QID: Efficient Query-Informed ViTs in Data-Scarce Regimes for OCR-free Visual Document Understanding

Binh M. Le[1]*†, Shaoyuan Xu[2]*, Jinmiao Fu[2], Zhishen Huang[2], Moyan Li[2], Yanhui Guo[2],
Hongdong Li[2,3], Sameera Ramasinghe[4], Bryan Wang[2]
[1]Sungkyunkwan University, S. Korea [2]Amazon, USA [3]ANU, Australia [4]Pluralis Research, Australia
bmle@g.skku.edu {shaoyux,jinmiaof,hzs,moyanli,yanhuig,hongdli}@amazon.com
samramasinghe@gmail.com brywan@amazon.com

Abstract

In Visual Document Understanding (VDU) tasks, fine-tuning a pre-trained Vision-Language Model (VLM) with new datasets often falls short in optimizing the vision encoder to identify query-specific regions in text-rich document images. Existing methods that directly inject queries into model layers by modifying the network architecture often struggle to adapt to new datasets with limited annotations. To address this, we introduce QID, a novel, streamlined, architecture-preserving approach that integrates query embeddings into the vision encoder, leading to notable performance gains, particularly in data-scarce fine-tuning scenarios. Specifically, our approach introduces a dual-module framework: a query-aware module that generates a unique query vector to precisely guide the model's focus, as well as a query-agnostic module that captures the positional relationships among tokens, ensuring robust spatial understanding. Notably, both modules operate independently of the vision attention blocks, facilitating targeted learning of query embeddings and enhancing visual semantic identification. Experiments with OCR-free VLMs across multiple datasets demonstrate significant performance improvements using our method, especially in handling text-rich documents in data-scarce environments.

1. Introduction

Recent advancements in vision-language models (VLMs) have significantly impacted Visual Document Understanding (VDU), enabling models to interpret text-rich document images across various tasks. VDU methods are generally divided into two categories: optical character recognition (OCR)-dependent methods and OCR-free methods. Although OCR-dependent methods [2, 13, 33, 36, 38, 39] have obtained robust results, they are often bottlenecked by the computational latency and errors introduced by the OCR engines. [15, 32]. In contrast, OCR-free approaches [3, 11, 15, 40, 41] bypass these challenges, directly modeling the visual context of documents. These techniques generally utilize large VLMs trained in extensive VDU tasks [24,26,27,30], where visual and textual embeddings are jointly processed by language models to answer questions about an image.

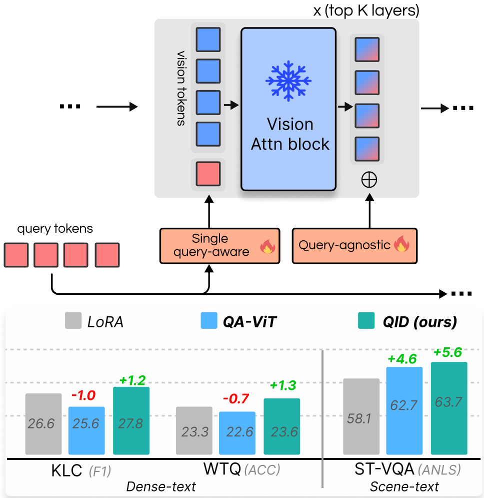


Figure 1. **Illustration of our approach. Top:** Unlike previous work (e.g., QA-ViT [9]), our method detaches the query-informed module *independently* from the attention block, decomposing it to a query-aware module learning single embedding vector and a query-agnostic module, thereby reducing computational demands during training and inference. **Bottom**: Comparative results of our proposed method on two dense-text and one scene-text image datasets versus baseline methods, applying fine-tuning to the Qwen-VL-Chat model with only 1,000 samples per dataset.

However, fine-tuning pre-trained VLMs for VDU, especially with limited data, presents unique challenges. Unlike general visual question answering (VQA) tasks, VDU of-

[*] These authors contributed equally.
[†] Work done during internship at Amazon.

ten requires a deeper contextual understanding and domain-specific knowledge, especially in dense text documents, where the annotation of logical inference can be complex [25]. This challenge underscores the need for efficient adaptation in data-scarce settings, critical for VDU applications in fields like medical or legal document analysis, with minimal additional parameters. Existing Parameter-efficient fine-tuning (PEFT) techniques, such as LoRA [12] and DoRA [22], are popular for adapting language models but struggle with high-complexity VDU tasks since they only enhance the linguistic components. More recently, a stream of research called query-aware (QA) [1, 9, 37] has shown promise by injecting query embeddings into the visual encoding process, enhancing query-specific visual attention. However, these methods struggle with dense text documents under data-scarce conditions, where they introduce inefficiencies and fail to generalize effectively.

To address these limitations, we present **QID**, a novel, lightweight Query-Informed Vision Transformer (ViT) designed particularly in Data-scarce regimes. Our method introduces two modules: (1) a query-aware module, which generates a single, robust query embedding vector to align the model's focus with relevant document regions, and (2) a query-agnostic module, which captures positional dependencies across visual tokens and mitigates distribution shifts introduced by the query, to help the model maintain consistency in general layout patterns across diverse documents. To enhance local cross-attention, we also introduce fuse and defuse learning steps in the query-aware module, which refine query embeddings through spherical augmentation and entropy regularization. These steps enable precise query alignment with visual elements, even in data-limited scenarios. Both query-aware and query-agnostic modules operate independently of the core attention blocks, enabling efficient adaptation to query-relevant features with minimal data. As illustrated in Fig. 1, our approach not only maintains architectural simplicity but also delivers substantial performance gains across OCR-free VDU tasks, particularly in data-scarce regimes.

Our contributions are summarized as follows:

- We propose a lightweight, architecture-preserving approach that integrates query embeddings into the vision encoder without modifying the core attention blocks. This method enhances the VLM's ability to focus on query-relevant regions in text-rich documents, particularly beneficial in data-scarce scenarios.

- Our framework introduces a query-aware module for augmenting question representation and a query-agnostic module to address positional dependencies and vision distribution shift. We further enhance the query-aware module with fuse and defuse learning steps, using spherical augmentation and entropy regularization to improve query alignment and robustness.

- Extensive experiments across multiple VDU datasets with pre-trained VLMs validate the efficiency and effectiveness of our approach, demonstrating consistent improvements over State-of-The-Art (SoTA) QA and PEFT methods in data-scarce settings.

2. Related Work

2.1. Visual Document Understanding

Visual Document Understanding (VDU) seeks to interpret and reason logically about a wide range of digitalized document images. Strategies in VDU are categorized into two main approaches. OCR-dependent approaches integrate images with an external OCR engine to annotate textual content, exemplified by works such as [2, 13, 38]. Alternatively, OCR-free approaches train large vision-language models (LVLMs) on extensive datasets, as seen in [3,15,40], enhancing capabilities without relying on OCR. Notably, mPLUG-DocOwl [40] enhances the capabilities of LVLMs for VDU by introducing a modular model based on mPLUG-Owl [42] designed specifically for OCR-free document understanding. Pix2Struct [17] pioneers an approach of screenshot parsing objectives, and UReader [41] innovates with a shape-adaptive cropping module that precedes the encoder-decoder architecture, leveraging a frozen low-resolution vision encoder for processing high-resolution images. However, a common limitation of these OCR-free approaches is that the vision encoder processes images without contextual knowledge of the textual prompts, which can lead to suboptimal vision representation for VDU tasks.

