Question Aware Vision Transformer for Multimodal Reasoning

Supplementary Material

A. Implementation Details

Overall Training Protocol For all of the considered architectures, we follow the same general training procedure in which we apply LoRa [23] to the LLM and finetune the projection module. When applying QA-ViT, we also finetune the instruction representation projection MLPs. In particular, we employ LoRa (α=32, r=16, dropout=0.05, and the queries and keys as the target modules) and utilize an AdamW [36] optimizer ($\beta_1, \beta_2 = 0.9, 0.999$ and $\epsilon = 1e-08$) with cosine annealing scheduler [35] that decays to $\times 0.01$ from the base learning rate. In addition, we perform 1000 warm-up steps. We use 8 Nvidia A100 (40G) GPUs in all of our experiments with bfloat16. Next, we provide the specific implementation details regarding ViT+T5, BLIP2, InstructBLIP, and LLaVA-1.5.

ViT+T5 ViT+T5 is comprised of a CLIP [41] ViT-L vision encoder that operates in a 336×336 resolution, coupled with a FLAN-T5 encoder-decoder model [14] using an MLP projection module. The projection component consists of two linear layers that map from the ViT's dimension D_1 into the LLM's one D_2 ($D_1 \rightarrow D_2 \rightarrow D_2$). We train three variants of ViT+T5, which differ in the LLM scale, where we consider `base`, `large`, and `xl`. We use the LLM's encoder as the question encoder and train the models on our multi-task dataset (Sec. 4.1) for 5, 2, and 2 epochs, using a batch size per GPU of 16, 8, and 6, with a learning rate of $1e-4$, $5e-5$ and $1e-5$, respectively. QA-ViT introduces 38M, 45M, and 66M trainable parameters out of the overall 589M, 1,132M, and 3,220M. In addition, when applying QA-ViT to a pretraining-free setup, we observe that using a higher learning rate ($\times 100$) for the projection module stabilizes the training. We hypothesize that while the vision encoder and LLM are pretrained separately, the projection module is randomly initialized, and thus, its weights should be adjusted more than the former counterparts.

BLIP2 and InstructBLIP We experiment with both the `xl` and `xxl` models, and similar to the ViT+T5, we use the LLM's encoder for processing the question before feeding it into QA-ViT . We use a single learning rate group for all models for all the trainable parameters. For the `xl` models, we train for 2 epochs, with a batch size of 8 per GPU with a base learning rate of $2e-5$. For the `xxl` ones, we reduce the batch size to 4 per GPU. In addition, we employ a weight decay of 0.05 for all models.

Template
<image>"A short image caption:"
<image>"A short image description:"
<image>"A photo of"
<image>"An image that shows"
<image>"Write a short description for the image."
<image>"Write a description for the photo."
<image>"Provide a description of what is presented in the photo."
<image>"Briefly describe the content of the image."
<image>"Can you briefly explain what you see in the image?"
<image>"Could you use a few words to describe what you perceive in the photo?"
<image>"Please provide a short depiction of the picture."
<image>"Using language, provide a short account of the image."
<image>"Use a few words to illustrate what is happening in the picture."

Table 5. **Captioning instruction templates**. The instruction templates used for the captioning datasets. For VQA, we simply use the provided question.

LLaVA-1.5 As LLaVA-1.5 is based on a decoder-only LLM, we use the model's embedding module to process the questions when applying QA-ViT . We train for one epoch with an effective batch size of 4 per GPU (using 2-step gradient accumulation) and a base learning rate of $5e-5$.

B. Multi-Task Training Dataset and Evaluation

As stated in Sec. 4.1, we utilize a multi-task dataset that contains multiple benchmarks of different tasks. In Tab. 6, we provide a detailed list of the training datasets and the evaluation metric and split used for reporting results throughout the paper.

C. Image Captioning Templates

For the VQA-based datasets, we simply utilize the provided question to guide QA-ViT. However, in the captioning case, it is infeasible. Thus, we use the captioning templates used in InstructBLIP [15] and provide them in Tab. 5 for completeness. These captions are sampled uniformly during training and inference.

D. Additional OCR Results

D.1. In-Depth Scene-Text analysis

As explained in Sec. 4.5, we view the scene-text benchmarks as an interesting testing bed for our approach. To understand the contribution of QA-ViT for scene-text understanding, we follow the analysis of Ganz et al. [20] and decompose the results of $\mathrm{VQA}^{\mathrm{T}}$ into two non-overlapping subsets – i) $\mathrm{VQA}^{\mathrm{T}}_{\mathrm{See}\cap\mathrm{Read}}$ is the manually curated subset which contains questions that require reasoning over OCR

Task	Dataset	Description	Eval split	Metric
Image Caption	COCO	Captioning of natural images	karpathy-test	CIDEr(↑)
Scene-Text Caption	TextCaps	Text-oriented captioning of natural images	validation	CIDEr(↑)
General VQA	VQA^{v2}	VQA on natural images	test-dev	vqa-score(↑)
	Visual Genome	VQA on natural images	-	-
Scene-Text VQA	VQA^{T}	Text-oriented VQA on natural images	validation	vqa-score(↑)
	VQA^{ST}	Text-oriented VQA on natural images	test	ANLS(↑)
	VQA^{OCR}	Text-oriented VQA on book covers	-	-
Documents Understanding	DocVQA	VQA on scanned documents	test	ANLS(↑)
	InfoVQA	VQA on infographic images	test	ANLS(↑)
	ChartQA	VQA on chart images	-	-

Table 6. **Training datasets and evaluation**. The datasets used for training alongside their evaluation split and metric, if applicable.

