DocVXQA: Context-Aware Visual Explanations for Document Question Answering

Mohamed Ali Souibgui [*1] **Changkyu Choi** [*2] **Andrey Barsky** [1] **Kangsoo Jung** [3] **Ernest Valveny** [1]
Dimosthenis Karatzas [1]

Abstract

We propose **DocVXQA**, a novel framework for visually self-explainable document question answering. The framework is designed not only to produce accurate answers to questions but also to learn visual heatmaps that highlight contextually critical regions, thereby offering interpretable justifications for the model's decisions. To integrate explanations into the learning process, we quantitatively formulate explainability principles as explicit learning objectives. Unlike conventional methods that emphasize only the regions pertinent to the answer, our framework delivers explanations that are *contextually sufficient* while remaining *representation-efficient*. This fosters user trust while achieving a balance between predictive performance and interpretability in DocVQA applications. Extensive experiments, including human evaluation, provide strong evidence supporting the effectiveness of our method. The code is available at https://github.com/dali92002/DocVXQA.

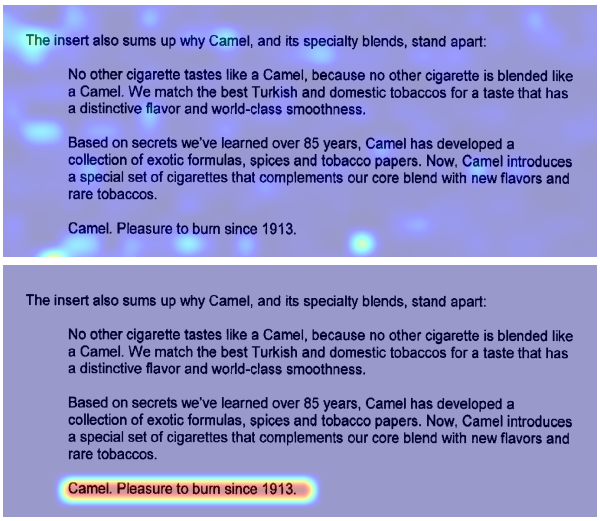


Figure 1. An illustration of the relevant regions in a DocVQA model (highlighted zones), produced by extracting the raw attention maps from the last layer (top) and by using our method (bottom) for the question " 'Pleasure to burn since 1913', Which cigarette's tagline is this?". Here, the answer given correctly by the model is "Camel".

1. Introduction

Document visual question answering (DocVQA) is essential for automatically extracting information from visually complex, multi-modal documents including invoices, reports, and contracts (Appalaraju et al., 2024; Blau et al., 2024; Tito et al., 2023; Mathew et al., 2021). This task requires interpreting the question and generating an answer through natural language, given a specific document. Recently, DocVQA models have increasingly leveraged large vision-language models for their ability to process both textual and visual modalities at scale (Wang et al., 2024; Zhao et al., 2024; Rasheed et al., 2024; Parashar et al., 2024).

Despite these promising advances in terms of utility, DocVQA relies on large, opaque neural network architectures that function as "black boxes," offering limited transparency into how they generate answers. This lack of transparency is particularly concerning in high-stakes domains such as finance (Wu et al., 2023), healthcare (Li et al., 2024), and law (Abdallah et al., 2023), where trust in AI-driven decision-making is crucial. In addition, existing explainable methods for DocVQA primarily rely on visualizing attention maps to highlight relevant regions influencing the model's predictions. However, as illustrated in Figure 1 (above), these attention-based explanations are often noisy or imprecise, failing to justify the reasoning behind the answer. Instead of offering clear insights, they tend to broadly highlight all occurrences of the answer without

[*]Equal contribution [1]Computer Vision Center, Universitat Autònoma de Barcelona, Spain [2]UiT The Arctic University of Norway, Norway [3]Inria, France. Correspondence to: Mohamed Ali Souibgui <msouibgui@cvc.uab.cat>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

contextualizing why it was selected.

To address this limitation and enhance the transparency of DocVQA models, we propose a novel *self-explainable* framework called **DocVXQA**, which learns to provide context-aware explanations to the answers in the form of relevance maps. Note that self-explainable models generate interpretable outputs by design, in contrast to post-hoc approaches that provide explanations retrospectively (Møller et al., 2024; Choi et al., 2024; Gautam et al., 2023; 2022). While these methods have been explored from a methodological perspective, their application in the context of DocVQA remains largely unexplored. To the best of our knowledge, our framework is the first self-explainable DocVQA model designed to generate visual explanations that align with textual answers. This approach enhances both transparency and user trust in model predictions.

To enhance transparency without disrupting established DocVQA models, we propose a strategy that learns generating explanations while maintaining compatibility with available pretrained models. We build upon the widely used Pix2Struct (Lee et al., 2023) model, trained for DocVQA task. With minimal alterations to the existing architecture– thus avoiding significant impacts on performance or the need for extensive retraining–our method first learns a mask over the document image as an explanation, and then uses the masked image as an input to the same network to answer the question. This sequential learning process ensures that the generated explanations directly contribute to the answer.

The core contribution of our DocVXQA lies in learning an effective mask designed to enhance human understanding. Grounded in a philosophical foundation of good explanation (Choi et al., 2024; Sokol & Flach, 2020), our approach aims to develop a method that learns mask representations that are both *contextually sufficient* and *representation-efficient*. To achieve this, we frame the trade-off between these objectives within the information bottleneck principle (Tishby & Zaslavsky, 2015), ensuring a balance between retaining relevant contextual information and minimizing redundancy in the representation. Additionally, to prevent overfitting in mask learning and to enhance the generalizability of the explanations, we integrate additional pretrained models (Faysse et al., 2024; Yu et al., 2024a), which infer the mask representation as a prior and interact with the main model to refine the explanation learning process, ensuring robust and reliable outcomes. This approach offers more precise and contextually relevant justifications. Notably, although our focus is on Pix2Struct (Lee et al., 2023), our method is inherently model-agnostic, allowing seamless integration with various DocVQA architectures.

Our main contributions are as follows: (1) We introduce DocVXQA, the first self-explainable DocVQA framework that learns to generate visually grounded context-aware explanations along with answer predictions. (2) We formalize explainability as an explicit learning objective with the information bottleneck principle. (3) We make our design model-agnostic and lightweight, thus, compatible with existing pretrained DocVQA models and requiring minimal architectural changes. (4) We conduct extensive experiments, including human evaluations, to validate the effectiveness of the proposed approach.

2. Related Work

2.1. Document visual question answering

Typical DocVQA models are reaching good performances using visual and textual features within transformer-based encoder-decoder architectures to generate answers (Appalaraju et al., 2024; Tito et al., 2023; Powalski et al., 2021; Xu et al., 2020). In this setup, text is extrtacted with OCR systems. However, despite their strong best-case performance, these methods are prone to errors introduced by OCR inaccuracies and face difficulty in domains that are more reliant on visual features.

In contrast, OCR-free end-to-end models (Lee et al., 2023; Aggarwal et al., 2023; Kim et al., 2022; Davis et al., 2022; Kim et al., 2021) predict answers directly from document images, utilizing pre-training objectives to interpret text without relying on OCR. More recently, Large Vision-Language Models (VLMs) like GPT-4 (Achiam et al., 2023) set new benchmarks in DocVQA performance, but their high computational requirements and closed-source nature limit their accessibility. There has also been a gradual emergence of smaller, open source models trained with instruction tuning (Wang et al., 2024; Zhang et al., 2023; Ye et al., 2023), but all these models fall short in providing explanations that support the reason behind their predictions. Among the OCR-free models, Pix2Struct (Lee et al., 2023) is designed for visually-situated language understanding tasks. It uses a pretraining strategy called screenshot parsing, converting masked web page screenshots into simplified HTML representations. The model incorporates variable-resolution input, preserving the aspect ratios for diverse document layouts. Thus, avoiding the significant reduction in original image resolution, which make it successful in performing DocVQA without requiring external OCR systems.

2.2. Explainability in vision and language

Current vision-and-language model explanations rely heavily on self-attention mechanisms to generate relevance maps, highlighting the regions that contribute to the model's decision (Bousselham et al., 2024; Chefer et al., 2021b;a). However, attention maps are often noisy (Abnar & Zuidema, 2020), making it challenging to extract meaningful insights into the model's reasoning. Recently, in the domain of

DocVQA, new approaches have been introduced to provide more fine-grained grounding by explicitly localizing answer regions (Mohammadshirazi et al., 2024; Zhou et al., 2024). These methods, however, require annotated answer locations during training, leading to the development of new datasets containing such annotations (Giovannini et al., 2025). While effective in the case of extractive DocVQA tasks (where the answer string is found explicitly in the document), these approaches introduce a significant annotation cost and remain limited in their ability to capture the context that explain why an answer is correct. Consequently, there remains a need for explanation techniques that not only identify specific answer locations but also provide a reasoning context beyond simple spatial answer grounding.

2.3. Learning to explain

Model explainability has gained significant attention in recent years, aiming to provide insights into neural network decision-making (Covert et al., 2021). A common approach is to generate saliency maps that highlight input regions relevant to predictions (Schulz et al., 2020; Selvaraju et al., 2017). In this context, the saliency map is a set of relevance scores over input pixels, intended to reveal the model's reasoning. Earlier explainability methods provide post-hoc explanations, i.e., get explanation without changing the trained model weights or architecture (Shrikumar et al., 2017; Selvaraju et al., 2017; Ribeiro et al., 2016; Bach et al., 2015). Gradient-based methods such as Grad-CAM (Selvaraju et al., 2017), LRP (Bach et al., 2015), and DeepLIFT (Shrikumar et al., 2017) calculate saliency maps through backpropagation of model output gradients to the input or feature space. While these methods are easy to implement, they often suffer from gradient shattering, leading to noisy and imprecise explanations. Another line of post-hoc explanation research is perturbation-based methods. This approach observes output changes by processing a set of perturbed images, each occluding different regions of the original image. Examples include LIME (Ribeiro et al., 2016), Occlusion (Zeiler & Fergus, 2014), and RISE (Petsiuk, 2018). While these methods often produce more reliable attributions, their effectiveness is constrained by the high computational cost of sampling a sufficient number of meaningful perturbations.