2.2. Query-Informed ViTs

InstructBLIP [37] attempted early textual instruction integration into vision embeddings via a QFormer atop the vision encoder, potentially overlooking image representation nuances. Subsequently, VisFocus [1] was proposed to encode prompts and perform cross-attention with vision tokens at every layer, but its performance did not surpass that of baseline models, as it was targeted at developing a lightweight model. Recently, QA-ViT [9] advocates the idea of appending the prompt tokens directly to the vision tokens in some of the final vision attention blocks. While this method has shown promise in simple VQA tasks, it necessitates modifications to the original vision blocks, resulting in a much-increased number of training parameters. Furthermore, when fine-tuning with dense-text datasets, applying the QA-based methods on top of LoRA [12] degrades the performance considerably, compared to using LoRA alone, especially in data-scarce regimes.

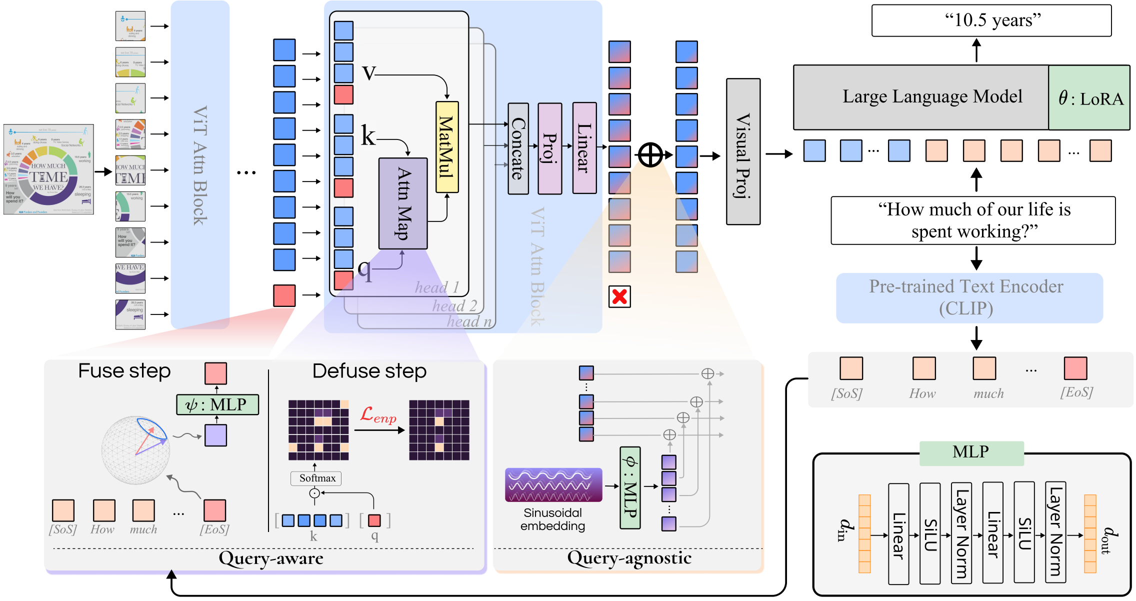


Figure 2. **Illustration of our end-to-end training procedure.** For simplicity, this figure demonstrates how our approach is integrated with the last ViT attention block. Note it can also be applied to other layers of an vision encoder. During fine-tuning stage, only the green modules are optimized. Our query-aware module, enhanced by the fuse and defuse learning steps, makes the query embedding more robust for the vision encoder. Our query-agnostic module offsets distribution shifts caused by the query information and, as it operates independently from the query vector, can be precomputed and saved as a bias term post-training. This efficient learning approach on a single query vector makes our proposed method lightweight and highly effective for VLMs in VDU tasks.

3. Methods

3.1. Overall Architecture

Figure 2 gives an overview of our method, showing how it integrates the query embedding into the vision attention block. Unlike previous methods [1, 9], we adopt the [EoS] (end-of-sentence) token of a query embedding from a pre-trained text encoder and make no modification to the network architecture of the vision attention block. This token vector is initially processed through a query-aware module, which injects the question embedding into the vision block. This query-aware module consists of two steps, i.e., fuse and defuse steps, aiming to enhance the query embedding under data-scarce conditions. The fuse step augments the query embedding on a hypersphere, while the defuse step focuses its attention on the most relevant visual areas. The vision tokens are subsequently adjusted by a query-agnostic module that learns a sinusoidal embedding and adds it directly to the vision tokens. Such query-agnostic components can be pre-computed after training, and used directly during inference to save computation.

The notations used in this paper are defined as follows: Given an image $\mathcal{I}$ and a query (i.e. the question) $\mathcal{Q}$, the vision encoding output from a pre-trained vision encoder V, which consists of L layers and is conditionally dependent on $\mathcal{Q}$, can be formalized as follows:

$$\mathbf{z}^L = \{z_i^L\}_{i=1}^{T_v} = V(\mathcal{I} \mid \mathcal{Q}), \quad z_i^L \in \mathbb{R}^{d_v}, \tag{1}$$

where T_v and d_v represent the number and the dimension of vision tokens, respectively. Typically, V is represented as a stack of L identical attention blocks, where each block is defined by:

$$\bar{\mathbf{z}}^l = \texttt{Proj}(\texttt{MSA}(\mathbf{z}^{l-1})) + \mathbf{z}^{l-1} \tag{2}$$

$$\mathbf{z}^l = \texttt{FFN}(\bar{\mathbf{z}}^l) + \bar{\mathbf{z}}^l, \tag{3}$$

where MSA, Proj, and FFN denote the multi-head self-attention, projection, and feed forward multi-layer perceptron layers, respectively.

3.2. Query-Aware Module

3.2.1 Single Query Embedding

With the increasing prevalence of decoder-only large language models, such as GPT-3 [6] and LLaMA [34], encoding a query (or question) often requires an external pre-trained text encoder. In this paper, we employ the text encoder from CLIP [29], known for its effectiveness in integration with multi-modal models for generation tasks [16] and segmentation tasks [5].

Let $\mathbf{q} \in \mathbb{R}^{d_t \times T_t}$ represent the text embedding of a query generated by the pre-trained text encoder, where d_t is the dimension of the text embedding space, and T_t is the maximum number of text tokens. While previous work [9] uses full text embeddings to train the model to capture the entire query representation, we find this approach suboptimal in data-scarce settings. Furthermore, [20] suggests that in

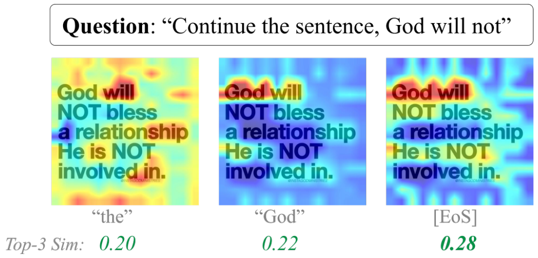

Question: “Continue the sentence, God will not”
God will NOT bless a relationship He is NOT involved in.
God will NOT bless a relationship He is NOT involved in.
God will NOT bless a relationship He is NOT involved in.
“the”
“God”
[EoS]
Top-3 Sim:
0.20
0.22
0.28

Figure 3. Effect of the [EoS] token in the query for highlighting semantic areas in the image [20]. The green numbers at the bottom indicate the top-3 cosine similarities between tokens and the image, as computed using CLIP embeddings [29].

models like CLIP, the end-of-sentence ([EoS]) token, punctuation marks, and non-object words often highlight corresponding semantic regions in images, whereas other tokens typically do not. Our findings confirm this, as illustrated in Fig. 3 , where the top three text tokens most similar to the image in the CLIP model [29] reveal that [EoS] highlights the most relevant visual area, while tokens with lower similarity, such as “the,” may even highlight background regions. This outcome arises from the fact that the [EoS] token serves as a holistic sentence representation in contrastive learning with vision tokens within the CLIP model. Therefore, using a single [EoS] token, denoted as $q_{\text{eos}} \in \mathbb{R}^{d_t}$, from the query embedding offers two key benefits. First, it is the most efficient token for highlighting important spatial areas in an image. Second, with limited fine-tuning data, optimizing the downstream module to integrate query information into the vision model using a single token is both more efficient and robust than using all text tokens, which often introduce additional noise.