Method	LLM	**Scene-Text**			**Documents**		
		VQA^{T}	VQA^{T}_{Read}	$VQA^{T}_{See\cap Read}$	DocVQA	InfoVQA	Average
ViT+T5-xl	Flan-T5-xl	48.0	49.3	35.6	42.3	26.4	34.4
+ QA-ViT		50.3	51.8	36.2	44.2	27.1	35.7
△		**+2.3**	**+2.5**	**+0.6**	**+1.9**	**+0.7**	**+1.3**
BLIP2	Flan-T5-xl	34.5	36.1	18.7	16.1	21.1	18.6
+ QA-ViT		36.6	38.3	20.4	17.1	21.2	19.2
△		**+2.1**	**+2.2**	**+1.7**	**+1.0**	**+0.1**	**+0.6**
InstructBLIP	Flan-T5-xl	36.2	37.9	19.3	17.3	19.9	18.6
+ QA-ViT		37.4	39.0	22.5	18.2	20.5	19.3
△		**+1.2**	**+1.1**	**+3.2**	**+0.9**	**+0.6**	**+0.7**
LLaVa-1.5	Vicuna-7B	57.4	59.0	42.5	44.1	32.1	38.1
+ QA-ViT		59.1	60.7	43.5	45.4	32.1	38.8
△		**+1.7**	**+1.7**	**+1.0**	**+1.3**	**0.0**	**+0.7**

Table 7. **Additional OCR Results.** Results on documents understanding and comprehensive VQA^{T} analysis.

and visual information simultaneously. We view this subset as the most challenging one. ii) VQA^{T}_{Read} is composed of questions that can be answered solely by using the OCR information. The unification of these subsets results in the entire VQA^{T} validation set. We provide the results on these subsets on the middle section of Tab. 7. As can be seen, QA-ViT improves the results on VQA^{T}_{Read} in all the models. This highlights the ability of our method to better harness some of the overlooked OCR information. In addition, it leads to consistent improvements on the $VQA^{T}_{See\cap Read}$, which requires cross-modal reasoning over the OCR and visual cues.

D.2. Documents Understanding

In this section, we present the performance results of both QA-ViT and the various baseline models in the context of document understanding, evaluated on DocVQA and InfoVQA, as detailed in the right section of Tab. 7. DocVQA encompasses questions related to dense-text scanned documents, while InfoVQA is designed for reasoning over infographics. Operating in these domains is highly challenging as it constitutes a substantial domain shift for the CLIP vision encoder (from natural images to documents and inforgraphichs). Moreover, as CLIP is inherently limited in dense-text scenarios, the application of QA-ViT, which specifically targets existing visual features, is not anticipated to yield a significant performance boost in such settings. Despite these challenges, our results, while far from state-of-the-art levels, consistently demonstrate improvements over baseline performance. This underscores the effectiveness of our method in directing visual attention towards OCR information within the given constraints.

E. Additional Qualitative Results and Analysis

In Fig. 6, we extend the visualizations conducted in the main paper to focus on the alignment of the text queries and visual features and provide additional demonstrations:

- We provide attention visualizations at three levels of granularity within the ViT: (i) before the question fusing, (ii) immediately after it, and (iii) at the final layer. Illustrated in Fig. 6, in (i), the network's attention spans across the entire visual content, while in (ii) and (iii), it focuses

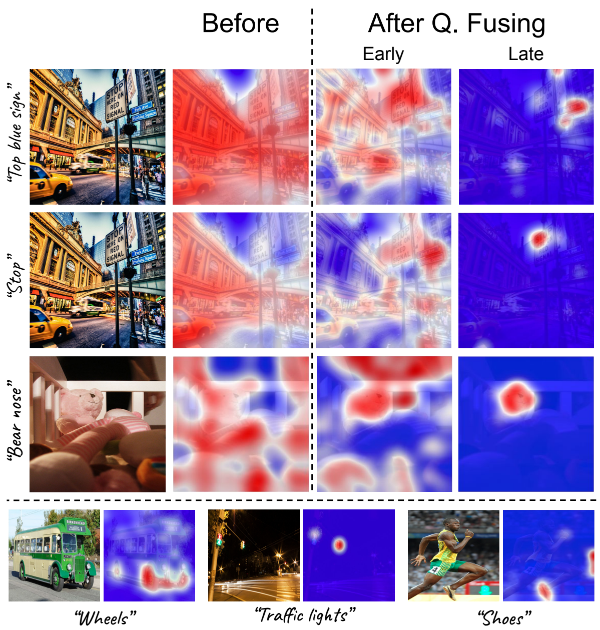

Figure 6. **Elaborated interpretations of QA-ViT.** Additional visual and textual features interaction demonstrations, including visualizations at different granularity levels within the ViT.

on fine-grained details according to the provided text. Specifically, the interaction of the text and vision throughout QA-ViT leads to more focused attention maps, as can be seen in the rightmost two columns.