Unlike post-hoc approaches, self-explainable methods (Choi et al., 2024; Møller et al., 2024; Rudin, 2019; Alvarez Melis & Jaakkola, 2018) embed explainability directly into the training process, ensuring that models learn to generate interpretable explanations alongside predictions. Instead of relying on external attribution techniques, these models are designed to provide built-in reasoning. Among the various approaches to self-explainable deep learning methods, two main strategies stand out: the integration of mathematically formulated explainability principles into the objective function (Bang et al., 2021; Rudin, 2019; Alvarez Melis & Jaakkola, 2018), and architectural adjustments of the network to enhance transparency (Gautam et al., 2023; 2022; Alvarez Melis & Jaakkola, 2018). These approaches mark a shift from the traditionally opaque 'black-box' neural networks to a more transparent and understandable framework.

2.4. Information theoretic learning

Information theory provides a robust foundation for quantifying data representation through metrics such as entropy, divergence, and mutual information (Principe, 2010; Shannon, 1948). Building on this foundation, recent learning systems have embraced these principles (Skean et al., 2025; Yu et al., 2024b; 2019; Tishby & Zaslavsky, 2015) and evolved through various innovative approaches (Jenssen, 2024; Choi et al., 2024).

3. Method

3.1. Visually self-explainable DocVQA: Enhancing context awareness through a learnable mask

We propose DocVXQA, a novel framework for visually self-explainable document question answering. At a high level, this framework provides not only the answer to a question but also a visual explanation in the form of a relevance map (Müller, 2024; Zhang et al., 2024). The core innovation of our method lies in learning the relevance map-based explanations while simultaneously improving their reliability. To achieve this, we quantitatively formalize philosophically defined explainability principles (Sokol & Flach, 2020) and integrate them into a unified objective function, enabling end-to-end training. Our framework builds upon the widely recognized Pix2Struct architecture (Lee et al., 2023), adapting it effectively for this domain.

Minimality-sufficiency trade-off In formulating explainability principles, DocVXQA draws inspiration from the information bottleneck (IB) framework (Tishby & Zaslavsky, 2015), which seeks to construct a compact representation T of input data X that preserves only the information essential for predicting a target variable Y. This framework balances the trade-off between reducing the mutual information $I(X;T)$ to compress irrelevant data and maximizing $I(T;Y)$ to retain predictive power. To implement self-explainability within the conventional DocVQA setting, we strategically define the bottleneck representation T as the masked input, $T = X \odot M$, where $\odot$ represents the Hadamard product that masks out irrelevant information within X, and M is a learnable mask. This formulation provides a self-explaining mask in the form of region discovery, highlighting the most relevant regions of the input (Choi et al., 2024; Zhmoginov et al., 2021). With these terms, the

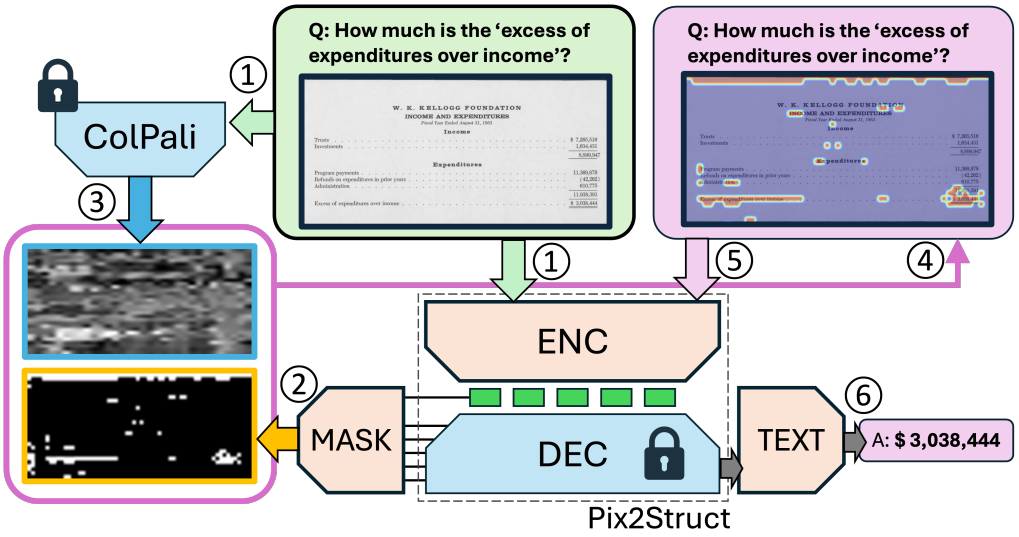


Figure 2. **Overview of the proposed DocVXQA framework:** ① The input question and the full document image are sent to both the pretrained Pix2Struct ENC-DEC and a pretrained ColPali model. ② The mask head (MASK) generates a learnable mask based on the decoder output and positional embeddings. ③ ColPali provides a mask prior, highlighting the relevant regions of the input document in relation to the question. ④ The learnable mask, guided by the mask prior, is combined with the original document to create the masked image, where only the highlighted parts are kept visible. ⑤ The masked image is processed through the Pix2Struct network. ⑥ The text head (TEXT) predicts the answer to the question based on the masked input.

objective function $\mathcal{L}_{\mathcal{IB}}$ is expressed as:

$$\min \mathcal{L}_{IB} = \beta I(X; X \odot M) - I(X \odot M; Y), \quad (1)$$

with hyperparameter β balancing the contrasting terms.

These two terms reflect the philosophical foundation of explanation, and the principles that define what an effective explanation should be (Choi et al., 2024; Sokol & Flach, 2020). We frame these principles as *minimality* and *sufficiency*, using them as learning criteria to guide the model as direct learning signals. *Minimality* emphasizes presenting only the most pertinent information for understanding a model's decision, while *sufficiency* ensures that explanations remain consistent and contribute decisively to predicting a correct answer.

Figure 2 provides an overview of our framework. We build on a pretrained Pix2Struct model as the baseline, updating the encoder, the newly added mask head, and the text head to enable the learning of a self-explainable mask representation. When a question and a document image are provided as inputs, the model first predicts a mask, introducing the minimality term $I(X; X \odot M)$ to encourage the removal of irrelevant information. The masked input is then processed by the model, together with the same question, to predict the answer based on the masked representation, captured by $I(X \odot M; Y)$. Note that the decoder of the Pix2Struct model remains frozen to ensure consistent text prediction.

As shown in the figure, our framework also incorporates a pretrained ColPali model (Faysse et al., 2024). The details of its role and integration are discussed in Sec. 3.4.

3.2. Learning sufficient explanation via cross-entropy

Upon learning the bottleneck representation, the mask M must forward contextually sufficient information to the model to answer the question by maximizing the mutual information $I(X \odot M; Y)$. Following (Amjad & Geiger, 2019), we approximate this by minimizing the cross-entropy loss $CE(\hat{Y}, Y)$,

$$\begin{aligned} \max I(X \odot M; Y) &= \underbrace{H(Y)}_{\text{constant}} - H(Y \mid X \odot M) \\ \iff \min H(Y \mid X \odot M) &\simeq CE(Y; \hat{Y}), \end{aligned} \quad (2)$$

where $H(Y)$ and $H(Y \mid X \odot M)$ are the marginal and conditional entropies of Y, respectively. $\hat{Y}$ denotes the model prediction. Note that $H(Y)$ is constant because the label Y is fixed.

3.3. Learning minimal explanation via continuity loss with L1 norm

Minimality is achieved by reducing the mutual information between the full document input X and the masked input $X \odot M$, resulting in a representation-efficient mask.

Contrary to the belief that estimating mutual information is inherently challenging (Poole et al., 2019; Belghazi et al., 2018), recent studies (Skean et al., 2024; Yu et al., 2019) have demonstrated that it can be computed deterministically when combined with kernel density estimation (Yu et al., 2019; Giraldo et al., 2014). This approach to quantifying mutual information is deterministic and integrates seamlessly with gradient descent over mini-batches. However, despite extensive experimentation, this approach did not yield the expected results. A potential limitation may stem from the model's sensitivity to hyperparameter tuning, particularly when dealing with high-dimensional image data, which could have restricted its capacity to reach optimal performance.

As an alternative, we propose a novel step by integrating a continuity loss, as known as anisotropic total variation (Bui et al., 2023; Johnson et al., 2016) combined with the L1 norm to enforce sparsity and smoothness in the generated relevance masks. The L1 norm promotes sparsity by penalizing the magnitude of mask values, ensuring that only the most essential regions of the input are highlighted, thereby producing concise and focused explanations. In parallel, the continuity loss enhances spatial coherence by minimizing abrupt transitions between adjacent regions in the mask. This is achieved by penalizing the L1 norm of horizontal and vertical differences in the reconstructed 2D relevance mask. Together, these complementary objectives enable the model to highlight minimal yet structurally coherent regions of importance. By fostering smoothness and minimality, this approach mitigates the risk of producing noisy or fragmented masks, ensuring that the explanations are not only interpretable but also intuitively aligned with human understanding.