3.2.2 Fuse and Defuse Learning Steps

Through experiments, we have found that combining QA-ViT [9] with PEFT methods, such as LoRA [12], often yields inferior results compared to using LoRA alone, particularly with limited fine-tuning data. For example, as shown in Table 1 and further detailed in Section 4, our experiments with the Qwen-VL-Chat model on 1,000 fine-tuning samples show degraded performance using a single token, QA-ViT+q_{eos}, compared with that with full query token, QA-ViT+**q**. We hypothesize that as the QA-ViT method [9] introduces a new Proj layer into the ViT block, it modifies the pre-trained model's attention dynamics without sufficient samples to optimize the new parameters, thereby impairing visual representation quality.

Building on the observed shortcomings of QA-ViT [9], we introduce a training paradigm that preserves the original architecture of the ViT model's attention block while utilizing a single token embedding enhanced by “fuse” and “defuse” learning steps, which are illustrated in Fig. 2. The “fuse” step enriches the query embeddings by augmenting them with a random noise vector from a Gaussian distribution around the original vector. On the other hand, the “defuse” step eliminates unrelated visual information (noises) while retaining the most relevant elements through entropy regularization.

Methods	InfoVQA	KLC	WTQ	VizWiz	ST-VQA
	ANLS ↑	*F1* ↑	*ACC* ↑	*VQA Score* ↑	*ANLS* ↑
			1K fine-tuning samples		
LoRA	34.38	26.62	23.30	38.46	58.07
QA-ViT +**q**	33.80	25.64	22.08	40.40	62.73
QA-ViT +q_{eos}	34.01	25.96	21.69	38.55	61.58
QID (*ours*)	34.18	27.81	23.56	42.08	63.69

Table 1. Comparison between the baseline method QA-ViT + q_{eos} and our approach, QID + q_{eos}, as experimented with the Qwen-VL-Chat model fine-tuned with 1,000 samples. Experimental settings are further detailed in Section 4.

Fuse step: spherical augmentation. Pre-trained CLIP models [29] are trained by maximizing the cosine similarity of text and image features for matching text-image pairs while minimizing it for mismatched pairs. Despite their extensive training on large datasets, there is still a gap remaining between image embeddings and corresponding text embeddings, rooting from the cone effect inherent in each modality's distribution [21]. Moreover, in our use case, where there is no image caption in the input but only a question about it, relying on separately pre-trained text and image encoders may underperform for VDU tasks. These observations motivate us to explore the potential of embedding space of q_{eos} to alleviate the narrow cone distribution and enhance the semantic representation of question embeddings. Meanwhile, this augmentation step fosters robust learning in scenarios with limited fine-tuning data by encouraging it to learn more generalized and resilient visual representations with respect to the query. Formally, we form pseudo text features $q'_{\text{eos}} \in \mathcal{S}(\mathcal{Q})$ for a given question $\mathcal{Q}$ on the hypersphere:

$$\mathcal{S}(\mathcal{Q}) = \{q'_{\text{eos}} | \text{Sim}(q'_{\text{eos}}, q_{\text{eos}}) > \tau\}, \tag{4}$$

where Sim denotes cosine similarity and τ is a threshold.

To generate a pseudo query feature q'_{eos}, we introduce a method to perturb the original query feature q_{eos} using adaptive Gaussian noise:

$$q'_{\text{eos}} = \frac{\widetilde{q}_{\text{eos}}}{\|\widetilde{q}_{\text{eos}}\|_2}, \quad \widetilde{q}_{\text{eos}} = q_{\text{eos}} + \sigma \cdot \|q_{\text{eos}}\|_2 \cdot \frac{\epsilon}{\|\epsilon\|_2}, \tag{5}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ represents Gaussian noise, $\sigma > 0$ is the hyperparameter that dictates the perturbation magnitude, and $\| \cdot \|_2$ denotes the ℓ_2 norm. The normalization of the Gaussian noise onto a hypersphere, followed by rescaling according to the norm of the query feature, ensures that the noise addition is adaptive. Subsequently, q'_{eos} is processed through a multi-layer perceptron (MLP) ψ to project it into the vision space $\mathbb{R}^{d_v}$. With the output $\mathbf{z}^{l-1}$ at the layer $(l-1)$-th in the vision model, the input for the next layer

l-th is formed as $\mathbf{z}^{l-1} = \text{Cat}\{\mathbf{z}^{l-1}, \psi(q'_{\text{eos}})\} \in \mathbb{R}^{(T_v+1)\times d_v}$, where Cat denotes a simple concatenation operation. This step in the process is illustrated in Fig. 2 - bottom left.

Defuse step: entropy regularization. In text-centric images, relevant visual areas are usually confined to small, localized regions (as illustrated in Fig. 4). Therefore, this step aims to constrain the query embedding to activate only specific local vision tokens via their cross-attention matrices. In the l-th vision attention block, the multi-head self-attention (MSA) layer consists of H distinct heads. For each head h $(1 \leq h \leq H)$, the token embedding z_i^{l-1} from the previous layer is projected into triplet forms: query, key, and value (note that this query is different from the query referring to the original question $\mathcal{Q}$). The matrices for query (Q_h^l), key (K_h^l), and value (V_h^l) contain corresponding elements. The self-attention matrix for the h-th head is:

$$A_h^l = \text{softmax}\left(Q_h^l(K_h^l)^T/\sqrt{d_v}\right) \in \mathbb{R}^{(T_v+1)^2}, \quad (6)$$

where softmax is applied across the columns of the inner dot product. The cross-attention between the q'_{eos} question embedding and the vision tokens is defined as:

$$A_{h|\text{cross}}^l = A_h^l[T_v+1, : T_v] \in \mathbb{R}^{T_v}. \quad (7)$$

We measure the uncertainty of a distribution using entropy over $A_{h|\text{cross}}^l$:

$$\mathcal{H}_h^l = -\sum_{i=1}^{T_v}(A_{h|\text{cross}}^l)_i \cdot \log\left[(A_{h|\text{cross}}^l)_i\right]. \quad (8)$$

To discourage a uniform distribution of $A_{h|\text{cross}}^l$ over the spatial dimension, we apply entropy regularization as follows:

$$\mathcal{L}_{\text{enp}} = \frac{1}{|L_q|}\sum_{l\in L_q}\sum_{h=1}^{H}\mathcal{H}_h^l, \quad (9)$$

where L_{q} are the predefined visual layers to which we project the query embedding q'_{eos}. The end-to-end of this process is depicted in Fig. 2.

