detailed visual representations that are overlooked by current MLLMs with CLIP-ViT as the visual backbone. It does not refer to traditional fine-grained vision tasks



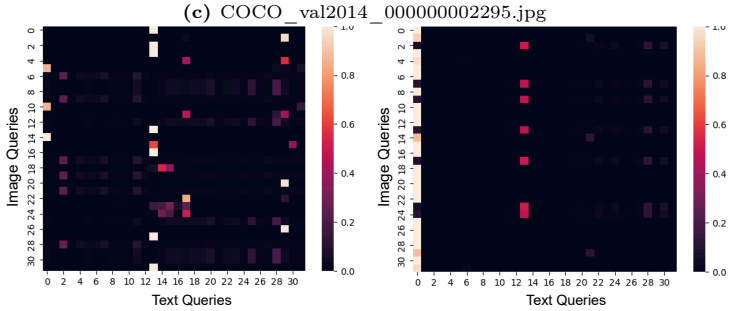
(a) COCO_val2014_000000001757.jpg



(b) COCO_val2014_000000031442.jpg



(c) COCO_val2014_000000002295.jpg



(d) COCO_val2014_000000002240.jpg

Fig. 8: Qualitative comparison of the queries for our model (left) with BLIP-2 (right). Our model learns diverse queries compared to BLIP-2.

(e.g., bird species classification) that require fine-grained *annotations*. We further evaluate our model for more fine-grained tasks, although our model is not dedicated to traditional "fine-grained tasks". We evaluate on *SugarCrepe (SC)* [5], *Flowers102* [9], and *Food-101* [2]. As shown in Tab. 7, 8, XFormer outperforms BLIP-2 by a large margin, indicating XFormer is also good at distinguishing visually similar objects.

References

1. 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)
2. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: European Conference on Computer Vision (2014)
3. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022)
4. 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)
5. Hsieh, C.Y., Zhang, J., Ma, Z., Kembhavi, A., Krishna, R.: Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In: Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023)
6. 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)
7. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023)
8. 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)
9. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian conference on computer vision, graphics & image processing. IEEE (2008)
10. Tiong, A.M.H., Li, J., Li, B., Savarese, S., Hoi, S.C.: Plug-and-play VQA: Zero-shot VQA by conjoining large pretrained models with zero training. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2022. Association for Computational Linguistics (2022)
11. 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)
12. Wang, G., Ge, Y., Ding, X., Kankanhalli, M., Shan, Y.: What makes for good visual tokenizers for large language models? (2023)
13. Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H.: Ofa: Unifying architectures, tasks, and modalities through a simple

sequence-to-sequence learning framework. In: International Conference on Machine Learning. PMLR (2022)