3.4. Learning context-aware explanations through token interactions

Through our experiments with training the model using the objective function in Equation (1), we identified that the learned mask often overfits to regions directly corresponding to the answer (see Figure 7 in Appendix). This overfitting reduces the mask's utility, rendering it a mere visual reproduction of the textual answer rather than providing meaningful, context-aware insights. Such masks lack the ability to highlight broader, equally relevant information related to the question, leading to unreliable explanations.

To address this issue, we emphasize the need to regularize the mask-learning process, not only to mitigate overfitting but also to enable learning more generalizable explanations. Specifically, we aim to create context-aware explanations that identify and integrate the most relevant information across the broader input space. To achieve this, we incorporate a publicly available, pretrained vision-language understanding model as a source of prior knowledge to guide mask learning. We utilize its inference results of this model as prior information to facilitate interactive mask learning within our framework. While many models can be used for this purpose, we employ ColPali (Faysse et al., 2024) in our implementation. ColPali is a vision-language retrieval model. Its architecture integrates visual patches and textual tokens to generate multi-vector representations. These representations are then subjected to a late interaction matching mechanism (Liu et al., 2024; Khattab & Zaharia, 2020), which maximizes the interaction between each question token and its corresponding relevant tokens in the document image. This provides a degree of explainability by visualizing question-document interactions through heatmaps, offering insights into the model's decision-making process. Although these visualizations often fail to align with human intuition or provide comprehensive explanations due to their reliance on attention distributions, we use them as a prior mask and refine them within our DocVXQA framework to enhance relevance and interpretability.

3.5. Proposed learning objective

Our model achieves the following: (1) it learns to make predictions with masked input documents via cross-entropy loss $\mathcal{L}_{CE}$ (Sec.3.2), (2) it minimizes the highlighted regions to ensure concise and focused explanations via $\mathcal{L}_{L1}$ (Sec.3.3), and (3) it generates more context-aware explanations through tokens interaction with the ColPali signals via $\mathcal{L}_{MSE}$ (Sec.3.4). The complete objective function $\mathcal{L}_{VXQA}$ is defined as:

$$\begin{aligned}\min \mathcal{L}_{VXQA} =& \gamma \mathcal{L}_{MSE}(M_p; M) \\ &+ \beta \mathcal{L}_{L1}(X; X \odot M) + \mathcal{L}_{CE}(Y; \hat{Y}),\end{aligned} \tag{3}$$

where $\mathcal{L}_{MSE}$ aligns the learned mask with the prior M_p derived from ColPali, $\mathcal{L}_{L1}$ encourages sparsity in the highlighted regions, and $\mathcal{L}_{CE}$ represents the cross-entropy loss for prediction accuracy. The hyperparameters γ and β balance the contributions of each term. All components of the objective function are optimized simultaneously in an end-to-end fashion. As explanations M emerge naturally through this integrated optimization process, our framework inherently exhibits *self-explainability*.

3.6. Postprocessing

For a more human understandable relevance maps, a postprocessing stage is applied, consists in enclosing the connected relevance regions with bounding boxes and keep only k boxes with high relevance score. This step, applied only in the inference stage, is further detailed in Appendix A.1.

Method	Mask Thresh.	DocVQA (Mathew et al., 2021)			PFL-DocVQA (Tito et al., 2024)		
		Acc. ↑	ANLS ↑	Pix. Ratio ↓	Acc. ↑	ANLS ↑	Pix. Ratio ↓
Pix2Struct, unmasked (Lee et al., 2023)	–	0.56	0.68	1	0.80	0.92	1
Our model, unmasked	–	0.51	0.65	1	0.57	0.79	1
Raw Attention	0.05	0.34	0.46	0.38	0.06	0.13	0.27
	0.10	0.20	0.32	0.20	0.04	0.10	0.17
	0.25	0.07	0.13	0.08	0.01	0.04	0.09
	0.50	0.03	0.08	0.04	0.00	0.01	0.05
Attention Rollout (Abnar & Zuidema, 2020)	0.01	0.03	0.07	0.04	0.00	0.02	0.03
	1×10^{-5}	0.06	0.11	0.07	0.00	0.03	0.05
Grad-CAM (Selvaraju et al., 2017)	0.50	0.01	0.05	0.18	0.00	0.03	0.16
ColPali+Pix2Struct (Faysse et al., 2024)	0.50	0.38	0.50	0.23	0.28	0.39	0.18
Ours	0.70	**0.38**	**0.54**	**0.23**	**0.43**	**0.66**	**0.22**
	0.80	0.30	0.46	0.11	0.33	0.54	0.13

Table 1. Evaluation of various explainability methods under different mask threshold settings on the DocVQA task. The results demonstrate the trade-offs between the interpretability and utility of the different approaches. The unmasked approaches are set for reference as upper bounds for utility.

4. Experiments and Results

4.1. Experimental Setup

We posit that a good explanation map should highlight all the critical regions necessary to answer a question while minimizing irrelevant areas correctly, thus being both sufficient and minimal. We experimentally compare our method against several baseline techniques to evaluate both qualities. The evaluation process is structured as follows: explanation masks are generated using different methods, then postprocessed and binarized to keep only the relevant information when applied to the original document image. After that, masks are applied and the resulting image is passed to our fine-tuned Pix2Struct to predict the answer. Predicted answer is compared with the ground truth (GT) using accuracy and average normalized Levenshtein similarity (ANLS) to assess the sufficiency. Furthermore, the proportion of highlighted relevant regions in the mask relative to the total image area, referred to as the pixel ratio, is calculated to measure mask minimality. An ideal method achieves a favorable trade-off between explanation utility (accuracy and ANLS) and minimality (pixel ratio). The experiments are done on two datasets, DocVQA (Mathew et al., 2021) and PFL-DocVQA (Tito et al., 2024).

4.2. Baselines

We compare our approach against three categories of baseline methods:

- Attention Baselines: In this category, we include both raw attention (obtained from Pix2Struct's last layer) and attention rollout (Abnar & Zuidema, 2020), which use aggregated attention weights across all layers to provide a more holistic view of the model's focus.
- Gradient Baselines: We implemented Grad-CAM (Selvaraju et al., 2017), adapting it for the autoregressive sequence output layer by applying it to each output token from the decoder and aggregating the resulting maps.
- Retrieval Baseline: This approach utilizes ColPali (Faysse et al., 2024), a retrieval-based framework. Here, the input (question + answer) is first processed by ColPali on its own to generate a heatmap as the explanation. This heatmap is then applied to the image, which is subsequently passed to the Pix2Struct model for prediction. We call this baseline ColPali+Pix2Struct.

4.3. Results

Quantitative Evaluation The results obtained are summarized in Table 1, where different methods are evaluated to balance utility and explainability. As a reference for upper bound utility, we include results for unmasked images using the original Pix2Struct model and our trained model. Although our model exhibits a slight decrease in performance, this can be attributed to the challenge of simultaneously managing both masked and clear images during training.

Our approach stands out among the explainable methods, achieving the best utility in terms of accuracy and ANLS while maintaining a modest pixel ratio. This highlights the method's ability to focus on the most critical regions of the