3.3. Query-Agnostic Module

To offset the distribution shift caused by the introduction of an additional query token q'_{eos} to the frozen vision attention block, we implemented a query-agnostic module. This module functions as a bias term added to the output of the block. Unlike QA-ViT [9], which modifies the attention block based on the query token, our module operates independently, learning solely from an initiated sinusoidal signal and is inserted directly after the vision block as illustrated in Fig. 2. Formally, a fixed position embedding of dimension d_p employs a sinusoidal function [35]:

$$\mathbf{P}[i, 2k] = \sin\left(\frac{i}{10000^{\frac{2k}{d_p}}}\right), \mathbf{P}[i, 2k+1] = \cos\left(\frac{i}{10000^{\frac{2k}{d_p}}}\right). \quad (10)$$

Here, $\mathbf{P} \in \mathbb{R}^{T_v \times d_p}$, i denotes the index of the position, and k represents the index within the dimension of the embedding. We set the value of d_p to 64 in our work. Each spatial sinusoidal vector is further refined by a learnable, shallow feed-forward multi-layer perceptron (MLP) ϕ, and matched in dimension to the vision space $\mathbb{R}^{d_v}$. The output from the vision attention block $\mathbf{z}^l$ that is processed through a query-aware module. After removing the query token [9] (☒ in Fig. 2), this output is enhanced by adding the learnable embedding:

$$\mathbf{z}^l = \mathbf{z}^l + \phi(\mathbf{P}) \in \mathbb{R}^{T_v \times d_v}. \quad (11)$$

Adding this learnable embedding serves dual purposes. Firstly, it compensates for the distribution shift, as mentioned earlier. Secondly, the sinusoidal embedding enhances spatial position awareness within the vision model [8, 18], which is beneficial for relative position localization tasks, such as table question-answering in the WTQ dataset [28].

3.4. Training Objectives

Let θ represent optional fine-tune parameters of the LLM part (e.g., LoRA parameters [12]). Our training objective for efficiently injecting the query token into the vision model of the VLM is formulated as:

$$\min_{\psi,\phi,\theta} \mathcal{L}_{\text{Overall}} = \mathcal{L}_{\text{CE}} + \alpha\mathcal{L}_{\text{enp}}, \quad (12)$$

where $\mathcal{L}_{\text{CE}}$ is the negative log-likelihood loss for the predicted text tokens of the fine-tuning dataset, and α is a hyper-parameter balancing the effect of entropy regularization on the vision attention blocks.

During inference, neither the fuse nor defuse steps is employed. Instead, only the single query q_{eos} is used as $\psi(q_{\text{eos}})$. Since the query-agnostic module is independent of the query tokens, after training, the matrix $\mathbf{P}$ is processed through ϕ and saved as a bias matrix, significantly reducing computational overhead.

4. Experiments

4.1. Settings

VLM models. To demonstrate the generalization of our proposed method, we selected two SoTA large VLMs: mPLUG-Owl2 [43] and Qwen-VL-Chat [3]. Both models have approximately 7 billion parameters. The former model, mPLUG-Owl2, was trained on various VQA datasets and has shown SoTA performance on OCR tasks. The latter, Qwen-VL-Chat, in addition to being trained with VQA datasets, has been trained with a variety of text-centric datasets such as Doc-VQA [26], TextVQA [30], OCRVQA [27], and ChartQA [24].

Datasets and evaluation metrics. During the fine-tuning phase, we intentionally excluded datasets that were used in the pre-training phase (e.g., DocVQA, TextVQA, OCR-VQA, ChartQA) to avoid redundancy and ensure the robustness of the fine-tuning process. Consequently, the two VLMs, mPLUG-Owl2 and Qwen-VL-Chat, were fine-tuned with different text-centric datasets, including both dense text datasets and scene text datasets as outlined in Table 2. The dense text datasets, which primarily feature images captured from documents, include InfoVQA [25], KLC [31], and WTQ [28]. Scene text datasets feature images from natural scenes and include VizWiz [10] and ST-VQA [4]. More details of the datasets are provided in Supplementary Material - Section A.

In alignment with previous studies [1, 3, 19, 40], we report the performance of the InfoVQA and ST-VQA datasets using the Average Normalized Levenshtein Similarity (ANLS). For the KLC dataset, we use the F_1 score as the evaluation metric. The evaluation metric for WTQ is accuracy, while the VizWiz dataset is evaluated using the VQA score.

Baselines. To fine-tune the VLMs on target datasets, Parameter-Efficient Fine-Tuning (PEFT) methods are preferred. We compare our methods with LoRA [12] and its advanced version, DoRA [22]. Additionally, we have re-implemented Visual Prompt Tuning (VPT) employing 50 tokens on the vision encoders, and re-implemented QA-ViT [9] on both VLMs.

Implementation details. During the fine-tuning stages, all images are resized to 448 × 448 pixels. We set the values of α and σ to 10^{-2} and 0.16, respectively, by carefully tuning them to achieve an optimal balance. We note that setting them too high can result in either excessive noise (in case of σ) or overly aggressive filtering of visual information (in case of α)). Experiments are conducted on a computational platform equipped with four NVIDIA A100 40GB GPUs. The models undergo the fine-tuning process with a cumulative batch size of 128, targeting a maximum of five epochs for both Qwen-VL-Chat and mPLUG-Owl2 models. Early stopping is applied when there is no decrease in validation loss. The training employs the AdamW optimizer [23], starting with a learning rate of 2×10^{-5} and a linear warm-up over 100 steps.

Dataset		Task	Training Set	Test Set
Dense text	InfoVQA [25]	VQA	24K	3.3K
	KLC [31]	KIE	14K	4.9K
	WTQ [28]	Table	14K	4.3K
Scence text	VizWiz [10]	VQA	21K	4.3K
	ST-VQA [4]	VQA	20K	6K

Table 2. Statistics of the fine-tuning and evaluation datasets. For ST-VQA dataset, we split its public training set into training and test set. As per the experimental settings outlined in Table 3, only a subset of each dataset (e.g., 1,000 training samples) is utilized for fine-tuning, while test sets are hold the same.

For the main experimental results, we fine-tuned two models using a small number of samples from each dataset, specifically, 1,000 and 2,000 samples. In the ablation studies, we demonstrate that our method remains superior given very limited or full datasets.