- To better demonstrate the fine-grained interaction of text and vision in QA-ViT, we show the attention maps of the same image with respect to different text prompts (top two rows). This highlights QA-ViT's ability to shift the focus of the visual features based on the provided text.
- The bottom row contains additional visual-textual attention visualization, indicating QA-ViT's text-based focus.

In addition, we provide qualitative comparison between QA-ViT and and the baseline in Fig. 7.

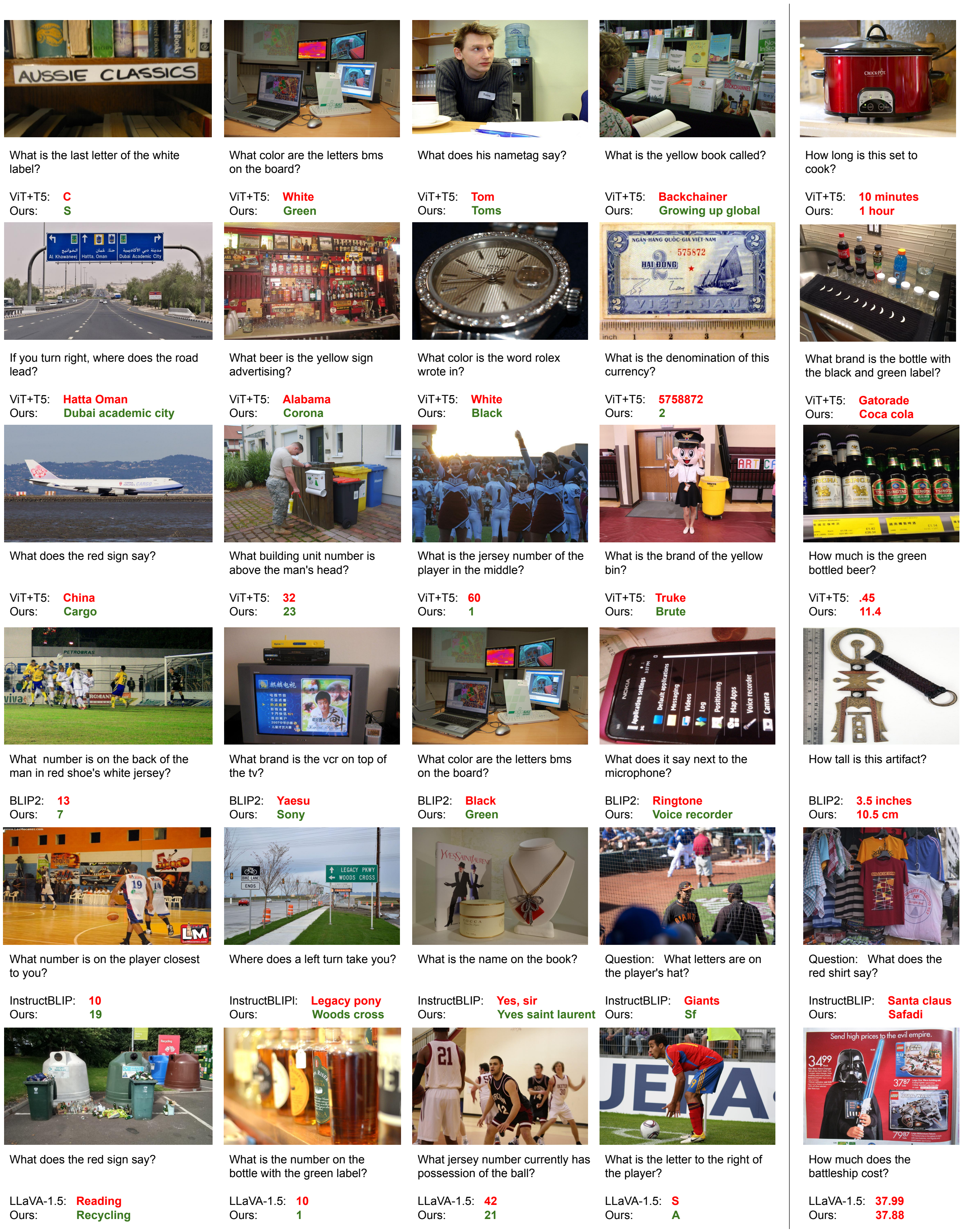

Figure 7. **Additional qualitative results.** Comparison between the baseline and our method on VQA^T validation set using ViT+T5 (`base`, `large`, `xl`), BLIP2 and InstructBLIP (`xxl`) and LLaVA-1.5. Success and fail cases are presented on the left and right, respectively.