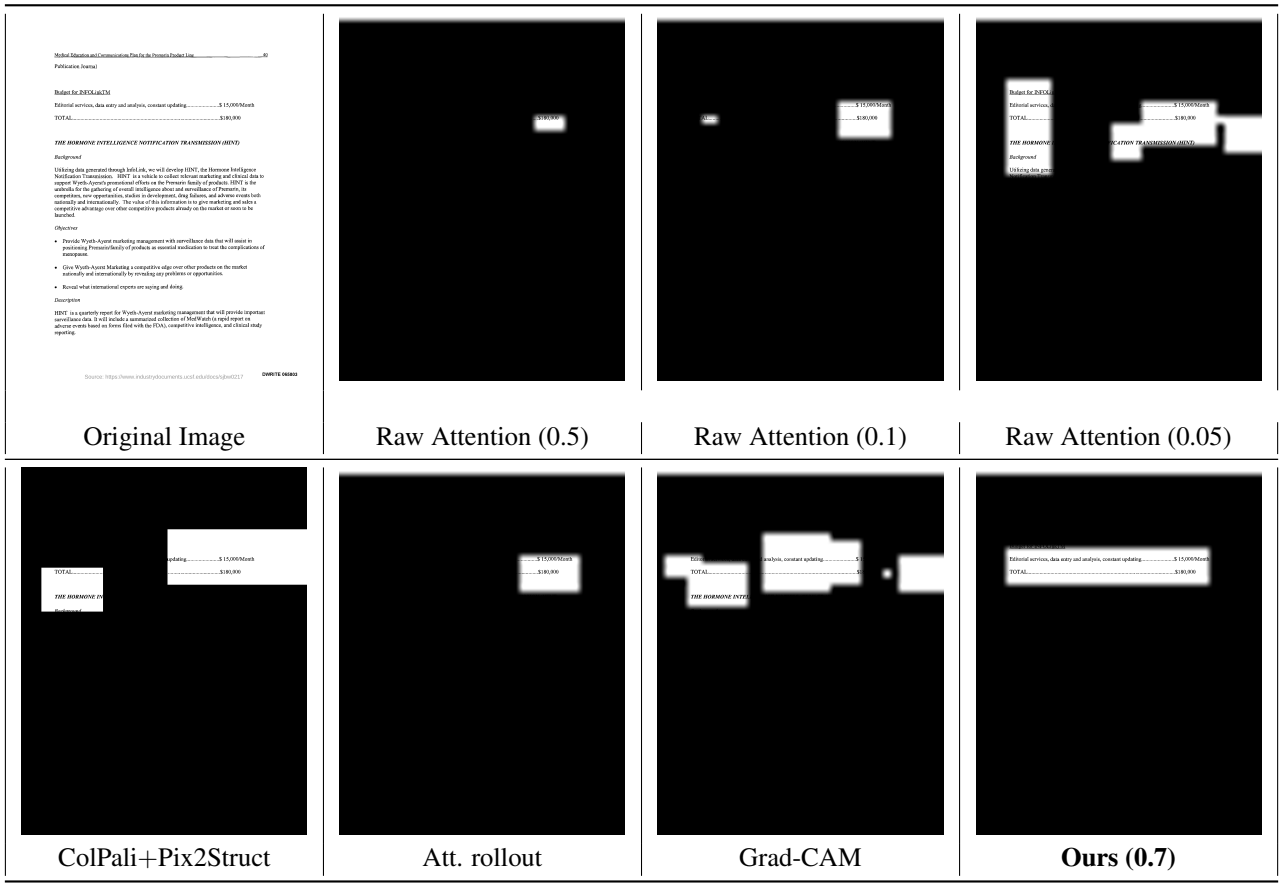


Figure 3. A comparison of explanations generated by different methods for the question **"What is the total amount?"** with the model's answer being **"$180,000"**. Relevance maps with different (thresholds) are applied to the input image to keep only the relevant regions. Best viewed at high zoom. Additional qualitative results across diverse contextual scenarios are provided in Appendix B.3.

input image, providing concise yet effective explanations. In contrast, raw attention exhibits reasonable performance when the mask threshold is set very low, as this results in large pixel ratios that cover a significant portion of the image. However, as the threshold increases, raw attention's utility metrics degrade significantly, suggesting it struggles to maintain performance when forced to operate with smaller, more answer-overfitted explanations. Attention rollout, while offering a more comprehensive aggregation of attention weights across layers, performs poorly overall. Its failure to accurately localize relevant regions is likely due to the propagation of errors (missing the contextual regions) across layers, which focused the relevant regions only around the answer. Grad-CAM, despite its popularity, also shows weak performance in this context. This can be attributed to the challenges of adapting Grad-CAM to autoregressive sequence models, where token-level aggregation may reduce its ability to generate precise explanations. Finally, the ColPali+Pix2Struct approach demonstrates competitive utility metrics with reasonable pixel ratios on the DocVQA dataset, however, our method achieves even better utility at the same ratio, as well as significantly better performance on the PFL-DocVQA dataset, which further shows the efficacy of our approach.

Qualitative Evaluation To reflect the results presented in Table 1, we illustrate the relevance masks generated by different methods in Figure 3. Each method was tasked with explaining the model's response to the question *"What is the total amount?"*, where the predicted answer is *"$180,000"*. The explanation masks are overlaid on the input image to highlight regions deemed relevant for the prediction. From the visualizations, it is evident that our method produces explanations that are both precise and comprehensive. Compared to baseline methods, such as ColPali+Pix2Struct, Attention Rollout, and Raw Attention, our approach successfully isolates all the textual elements contributing to the final answer in a relevance map that is easy to visually process by a human, while avoiding irrelevant regions. Notably, Grad-CAM focuses on certain parts of the document that do not contain all the required information to answer the question by the model.

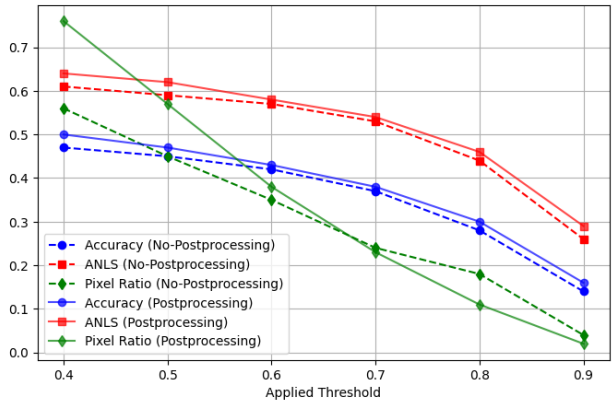


Figure 4. Performance of our method with and without postprocessing, under different thresholds applied to the relevance masks.

When these masked images were fed back into the model, all methods except Grad-CAM produced the correct answer. However, this does not necessarily indicate good relevance maps. In such scenarios, highlighting only any number will lead the DocVQA to use it as an answer, even without evidence that it corresponds to the "total". Our method leverages the most clear and *context-aware* explanation. This results in an accurate visualization of the reasoning process, which enhances users' ability to understand the model's decision-making, and fosters greater trust and confidence in the system.

Effect of Post-processing To study the impact of postprocessing and thresholding on the resultant masks, we perform the experiment presented in Figure 4. The experiment highlight the trade-offs between precision and coverage when applying different thresholds and when using postprocessing. Lower thresholds achieve higher Accuracy and ANLS but at the cost of significantly larger Pixel Ratios, indicating less precise localization of relevance regions. As the threshold increases, Pixel Ratio decreases, leading to more focused masks but with a corresponding decline in Accuracy and ANLS. When postprocessing is applied, the results show an improvement in the accuracy and ANLS with a reduction in Pixel ratio with a threshold ≥ 0.7. This improvement underscores the importance of postprocessing as a critical step in refining the outputs of our method, balancing relevance and mask quality.

Learning Objectives In our method, three learning objectives are used during training: Sufficiency (S), Minimality (M), and Token Interactions (TI). To evaluate their individual and combined contributions to performance, we conducted an ablation study, as shown in Table 2. The first row, where only the sufficiency loss is applied, achieves the highest Accuracy and ANLS but has the largest Pixel Ratio.

S	M	TI	Acc. ↑	ANLS ↑	Pixel Ratio ↓
✓			0.51	0.65	1
✓	✓		0.19	0.36	0.02
✓	✓	✓	0.38	0.54	0.23

Table 2. Ablation study on the impact of different learning objectives. S: sufficiency, M: minimality, TI: Token Interactions.

Method	Context ↑	Clarity ↑
Raw Attention (0.25)	2.90 ± 0.48	2.75 ± 0.42
Raw Attention (0.50)	2.26 ± 0.54	2.34 ± 0.60
ColPali+Pix2Struct	3.97 ± 0.25	3.02 ± 0.42
Ours (0.7)	$\mathbf{4.49 \pm 0.26}$	$\mathbf{3.56 \pm 0.41}$

Table 3. Results of the human evaluation of the explanation quality for different methods. Participants rated on a scale of 1 to 5.

This indicates that the model identifies the entire image as relevant, maximizing coverage but lacking precision. When both sufficiency and minimality losses are employed, the Pixel Ratio drastically decreases to 0.02, demonstrating the effectiveness of minimality in constraining the mask to small regions that are necessary to answer the question. However, this extreme reduction leads to overfitting, as the model disregards accompanying context that may be important for generalization. Consequently, both Accuracy and ANLS significantly decrease on the test data, highlighting a trade-off between precision and informativeness. Finally, the inclusion of token interactions loss (S+M+TI) balances these trade-offs, improving mask coverage on the context leading to achieving significant gains in Accuracy and ANLS.

Human Evaluation The goal of developing explainable DocVQA models is to enhance human trust in these systems. To experimentally assess the quality of the explanations, we conducted a human evaluation study with 42 impartial participants, blind to the details of our method. We used 10 example document images. For each image, participants were shown a corresponding question, the ground truth (GT) answer, and four different explanation masks generated from relevance maps of the top-performing methods seen in Table 1, based on the trade-offs between Accuracy, ANLS, and Pixel Ratio–this includes ColPali, raw attention, and our own method. Participants were asked to evaluate each explanation on two key criteria, rated on a scale of 1 to 5:

- **Context-awareness of the explanation:** Participants responded to the question: "How confident are you (1-5) to answer the question using only this masked image?" This question assesses whether the explanation directly supports the answer and provides all the necessary context

to answer.

- **Clarity of the explanation:** Participants responded to the question: "Does the masked image include only all the necessary information in a clear, concise, and comprehensive way to fully explain the model's reasoning? Rate on a scale (1-5)." This question evaluates the extent to which the explanation provides thorough, focused and human comprehensive coverage of the relevant content.

The results of the human evaluation are presented in Table 3. As shown, our method outperforms all baselines in both context-awareness and clarity, achieving the highest scores, compared to ColPali and Raw Attention variants. These results highlight our method's ability to generate clear explanations that enable participants to confidently answer questions without the original image, while also providing comprehensive and focused masks that include only the necessary information. Thus, using our method enhances user trust and understanding of the model's predictions.

Furthermore, we conducted another human preference study to directly compare the masks of our method against the top-performing baseline in the previous study, that is ColPali+Pix2Struct. In total, 12 participants evaluated 21 randomly selected question-answer pairs (252 trials in total). As a result, our method was preferred in 163 trials (64.7%; 95% CI [58.4%, 70.6%], $p << 0.001$), with all participants ($12/12$) favoring our approach overall. Thus, our method is offering a significantly more compact and interpretable explanations than ColPali+Pix2Struct.

Model Agnosticity Our DocVXQA framework is designed to be model-agnostic. To demonstrate its architectural flexibility, we implemented it with Donut (Kim et al., 2021) as an alternative to Pix2Struct. Notably, Donut and Pix2Struct differ fundamentally in how they incorporate the question. Pix2Struct renders the question directly onto the input image, while Donut tokenizes the question and uses it to condition the decoder during generation. To accommodate this difference, we adapt our method by modifying the mask head to also incorporate the encoded question. Qualitative results presented in Appendix B.4 illustrate that our method consistently identifies relevant regions across both backbones, validating its model-agnostic design.