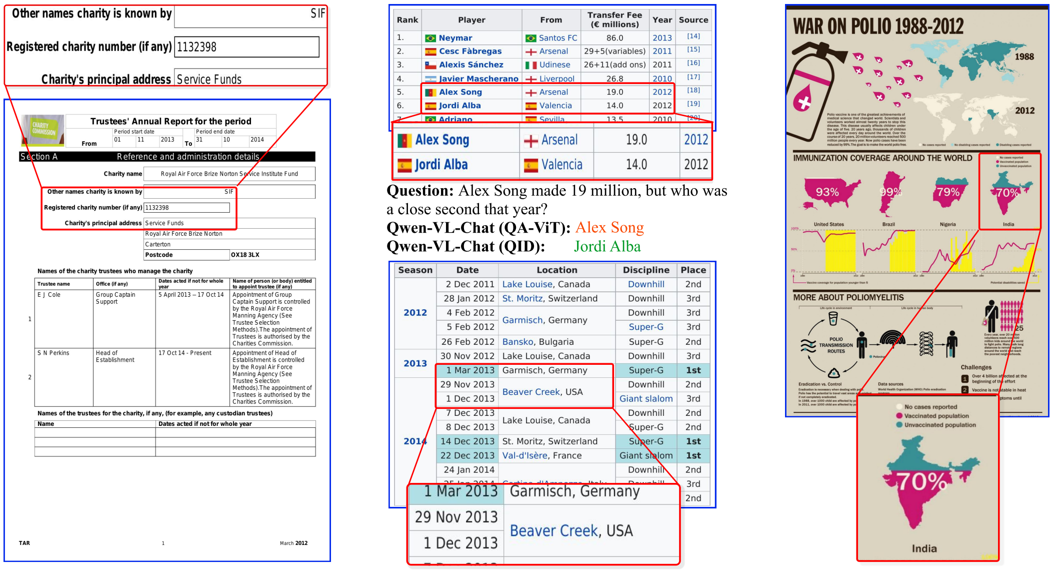
4.2. Comparisons with SoTA

The performance outcomes of various VLMs on different benchmarks related to visual document understanding are consolidated in Table 3. Specifically, for both models, mPLUG-Owl2 and Qwen-VL-Chat, our proposed method achieves the highest performance on average while introducing the smallest additional computational overhead on the Vision Transformer block, at 9×10^{-3} GFLOPs for mPLUG-Owl2 and 2×10^{-2} GFLOPs for Qwen-VL-Chat, respectively. With mPLUG-Owl2, we observe up to 0.7% and 0.7% improvements on average in the 1,000 and 2,000 sample settings, respectively, while Qwen-VL-Chat shows more substantial gains of 1.4% and 0.4% in the same settings, compared with the second best performance. Notably, these improvements are observed in both dense text and scene text datasets. In contrast, given the small number of fine-tuning samples, QA-ViT method reveals limited improvement on dense text datasets such as KLC and WTQ, and it even performs worse than solely fine-tuning with LoRA. This is attributed to the fact that QA-ViT utilizes all query tokens for interaction with the vision block, which becomes noisy and inefficient for learning in data-scarce scenarios. To ensure a fair and consistent comparison, we compare results within each model. Other baselines utilized higher resolutions and were fully trained on the datasets; we include them to provide the reader with more context regarding the current progress in solving the VDU task. Conclusively, across different experimental settings, our method consistently enhances the visual representations of the image encoder within VLMs in text-rich scenarios, thereby bolstering the performance of VDU with the smallest additional overhead in the inference phase.

Figure 4 showcases the qualitative results generated by the Qwen-VL-Chat model [3], comparing the version fine-tuned with QA-ViT [9] to our proposed QID approach across a varied collection of document images. More qualitative visualizations are provided in Supplementary Material - Section B. The results clearly demonstrate that our method is not only effective in extracting information from dense text documents (as shown in the left and right images) but also significantly enhances the reasoning capabilities of the VLMs (as evident in the middle images). The integration of a single query embedding with a query agnostic module, combined with our fuse and defuse training modules, significantly improves the efficacy of vision models. This enhancement facilitates a deeper understanding in text-rich environments for the VLMs.

Models	Resolution		+Ovh./ViTBlock	InfoVQA	KLC	WTQ	VizWiz	ST-VQA	*Avg.*
			GFLOP ↓	*ANLS* ↑	*F1* ↑	*ACC* ↑	*VQA Score* ↑	*ANLS* ↑	
UREADER [41]	(896 × 896)	*Fully trained*		42.2	32.8	29.4	-	-	-
Pix2Struct-L [17]	(1024 × 1024)			40.0	-	-	-	-	-
Pali-3 [7]	(1064 × 1064)			57.8	-	-	-	-	-
Qwen-VL-Chat† [3]	(448 × 448)			33.1	31.5	24.8	-	-	-
mPLUG-Owl† [42]	(448 × 448)			32.5	31.2	25.2	-	-	-
mPLUG-DocOwl [40]	-			38.2	30.3	26.9	-	-	-
Visfocus [1]	-			31.9	-	-	-	-	-
LLaVA+Vicuna+QA-ViT [9]	(336 × 336)			-	-	-	-	62.4	-
mPLUG-Owl2 [43]	(448 × 448)	*Zero-shot*		25.0	9.4	11.9	32.5	55.0	26.8
		PEFT on 1K							
		DoRA [22]		25.6	16.7	12.2	50.4	54.9	32.0
		LoRA [12]		25.5	15.7	12.6	51.5	54.7	32.0
		VPT [14]	+0.84	25.8	16.1	11.6	50.1	53.5	31.4
		QA-ViT [9]	+1.88	**26.7**	16.5	12.6	47.1	55.2	31.6
		QID (*ours*)	**+0.009**	**26.7**	**16.8**	**12.8**	**52.0**	**55.4**	**32.7**
		PEFT on 2K							
		DoRA [22]		25.9	16.4	12.9	51.6	55.2	32.4
		LoRA [12]		26.3	16.5	13.1	52.7	55.3	32.8
		VPT [14]	+0.84	26.0	17.1	12.1	51.3	54.6	32.2
		QA-ViT [9]	+1.88	25.6	16.7	12.6	52.5	55.4	32.6
		QID (*ours*)	**+0.009**	**27.3**	**17.4**	**13.4**	**53.9**	**55.7**	**33.5**
Qwen-VL-Chat [3]	(448 × 448)	*Zero-shot*		34.2	19.9	22.6	35.2	56.5	33.7
		PEFT on 1K							
		DoRA [22]		**34.6**	27.3	23.1	39.0	58.1	36.4
		LoRA [12]		34.4	26.6	23.3	38.5	58.1	36.2
		VPT [14]	+2.01	31.0	26.0	20.6	**43.3**	56.2	35.4
		QA-ViT [9]	+4.53	33.8	25.6	22.6	40.4	62.7	36.9
		QID (*ours*)	**+0.02**	34.2	**27.8**	**23.6**	42.1	**63.7**	**38.3**
		PEFT on 2K							
		DoRA [22]		34.5	27.8	23.4	39.2	58.0	36.6
		LoRA [12]		34.4	27.4	23.2	38.9	58.0	36.4
		VPT [14]	+2.01	31.9	26.9	20.8	**43.4**	56.8	36.0
		QA-ViT [9]	+4.53	34.5	28.4	24.5	42.0	64.2	38.7
		QID (*ours*)	**+0.02**	**34.8**	**28.8**	**24.9**	42.2	**65.1**	**39.1**

Table 3. **Experimental results across five datasets**. *Ovh./ViTBlock* denotes the computational overhead added to each Vision Transformer (ViT) block during the inference phase. † denotes models trained using the approach described in [19]. ↓ and ↑ indicate that lower and higher values are preferable, respectively. Our method offers a substantial improvement with modest overhead, with greater impact when dataset annotations require expert knowledge.



Question: What is the value of charity number?
Qwen-VL-Chat (QA-ViT): 1133238
Qwen-VL-Chat (QID): 1132398

Question: Alex Song made 19 million, but who was a close second that year?
Qwen-VL-Chat (QA-ViT): Alex Song
Qwen-VL-Chat (QID): Jordi Alba

Question: Where did this racer compete after Garmisch, Germany in 2013?
Qwen-VL-Chat (QA-ViT): Lake Louise, Canada
Qwen-VL-Chat (QID): Beaver Creek, USA

Question: What percentage of the Indian population is vaccinated against Polio in 2012?
Qwen-VL-Chat (QA-ViT): 79%
Qwen-VL-Chat (QID): 70%

Figure 4. **Qualitative results between QA-ViT and our QID**. Crucial regions are enlarged for better visualization. More visualizations are provided in Supp. Material - Section B.

Settings	InfoVQA	KLC	WTQ	VizWiz	ST-VQA	*Avg.*
	ANLS ↑	*F1* ↑	*ACC* ↑	*VQA Score* ↑	*ANLS* ↑	
w/o query-agnostic	34.60	27.71	22.96	39.89	63.46	37.72
w/o sinusoidal	34.60	27.31	22.45	40.97	63.96	37.85
w/o fuse step	33.84	27.44	23.50	39.02	63.92	37.54
w/o defuse step	34.77	27.60	23.58	37.94	62.90	37.36
with **q**	33.89	27.63	23.21	40.39	64.28	37.88
QID (***ours***)	34.18	27.81	23.56	42.08	63.69	**38.26**

Table 4. **Ablation studies**. Impacts of various proposed modules on the Qwen-VL-Chat models fine-tuned with 1,000 samples.

4.3. Ablation Study

Effect of different modules. Table 4 presents the effects of each module proposed in our paper: the query-agnostic module, spherical augmentation, and entropy regularization. Additionally, we experiment with all query token embeddings. The experiments are conducted using 1,000 samples on Qwen-VL-Chat model. Notably, without the query-agnostic module, the model performance decreases by 0.54% compared to the full QID model, due to the shift in vision representation caused by the incorporation of the query token as input. In the absence of noise augmentation and entropy regularization, the model typically performs worse, particularly on the KLC and VizWiz datasets, compared to the full QID model. This observation confirms that both fuse and defuse learning steps enable the model to learn from scarce data more efficiently. Lastly, using all query token embeddings (*with* **q**) tends to limit our ability to integrate fuse and defuse learning steps, and meanwhile to introduce redundant tokens in training, resulting in a 0.38% lower performance than our proposed QID. These findings validate the necessity of our learning strategies, which incorporate query tokens into vision models, thereby enhancing the comprehension of text-rich documents in VLMs.

Fine-tuning with extremely limited and full data. To demonstrate our method's effectiveness across learning conditions, we fine-tuned Qwen-VL-Chat using 500 samples per dataset and the full datasets, with consistent training settings except for a single epoch for full datasets. The results are shown in Table 5. In the extremely limited data scenario, with only 500 training samples, our method still achieved the best performance among all baselines, showing an average of 1.22% improvement over QA-ViT. Meanwhile, with full datasets, our QID achieved the highest performance across all five datasets, obtaining an average improvement of 1.38% compared to QA-ViT. Conversely, limitations of QA-ViT are revealed in some scenarios when trained with full datasets, typically underperforming compared to LoRA by 0.5% on KLC — a dense text dataset. These results highlight the superiority of our method over QA-ViT, not only in data-scarce regimes but entire datasets, in terms of enhancing the image encoder's ability to discern more effective cues in text-rich environments.

Method	InfoVQA	KLC	WTQ	VizWiz	ST-VQA	*Avg.*
	ANLS ↑	*F1* ↑	*ACC* ↑	*VQA Score* ↑	*ANLS* ↑	
PEFT on 500						
DoRA [22]	34.00	26.02	22.47	38.00	57.04	35.51
LoRA [12]	34.56	26.60	22.72	38.40	57.63	35.98
VPT [14]	31.00	24.96	19.34	42.92	55.76	34.80
QA-ViT [9]	33.24	25.98	21.60	39.72	61.82	36.47
QID (***ours***)	34.17	27.17	23.85	39.87	63.41	**37.69**
PEFT on Full						
DoRA [22]	34.56	28.78	24.25	40.19	58.89	37.33
LoRA [12]	34.55	29.15	24.27	40.51	58.50	37.40
VPT [14]	30.37	27.70	21.06	39.39	54.86	34.68
QA-ViT [9]	34.95	28.65	25.19	39.21	63.21	38.28
QID (***ours***)	35.24	30.17	26.02	41.48	65.39	**39.66**

Table 5. **Ablation studies**. Performance of our proposed method when integrated with the Qwen-VL-Chat model, trained on a very small number of tuning samples (500) as well as on the full dataset.

5. Limitation & Future Work

While our proposed QID method demonstrates superior performance across various datasets, we acknowledge two primary limitations that warrant further investigation. First, our approach depends on an external pre-trained CLIP model [20], constrained to 77 tokens, potentially limiting its ability to handle longer or more complex queries; we plan to mitigate this by exploring state-of-the-art solutions like Long-CLIP [44] and its variants to support extended text inputs. Secondly, our current experiments focus on single-hop QA tasks, whereas real-world VDU often involves multi-hop QA requiring reasoning across multiple document regions. Although our design—integrating query-aware and query-agnostic modules into the final ViT layers (Fig. 2)—efficiently avoids multiple forward passes per query, its effectiveness for multi-hop reasoning remains untested; we intend to investigate adaptive query propagation mechanisms to address this.

6. Conclusion

This paper presents **QID**, a novel fine-tuning approach of enhancing OCR-free Visual Document Understanding (VDU) for Vision-Language Models (VLMs) in data-scarce regimes. By integrating a single query vector into the vision encoder without modifying its core architecture, QID effectively directs attention to query-relevant visual regions in a computationally efficient manner. Our unique fuse and defuse learning steps strengthen the query representation in data-scarce settings, while our query-agnostic module ensures robust positional encoding, supporting the model's adaptability to various document layouts. Our experimental results demonstrate that VLMs equipped with QID achieve evident performance gains across various datasets compared to baseline models, particularly excelling for dense-text tasks, with minimal overhead. Future directions include exploring multi-turn query representations to enable interactive VDU and scaling QID for larger, more complex models in diverse application scenarios.

References

[1] Ofir Abramovich, Niv Nayman, Sharon Fogel, Inbal Lavi, Ron Litman, Shahar Tsiper, Royee Tichauer, Srikar Appalaraju, Shai Mazor, and R Manmatha. Visfocus: Prompt-guided vision encoders for ocr-free dense document understanding. In *The 18th European Conference on Computer Vision*. Springer, 2024. 2, 3, 6, 7

[2] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 993–1003, 2021. 1, 2

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1, 2, 5, 6, 7, 12

[4] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301, 2019. 6, 12

[5] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3828–3837, 2024. 3

[6] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 3

[7] Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*, 2023. 7

[8] Michael Dorkenwald, Nimrod Barazani, Cees GM Snoek, and Yuki M Asano. Pin: Positional insert unlocks object localisation abilities in vlms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13548–13558, 2024. 5

[9] Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben Avraham, Oren Nuriel, Shai Mazor, and Ron Litman. Question aware vision transformer for multimodal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13861–13871, 2024. 1, 2, 3, 4, 5, 6, 7, 8, 12

[10] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 6, 12

[11] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024. 1

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 4, 5, 6, 7, 8

[13] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091, 2022. 1, 2

[14] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 7, 8

[15] Geewook Kim, Teakgyu Hong, Moonbin Yim, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Donut: Document understanding transformer without ocr. *arXiv preprint arXiv:2111.15664*, 7(15):2, 2021. 1, 2

[16] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2426–2435, 2022. 3

[17] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR, 2023. 2, 7

[18] Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. *arXiv preprint arXiv:2402.12058*, 2024. 5

[19] Xin Li, Yunfei Wu, Xinghua Jiang, Zhihao Guo, Mingming Gong, Haoyu Cao, Yinsong Liu, Deqiang Jiang, and Xing Sun. Enhancing visual document understanding with contrastive learning in large visual-language models. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 15546–15555, 2024. 6, 7

[20] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023. 3, 4, 8

[21] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022. 4

[22] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*, 2024. 2, 6, 7, 8

[23] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[24] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 1, 5