5. Conclusion

This paper introduced DocVXQA, the first self-explainable model for DocVQA, enhancing interpretability by learning relevance maps that elucidate the model's decisions. Unlike conventional approaches that rely on post-hoc explanations or merely justify answers without deep contextual grounding, DocVXQA inherently generates visual and context-aware explanations to the answers, ensuring that the model's process remains transparent, interpretable, and aligned with human reasoning. This is achieved by formulating and including fundamental explainability principles directly into the learning process. Our extensive evaluations demonstrate the effectiveness of our framework.

Future research will explore the generalization of our method across diverse DocVQA architectures and datasets, as well as optimizing the trade-off between explainability and task performance. We envision DocVXQA as a stepping stone toward robust, transparent, and user-centric AI solutions in document intelligence.

Acknowledgments

We thank all participants for their contributions to the human preference study. This work has been supported by the Consolidated Research Group 2021 SGR 01559 from the Research and University Department of the Catalan Government, and by project PID2023-146426NB-100 funded by MCIU/AEI/10.13039/501100011033 and FSE+. This work has been funded by the European Lighthouse on Safe and Secure AI (ELSA) from the European Union's Horizon Europe programme under grant agreement No 101070617. The work of Changkyu Choi is supported by Visual Intelligence, a centre for research-based innovation funded by the Research Council of Norway and its consortium partners (RCN grant no. 309439).

Impact Statement

To the best of our knowledge, this work is the first to introduce explainability into the DocVQA applications. As DocVQA systems are increasingly deployed in high-stakes domains such as banking, healthcare, and public administration–where automated document understanding aims to reduce human involvement–the demand for transparent and trustworthy models becomes paramount. Our approach addresses this urgent gap by enabling interpretability of DocVQA models, which is essential for ensuring reliability, fostering user trust, and facilitating responsible deployment of AI in sensitive real-world contexts.

References

Abdallah, A., Piryani, B., and Jatowt, A. Exploring the state of the art in legal qa systems. *Journal of Big Data*, 10(1): 127, 2023.

Abnar, S. and Zuidema, W. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, 2020.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S.,

Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Aggarwal, K., Khandelwal, A., Tanmay, K., Khan, O. M., Liu, Q., Choudhury, M., Chauhan, H. H., Som, S., Chaudhary, V., and Tiwary, S. Dublin–document understanding by language-image network. *arXiv preprint arXiv:2305.14218*, 2023.

Alvarez Melis, D. and Jaakkola, T. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.

Amjad, R. A. and Geiger, B. C. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2225–2239, 2019.

Appalaraju, S., Tang, P., Dong, Q., Sankaran, N., Zhou, Y., and Manmatha, R. Docformerv2: Local features for document understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 709–718, 2024.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Bang, S., Xie, P., Lee, H., Wu, W., and Xing, E. Explaining a black-box by using a deep variational information bottleneck approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11396–11404, 2021.

Belghazi, I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, R. D. Mine: Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80, pp. 531–540. PMLR, 2018.

Blau, T., Fogel, S., Ronen, R., Golts, A., Ganz, R., Ben Avraham, E., Aberdam, A., Tsiper, S., and Litman, R. Gram: Global reasoning for multi-page vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15598–15607, 2024.

Bousselham, W., Petersen, F., Ferrari, V., and Kuehne, H. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3828–3837, 2024.

Bui, K., Lou, Y., Park, F., and Xin, J. Weighted anisotropic–isotropic total variation for poisson denoising. In *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 1020–1024. IEEE, 2023.

Chefer, H., Gur, S., and Wolf, L. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 397–406, 2021a.

Chefer, H., Gur, S., and Wolf, L. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 782–791, 2021b.

Choi, C., Yu, S., Kampffmeyer, M., Salberg, A.-B., Handegard, N. O., and Jenssen, R. Dib-x: Formulating explainability principles for a self-explainable model through information theoretic learning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7170–7174. IEEE, 2024.

Covert, I., Lundberg, S., and Lee, S.-I. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021.

Davis, B., Morse, B., Price, B., Tensmeyer, C., Wigington, C., and Morariu, V. End-to-end document recognition and understanding with dessurt. In *European Conference on Computer Vision*, pp. 280–296. Springer, 2022.

Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C., and Colombo, P. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*, 2024.

Gautam, S., Boubekki, A., Hansen, S., Salahuddin, S., Jenssen, R., Höhne, M., and Kampffmeyer, M. Protovae: A trustworthy self-explainable prototypical variational model. *Advances in Neural Information Processing Systems*, 35:17940–17952, 2022.

Gautam, S., Höhne, M. M.-C., Hansen, S., Jenssen, R., and Kampffmeyer, M. This looks more like that: Enhancing self-explaining models by prototypical relevance propagation. *Pattern Recognition*, 136:109172, 2023.

Giovannini, S., Coppini, F., Gemelli, A., and Marinai, S. Boundingdocs: a unified dataset for document question answering with spatial annotations. *arXiv preprint arXiv:2501.03403*, 2025.

Giraldo, L. G. S., Rao, M., and Principe, J. C. Measures of entropy from data using infinitely divisible kernels. *IEEE Transactions on Information Theory*, 61(1):535–548, 2014.

Jenssen, R. Map it to visualize representations. In *The Twelfth International Conference on Learning Representations*, 2024.

Johnson, J., Alahi, A., and Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 694–711. Springer, 2016.

Khattab, O. and Zaharia, M. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 39–48, 2020.

Kim, G., Hong, T., Yim, M., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., and Park, S. Donut: Document understanding transformer without ocr. *arXiv preprint arXiv:2111.15664*, 2021.

Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., and Park, S. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pp. 498–517. Springer, 2022.

Lee, K., Joshi, M., Turc, I. R., Hu, H., Liu, F., Eisenschlos, J. M., Khandelwal, U., Shaw, P., Chang, M.-W., and Toutanova, K. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pp. 18893–18912. PMLR, 2023.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.

Liu, Q., Guo, G., Mao, J., Dou, Z., Wen, J.-R., Jiang, H., Zhang, X., and Cao, Z. An analysis on matching mechanisms and token pruning for late-interaction models. *ACM Transactions on Information Systems*, 42(5):1–28, 2024.

Mathew, M., Karatzas, D., and Jawahar, C. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., and Jawahar, C. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1697–1706, 2022.

Mohammadshirazi, A., Neogi, P. P. G., Lim, S.-N., and Ramnath, R. Dlava: Document language and vision assistant for answer localization with enhanced interpretability and trustworthiness. *arXiv preprint arXiv:2412.00151*, 2024.

Møller, B. L., Igel, C., Wickstrøm, K. K., Sporring, J., Jenssen, R., and Ibragimov, B. Finding nem-u: Explaining unsupervised representation learning through neural network generated explanation masks. In *Forty-first International Conference on Machine Learning*, 2024.

Müller, R. How explainable ai affects human performance: A systematic review of the behavioural consequences of saliency maps. *International Journal of Human–Computer Interaction*, pp. 1–32, 2024.

Parashar, S., Lin, Z., Liu, T., Dong, X., Li, Y., Ramanan, D., Caverlee, J., and Kong, S. The neglected tails in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12988–12997, 2024.

Petsiuk, V. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

Poole, B., Ozair, S., Van Den Oord, A., Alemi, A., and Tucker, G. On variational bounds of mutual information. In *International Conference on Machine Learning*, pp. 5171–5180. PMLR, 2019.

Powalski, R., Borchmann, Ł., Jurkiewicz, D., Dwojak, T., Pietruszka, M., and Pałka, G. Going full-tilt boogie on document understanding with text-image-layout transformer. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pp. 732–747. Springer, 2021.

Principe, J. C. *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010.

Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R. M., Xing, E., Yang, M.-H., and Khan, F. S. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13009–13018, 2024.

Ribeiro, M. T., Singh, S., and Guestrin, C. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Schulz, K., Sixt, L., Tombari, F., and Landgraf, T. Restricting the flow: Information bottlenecks for attribution. *arXiv preprint arXiv:2001.00396*, 2020.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMlR, 2017.

Skean, O., Dhakal, A., Jacobs, N., and Giraldo, L. G. S. Frossl: Frobenius norm minimization for self-supervised learning. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Skean, O., Dhakal, A., Jacobs, N., and Sanchez Giraldo, L. G. Frossl: Frobenius norm minimization for efficient multiview self-supervised learning. In *European Conference on Computer Vision*, pp. 69–85. Springer, 2025.

Sokol, K. and Flach, P. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 56–67, 2020.

Tishby, N. and Zaslavsky, N. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pp. 1–5. IEEE, 2015.

Tito, R., Karatzas, D., and Valveny, E. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834, 2023.

Tito, R., Nguyen, K., Tobaben, M., Kerkouche, R., Souibgui, M. A., Jung, K., Jälkö, J., D'Andecy, V. P., Joseph, A., Kang, L., et al. Privacy-aware document visual question answering. In *International Conference on Document Analysis and Recognition*, pp. 199–218. Springer, 2024.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., and Mann, G. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., and Zhou, M. 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*, pp. 1192–1200, 2020.

Ye, J., Hu, A., Xu, H., Ye, Q., Yan, M., Xu, G., Li, C., Tian, J., Qian, Q., Zhang, J., et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Yu, S., Giraldo, L. G. S., Jenssen, R., and Principe, J. C. Multivariate extension of matrix-based rényi's α-order entropy functional. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2960–2966, 2019.