[25] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 2, 6, 12

[26] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 1, 5

[27] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 1, 5

[28] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*, 2015. 5, 6, 12

[29] 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. 3, 4

[30] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 1, 5

[31] Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. Kleister: key information extraction datasets involving long documents with complex layouts. In *International Conference on Document Analysis and Recognition*, pages 564–579. Springer, 2021. 6, 12

[32] Kazem Taghva, Russell Beckley, and Jeffrey Coombs. The effects of ocr error on the extraction of private information. In *Document Analysis Systems VII: 7th International Workshop, DAS 2006, Nelson, New Zealand, February 13-15, 2006. Proceedings 7*, pages 348–357. Springer, 2006. 1

[33] Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19254–19264, 2023. 1

[34] 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. 3

[35] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 5

[36] Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. Docllm: A layout-aware generative language model for multimodal document understanding. *arXiv preprint arXiv:2401.00908*, 2023. 1

[37] Dai Wenliang, Li Junnan, Li Dongxu, Tiong Anthony, Zhao Junqi, Wang Weisheng, Li Boyang, Fung Pascale, and Hoi Steven. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[38] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1192–1200, 2020. 1, 2

[39] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*, 2020. 1

[40] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*, 2023. 1, 2, 6, 7

[41] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023. 1, 2, 7

[42] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 2, 7

[43] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051, 2024. 5, 7

[44] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, pages 310–325. Springer, 2024. 8

A. Fine-tuning Datasets

During our fine-tuning stage, VLMs are optimized using a variety of datasets with different tasks. These datasets, introduced in Table 2 , include InfographicVQA [25], Kleister Charity [31], WikiTableQuestions [28], VizWiz-VQA [10], and ST-VQA [4], and are briefly described as follows:

InfographicVQA (InfoVQA) [25]: This dataset is a collection of over five thousand infographic images, along with a large number of question-answer pairs. These infographics are sourced from various web domains and feature diverse layouts and designs. The InfographicVQA challenges vision language models to interpret and reason over complex visual documents, often necessitating understanding of graphical elements, data visualization, reasoning, and arithmetic skills.

Kleister Charity (KLC) [31]: This dataset consists of annual financial reports from UK charity organizations. The task involves key information extraction (KIE) such as charity names, addresses, charity numbers, and reporting dates. Primarily comprising scanned documents, this dataset poses challenges due to its length, diverse layout, and the necessity to interpret both text and structural features.

WikiTableQuestions (WTQ) [28]: This dataset includes question and answer pairs collected from thousands of HTML tables extracted from Wikipedia. The questions are designed to be complex, requiring multi-step reasoning and various data operations such as comparison, aggregation, and arithmetic computation.

VizWiz-VQA (VizWiz) [10]: Comprising a large number of question-answer pairs, this dataset features images captured by blind individuals using mobile phones and spoken questions about those images. Unique in terms of its image quality, which is often blurred, and the nature of its questions, a significant portion of the images are unanswerable due to poor image quality.

ST-VQA [4]: Designed specifically for understanding textual information within natural images, this dataset requires models to read and interpret scene text to accurately answer questions. It includes a large collection of images sourced from various public datasets such as COCO-text, Visual Genome, and ImageNet, challenging models to comprehend images across a wide range of scenarios and textual appearances within images.

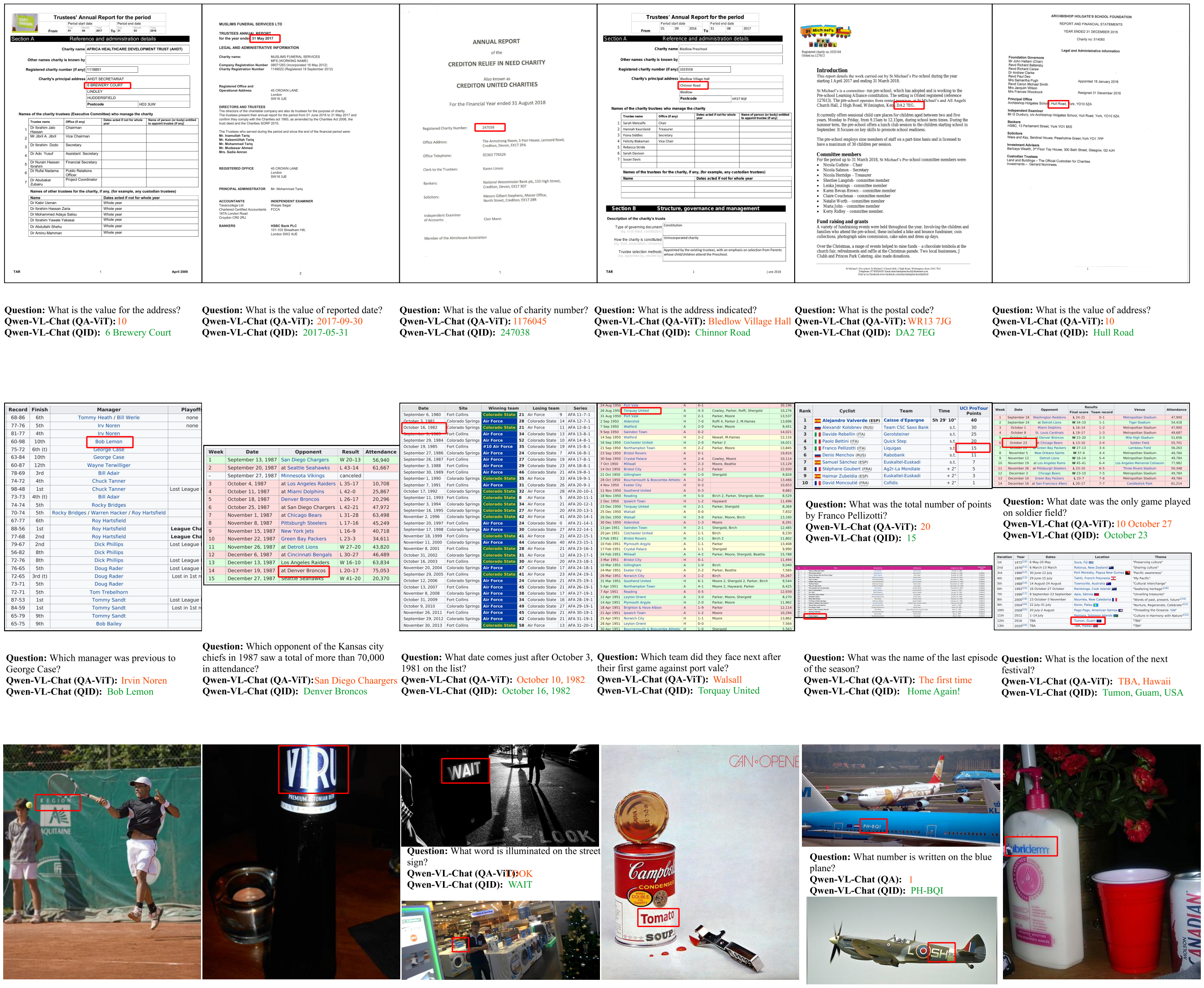
B. More Qualitative Results

Figure 5 presents additional quantitative results derived from various evaluation datasets comparing QA-ViT method [9] and our QID method, implemented with Qwen-VL-Chat [3]. These results highlight the effectiveness of our method in enhancing the vision model's ability to identify relevant visual cues and improve comprehension in both text-rich and natural scene environments.

Furthermore, we outline the limitations of our approach in Figure 6. Although our method aids in enhancing understanding of text-rich images, it does not significantly improve the model's reasoning and arithmetic capabilities. Consequently, our future research will focus on refining the model's ability to perform complex reasoning tasks more effectively in dense-text settings.