Yu, S., Tang, C., Xu, B., Cui, J., Ran, J., Yan, Y., Liu, Z., Wang, S., Han, X., Liu, Z., et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024a.

Yu, S., Yu, X., Løkse, S., Jenssen, R., and Principe, J. C. Cauchy-schwarz divergence information bottleneck for regression. In *The Twelfth International Conference on Learning Representations*, 2024b.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pp. 818–833. Springer, 2014.

Zhang, H., Torres, F., Sicre, R., Avrithis, Y., and Ayache, S. Opti-cam: Optimizing saliency maps for interpretability. *Computer Vision and Image Understanding*, 248:104101, 2024.

Zhang, Y., Zhang, R., Gu, J., Zhou, Y., Lipka, N., Yang, D., and Sun, T. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023.

Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.-M. M., and Lin, M. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhmoginov, A., Fischer, I., and Sandler, M. Information-bottleneck approach to salient region discovery. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III*, pp. 531–546. Springer, 2021.

Zhou, Y., Chen, Y., Lin, H., Yang, S., Zhu, L., Qi, Z., Ma, C., and Shan, Y. Doge: Towards versatile visual document grounding and referring. *arXiv preprint arXiv:2411.17125*, 2024.

Appendix

To ensure consistency, we continue the numbering of figures and tables from the main paper. The first references in the Appendix begin with Figure 5 and Table 4.

A. Implementation Details

A.1. Postprocessing

The relevance map produced by our method is passed throught a postprocessing step that envolve 2 stages: (1) background removal, (2) connected region bounding. In the following we detail each step, noting that for a fair comparison, these three steps are applied to all the methods presented in Table 1:

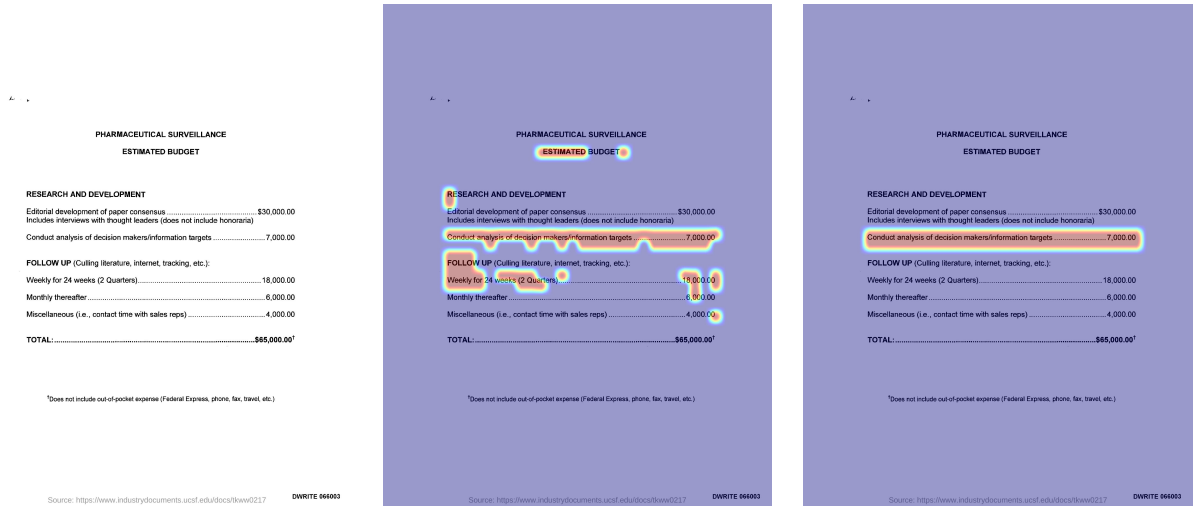
Background Removal. Since the output mask may highlight irrelevant background regions, we first filter out the highlighted regions that fall input image tokens that represent background. In other words, we discard tokens whose encoded patch variance are below a threshold of 0.01, as they are considered background in the document images, where in this domain the background is usually blank. This step ensures that only informative regions contribute to the final relevance map. In Figure 5, we show an extreme case of this scenario, where the produced relevance map contain a lot of background regions, after applying this postprocessing step, we can see that the quality of the map is significantly improved.

Original | Predicted Mask (with bg regions) | Predicted Mask (without bg regions)

Figure 5. Visualization of the background removal step in our methods. Question: "What is the estimated budget of 'conduct analysis of decision makers/ information targets' in research and development?". Model answer "$ 7,000.00".

Bounding connected regions. After filtering out background regions, the remaining mask effectively highlights relevant information, as shown in Figure 5. However, the raw mask may still appear fragmented and less intuitive for users. To enhance interpretability, we apply an additional postprocessing step that identifies connected components within the relevance map and encloses them in bounding boxes. We chose this approach because text in documents is best highlighted using bounding boxes, ensuring clear visibility and structured localization. Unlike other forms of heatmaps, which may appear vague or ambiguous, bounding boxes provide a precise, easily interpretable representation of the model's focus which facilitate its processing by human users.

After that, we rank these regions based on their confidence scores and retain only the top k most relevant boxes. This step, shown in Figure 6, ensures that the final explanation is both compact and user-friendly. In our study, we experimented with different values of k and found that $k = 3$ provided the best balance between conciseness and completeness.



Original | Predicted Mask | Predicted Mask (with full postprocessing)

Figure 6. Visualization of connecting the relevance regions in the heatmap and keeping only 1 region with top confidence score ($k = 1$). Question: "What is the estimated budget of 'conduct analysis of decision makers/ information targets' in research and development?". Model answer "$ 7,000.00".

A.2. Mask head

The implemented mask head is a fully connected neural network (MLP-based decoder) designed to process an input consists of the concatenation of the encoded tokens (features), decoder attentions and the positional encoding. The mask head generates a relevance score for each token, producing a relevance map that highlights the most important regions in terms of influence on the final answer. During training, the output of the mask head is passed through a Sigmoid activation function to ensure that the values are normalized between 0 and 1, applied to the input image, and then passed to the Pix2Struct component for the masked image prediction.

A.3. Hyperparameters

We summarize the values of the hyperparameters used in this paper. We not that the Pix2Struct component is first fine-tuned on the DocVQA dataset (Mathew et al., 2021), then our proposed model is trained according to the settings presented in Table 4. Note that in our work, we use the base version of Pix2Struct with 282M parameters. The model is composed of 12 encoder and 12 decoder layers with a hidden size of 768.

HYPERPARAMETER	VALUE
LEARNING RATE	1×10^{-7}
BATCH SIZE	5
OPTIMIZER	ADAMW
γ	0.5
β	5
THRESHOLD (k) FOR POSTPROCESSING	3

Table 4. HYPERPARAMETER SETTINGS

B. Additional Results

B.1. Selection of k.

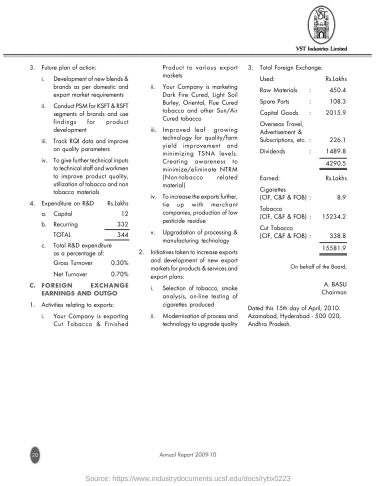
As illustrated in Figure 6, during the post-processing step, we retain only the top-k most relevant boxes. We conducted an ablation study to evaluate the impact of different k values and found that $k = 3$ offers the best trade-off between conciseness and completeness. The full results across various k values are reported in Table 5. As shown, $k = 3$ is considered the best choice for a good trade-off between interpretability and utility.

k	ACCURACY	ANLS	PIXEL RATIO
1	0.36	0.52	0.21
2	0.38	0.53	0.23
3	0.38	0.54	0.24
4	0.39	0.55	0.25
5	0.39	0.55	0.26
8	0.40	0.56	0.28
10	0.41	0.57	0.29
15	0.42	0.58	0.31
20	0.43	0.58	0.32
50	0.45	0.59	0.38
100	0.45	0.60	0.40

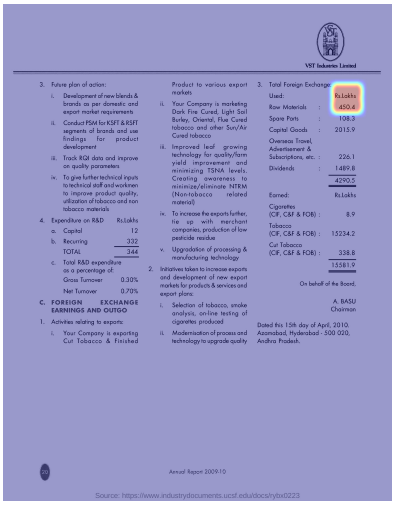
Table 5. A study on varying the number of top relevant boxes during post-processing.

B.2. The effect of using token interactions.

As we discussed in Section 3.4, training the model using the objective function in Equation (1) (without the use of token interactivity loss), lead to overfitting to the regions that directly corresponding to the answer. This overfitting reduces the mask's utility, rendering it a mere visual reproduction of the textual answer rather than providing meaningful, context-aware insights. This can be seen in Figure 7, where we show qualitative results of our method.

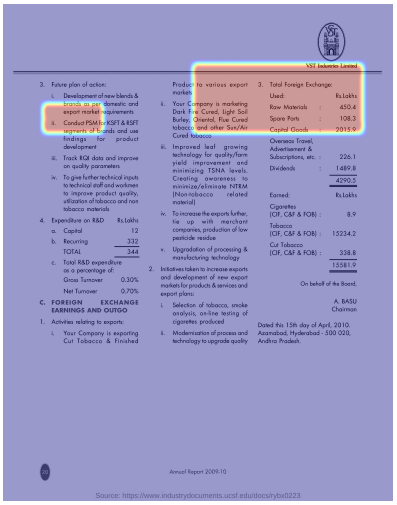


What is the total foreign exchange used for raw materials(Rs.lac)??
Answer: **“450.4”**



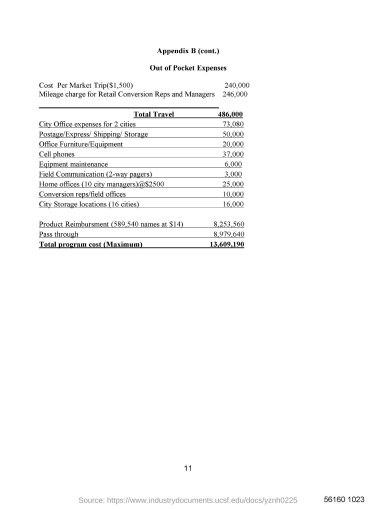
Clear Answer: “450.4”

Masked Answer: “450.4”

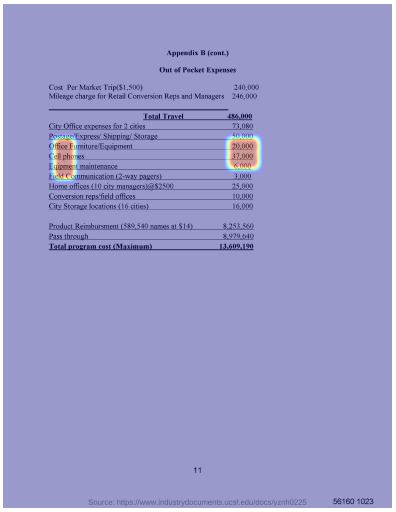


Clear Answer: “450.4”

Masked Answer: “450.4”

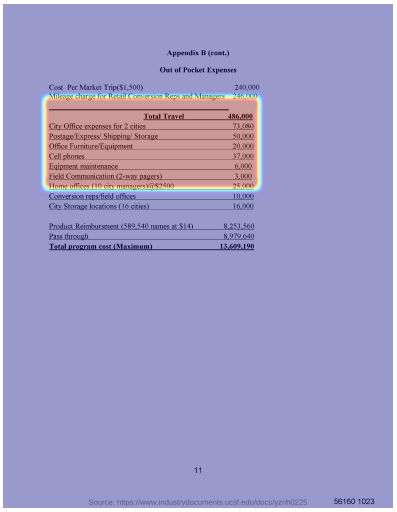


What is the equipment maintenance expenses?
Answer: **“6,000”**



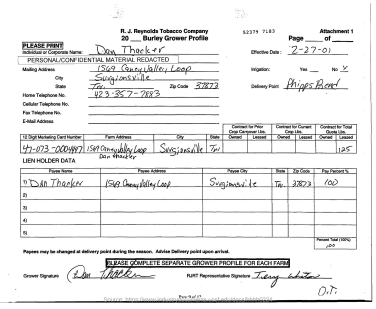
Clear Answer: “6,000”

Masked Answer: “6,000”

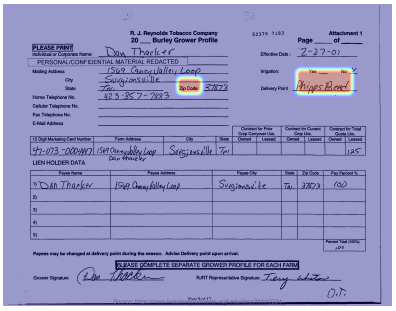


Clear Answer: “6,000”

Masked Answer: “6,000”

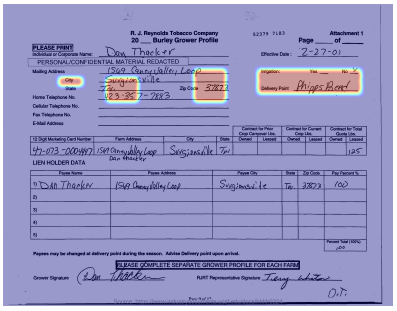


What is the delivery point mentioned in the form?
GT Answer: **“phipps bend”**



Clear Answer: “philips bank”

Masked Answer: “philips bank”



Clear Answer: “miops Reid”

Masked Answer: “shipping pound”

Figure 7. Visualization of the effect of using token interactions in our method. Left: Original Image. Middle: Masked image with relevance map learned without token interactions loss. Right: Masked image with relevance map learned with token interactions loss.

B.3. Additional qualitative results

In this section, we present the results of the applied relevant regions masks on the images (right) across diverse contextual scenarios, in comparison to the corresponding full images (left).

Complex layouts We test our approach on documents with complex layouts from the InfographicVQA dataset (Mathew et al., 2022).

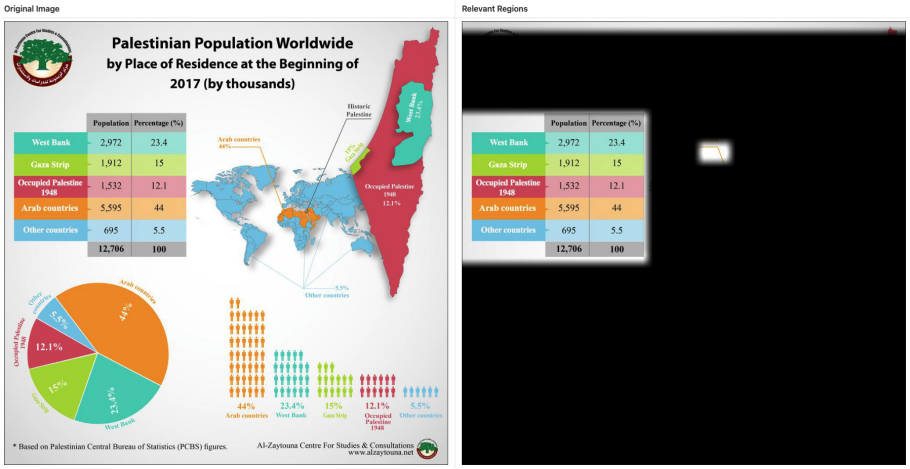


Figure 8. **Question:** What is the total percentage of Palestinians residing at places other than West Bank and Arab countries?
Answer: 32.6 % (correct).

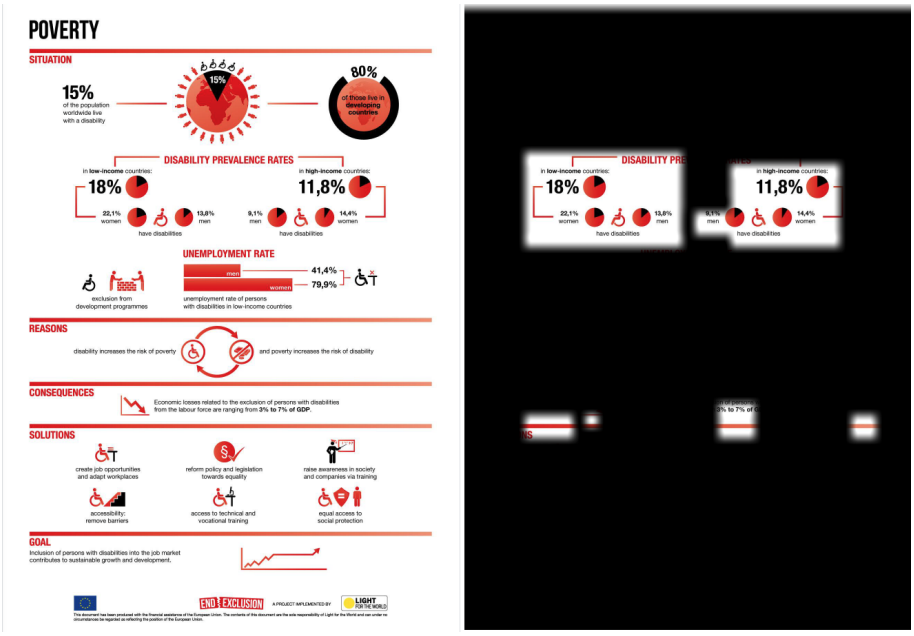
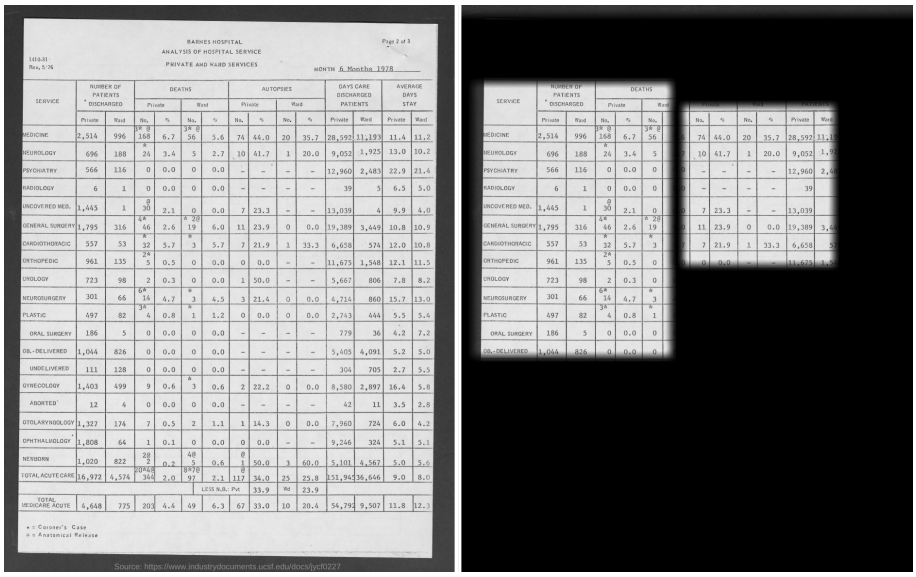


Figure 9. **Question:** What is the percentage of women with disabilities in low income countries?
Answer: 22.1 % (correct).