C. Broader Impact

The enhanced capabilities of vision-language models (VLMs) offer substantial promise for improving document comprehension in environments with extensive textual content. However, the interaction between question embeddings and vision representations remains relatively unexplored. Our approach encourages this interaction with limited fine-tuning samples while preserving the structural integrity of pre-trained VLMs. It also minimizes the necessity for extensive retraining, thereby reducing the computational resources required for deploying sophisticated AI solutions. Additionally, our method's efficiency with limited data can decrease both the time and costs involved in annotating large datasets, enhancing the accessibility and affordability of advanced document understanding technologies. We encourage the research community to further explore and adopt our QID for text-intensive tasks, anticipating significant benefits in various applications.



Question: What is the value for the address?
Qwen-VL-Chat (QA-ViT): 10
Qwen-VL-Chat (QID): 6 Brewery Court

Question: What is the value of reported date?
Qwen-VL-Chat (QA-ViT): 2017-09-30
Qwen-VL-Chat (QID): 2017-05-31

Question: What is the value of charity number?
Qwen-VL-Chat (QA-ViT): 1176045
Qwen-VL-Chat (QID): 247038

Question: What is the address indicated?
Qwen-VL-Chat (QA-ViT): Bledlow Village Hall
Qwen-VL-Chat (QID): Chinnor Road

Question: What is the postal code?
Qwen-VL-Chat (QA-ViT): WR13 7JG
Qwen-VL-Chat (QID): DA2 7EG

Question: What is the value of address?
Qwen-VL-Chat (QA-ViT): 10
Qwen-VL-Chat (QID): Hull Road

Question: Which manager was previous to George Case?
Qwen-VL-Chat (QA-ViT): Irvin Noren
Qwen-VL-Chat (QID): Bob Lemon

Question: Which opponent of the Kansas city chiefs in 1987 saw a total of more than 70,000 in attendance?
Qwen-VL-Chat (QA-ViT): San Diego Chaargers
Qwen-VL-Chat (QID): Denver Broncos

Question: What date comes just after October 3, 1981 on the list?
Qwen-VL-Chat (QA-ViT): October 10, 1982
Qwen-VL-Chat (QID): October 16, 1982

Question: Which team did they face next after their first game against port vale?
Qwen-VL-Chat (QA-ViT): Walsall
Qwen-VL-Chat (QID): Torquay United

Question: What was the total number of points by Franco Pellizotti?
Qwen-VL-Chat (QA-ViT): 20
Qwen-VL-Chat (QID): 15

Question: What date was the only game played on soldier field?
Qwen-VL-Chat (QA-ViT): 10 October 27
Qwen-VL-Chat (QID): October 23

Question: What was the name of the last episode of the season?
Qwen-VL-Chat (QA-ViT): The first time
Qwen-VL-Chat (QID): Home Again!

Question: What is the location of the next festival?
Qwen-VL-Chat (QA-ViT): TBA, Hawaii
Qwen-VL-Chat (QID): Tumon, Guam, USA

Question: What word is illuminated on the street sign?
Qwen-VL-Chat (QA-ViT): LOOK
Qwen-VL-Chat (QID): WAIT

Question: What number is written on the blue plane?
Qwen-VL-Chat (QA): 1
Qwen-VL-Chat (QID): PH-BQI

Question: What is written above the white square?
Qwen-VL-Chat (QA-ViT): AQUITAINE
Qwen-VL-Chat (QID): REGION

Question: What is the visible brand name of the product?
Qwen-VL-Chat (QA-ViT): VIRTU
Qwen-VL-Chat (QID): VITU

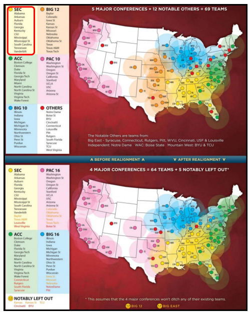
Question: What is the brand of washing machines behind the man?
Qwen-VL-Chat (QA-ViT): Indesit
Qwen-VL-Chat (QID): Samsung

Question: What is the word in red text on the front of the can?
Qwen-VL-Chat (QA-ViT): Campbell's
Qwen-VL-Chat (QID): Tomato

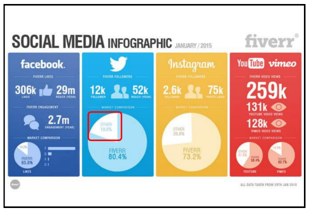
Question: What letters come after the red and blue target circle?
Qwen-VL-Chat (QA-ViT): k12
Qwen-VL-Chat (QID): SH

Question: What is the brand of the lotion?
Qwen-VL-Chat (QA-ViT): Eubriderm
Qwen-VL-Chat (QID): Lubriderm

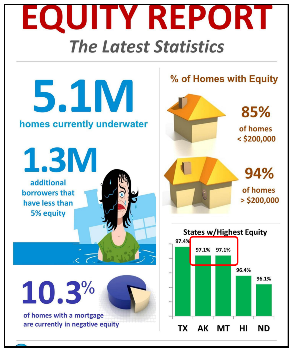
Figure 5. More qualitative results between between QA-ViT and our QID with Qwen-VL-Chat model. Image regions with answers are highlighted.



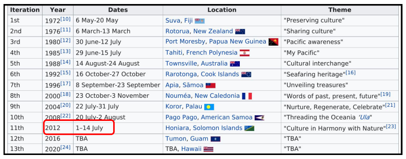
Question: How many teams did SEC conference have before re-alignment?
Qwen-VL-Chat (QID): 14
Ground Truth: 12



Question: What is the percentage market share others have in comparison to Fiverr in Twitter?
Qwen-VL-Chat (QID): 80.4%
Ground Truth: 19.6%



Question: How many states have 97.1% equity?
Qwen-VL-Chat (QID): 3
Ground Truth: 2



Iteration	Year	Dates	Location	Theme
1st	1972[10]	6 May-20 May	Suva, Fiji	"Preserving culture"
2nd	1976[11]	6 March-13 March	Rotorua, New Zealand	"Sharing culture"
3rd	1980[12]	30 June-12 July	Port Moresby, Papua New Guinea	"Pacific awareness"
4th	1985[13]	29 June-15 July	Tahiti, French Polynesia	"My Pacific"
5th	1988[14]	14 August-24 August	Townsville, Australia	"Cultural interchange"
6th	1992[15]	16 October-27 October	Rarotonga, Cook Islands	"Seafaring heritage"[16]
7th	1996[17]	8 September-23 September	Apia, Sāmoa	"Unveiling treasures"
8th	2000[18]	23 October-3 November	Nouméa, New Caledonia	"Words of past, present, future"[19]
9th	2004[20]	22 July-31 July	Koror, Palau	"Nurture, Regenerate, Celebrate"[21]
10th	2008[22]	20 July-2 August	Pago Pago, American Samoa	"Threading the Oceania *'Ula*"
11th	2012	1–14 July	Honiara, Solomon Islands	"Culture in Harmony with Nature"[23]
12th	2016	TBA	Tumon, Guam	"TBA"
13th	2020[24]	TBA	TBA, Hawaii	"TBA"

Question: What is the average length of the festival as of 2012?
Qwen-VL-Chat (QID): 10 years
Ground Truth: 14 days

Figure 6. Failure cases of QID on documents and questions require arithmetic and reasoning skills. Image regions with answers are highlighted.