Densely populated documents We present results on documents containing dense textual content.

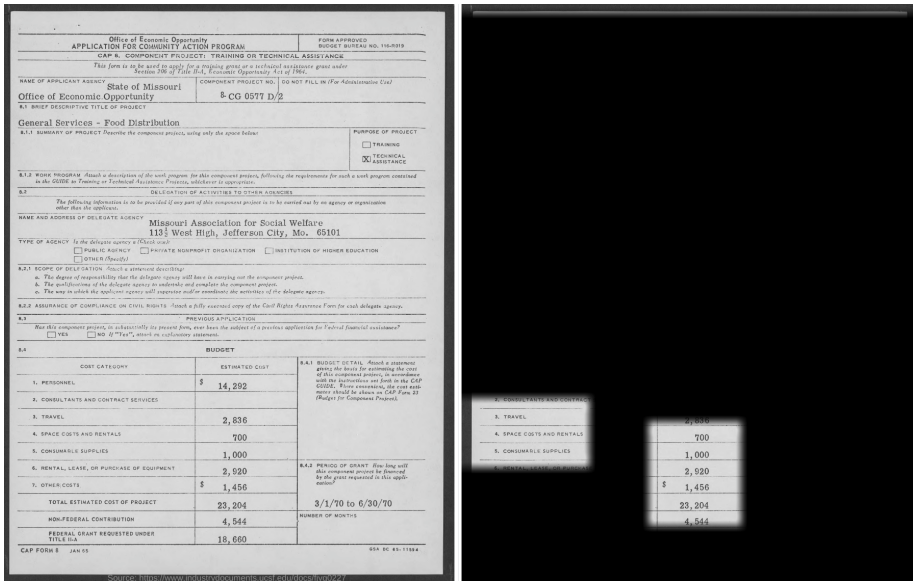


RARNES HOSPITAL
ANALYSIS OF HOSPITAL SERVICE
PRIVATE AND WARD SERVICES
MONTH 6 Months 1978

SERVICE	NUMBER OF PATIENTS DISCHARGED		DEATHS				AUTOPSIES				DAYS CARE DISCHARGED PATIENTS		AVERAGE DAYS STAY	
	Private	Ward	Private No.	Private %	Ward No.	Ward %	Private No.	Private %	Ward No.	Ward %	Private	Ward	Private	Ward
MEDICINE	2,514	996	1* 8 168	6.7	3* 8 56	5.6	74	44.0	20	35.7	28,592	11,193	11.4	11.2
NEUROLOGY	696	188	24	3.4	5	2.7	10	41.7	1	20.0	9,052	1,925	13.0	10.2
PSYCHIATRY	566	116	0	0.0	0	0.0	–	–	–	–	12,960	2,483	22.9	21.4
RADIOLOGY	6	1	0	0.0	0	0.0	–	–	–	–	39	5	6.5	5.0
UNCOVERED MED.	1,445	1	8 30	2.1	0	0.0	7	23.3	–	–	13,039	4	9.9	4.0
GENERAL SURGERY	1,795	316	4* 46	2.6	* 28 19	6.0	11	23.9	0	0.0	19,389	3,449	10.8	10.9
CARDIOTHORACIC	557	53	* 32	5.7	* 3	5.7	7	21.9	1	33.3	6,658	574	12.0	10.8
ORTHOPEDIC	961	135	2* 5	0.5	0	0.0	0	0.0	–	–	11,675	1,548	12.1	11.5
UROLOGY	723	98	2	0.3	0	0.0	1	50.0	–	–	5,667	806	7.8	8.2
NEUROSURGERY	301	66	6* 14	4.7	* 3	4.5	3	21.4	0	0.0	4,714	860	15.7	13.0
PLASTIC	497	82	3* 4	0.8	* 1	1.2	0	0.0	0	0.0	2,743	444	5.5	5.4
ORAL SURGERY	186	5	0	0.0	0	0.0	–	–	–	–	779	36	4.2	7.2
OB.-DELIVERED	1,044	826	0	0.0	0	0.0	–	–	–	–	5,405	4,091	5.2	5.0
UNDELIVERED	111	128	0	0.0	0	0.0	–	–	–	–	304	705	2.7	5.5
GYNECOLOGY	1,403	499	9	0.6	* 3	0.6	2	22.2	0	0.0	8,580	2,897	16.4	5.8
ABORTED	12	4	0	0.0	0	0.0	–	–	–	–	42	11	3.5	2.8
OTOLARYNGOLOGY	1,327	174	7	0.5	2	1.1	1	14.3	0	0.0	7,960	724	6.0	4.2
OPHTHALMOLOGY	1,808	64	1	0.1	0	0.0	0	0.0	–	–	9,246	324	5.1	5.1
NEWBORN	1,020	822	28 2	0.2	48 5	0.6	8 1	50.0	3	60.0	5,101	4,567	5.0	5.6
TOTAL ACUTE CARE	16,972	4,574	20468 344	2.0	8978 97	2.1	117	34.0	25	25.8	131,943	36,646	9.0	8.0
					LESS N.B. Pvt			33.9	Wd	23.9				
TOTAL MEDICARE ACUTE	4,648	775	203	4.4	49	6.3	67	33.0	10	20.4	54,791	9,507	11.8	12.3

* = Coroner's Case
# = Anatomical Release

Figure 10. **Question:** Under Private service how many patients were discharged in Neurology? **Answer:** 696 (correct).



Office of Economic Opportunity
APPLICATION FOR COMMUNITY ACTION PROGRAM
FORM APPROVED BUDGET BUREAU NO. 116-R019

CAP 6. COMPONENT PROJECT: TRAINING OR TECHNICAL ASSISTANCE

This form is to be used to apply for a ongoing grant or a technical assistance grant under Section 209 of Title IIA, Economic Opportunity Act of 1964.

NAME OF APPLICANT AGENCY: State of Missouri
Office of Economic Opportunity

COMPONENT PROJECT NO.: S-CG 0577 D/2

DO NOT FILL IN (For Administrative Use)

6.1 BRIEF DESCRIPTIVE TITLE OF PROJECT
General Services - Food Distribution

6.1.1 SUMMARY OF PROJECT *Describe the component project, using only the space below*

PURPOSE OF PROJECT
☐ TRAINING
☒ TECHNICAL ASSISTANCE

6.1.2 WORK PROGRAM *Attach a description of the work program for this component project, following the requirements for such a work program contained in the GUIDE to Training or Technical Assistance Projects, whichever is appropriate.*

6.2 DELEGATION OF ACTIVITIES TO OTHER AGENCIES

The following information is to be provided if any part of this component project is to be carried out by an agency or organization other than the applicant.

NAME AND ADDRESS OF DELEGATE AGENCY
Missouri Association for Social Welfare
1131 West High, Jefferson City, Mo. 65101

TYPE OF AGENCY *Is the delegate agency a (Check one)*
☐ PUBLIC AGENCY ☐ PRIVATE NONPROFIT ORGANIZATION ☐ INSTITUTION OF HIGHER EDUCATION
☐ OTHER *(Specify)*

6.2.1 SCOPE OF DELEGATION *Attach a statement describing:*
a. *The degree of responsibility that the delegate agency will have in carrying out the component project.*
b. *The qualifications of the delegate agency to undertake and complete the component project.*
c. *The way in which the applicant agency will supervise and/or coordinate the activities of the delegate agency.*

6.2.2 ASSURANCE OF COMPLIANCE ON CIVIL RIGHTS *Attach a fully executed copy of the Civil Rights Assurance Form for each delegate agency.*

6.3 PREVIOUS APPLICATION
Was this component project, in substantially its present form, ever been the subject of a previous application for Federal financial assistance?
☐ YES ☐ NO *If "Yes", attach an explanatory statement.*

6.4 BUDGET

COST CATEGORY	ESTIMATED COST
1. PERSONNEL	$ 14,292
2. CONSULTANTS AND CONTRACT SERVICES	
3. TRAVEL	2,836
4. SPACE COSTS AND RENTALS	700
5. CONSUMABLE SUPPLIES	1,000
6. RENTAL, LEASE, OR PURCHASE OF EQUIPMENT	2,920
7. OTHER COSTS	$ 1,456
TOTAL ESTIMATED COST OF PROJECT	23,204
NON-FEDERAL CONTRIBUTION	4,544
FEDERAL GRANT REQUESTED UNDER TITLE IIA	18,660

6.4.1 BUDGET DETAIL *Attach a statement giving the basis for estimating the cost of this component project, in accordance with the budget instructions in the CAP GUIDE. Where necessary, the cost estimates should be shown on CAP Form 23 (Budget for Component Project).*

6.4.2 PERIOD OF GRANT *How long will this component project be financed by the grant requested in this application?*
3/1/70 to 6/30/70
NUMBER OF MONTHS 4.544

CAP FORM 5 JAN 65 GSA BC 65-1184

Figure 11. **Question:** What is the estimated cost of Space costs and Rentals? **Answer:** 700 (correct).

Degraded quality We present results on documents with degraded visual quality.

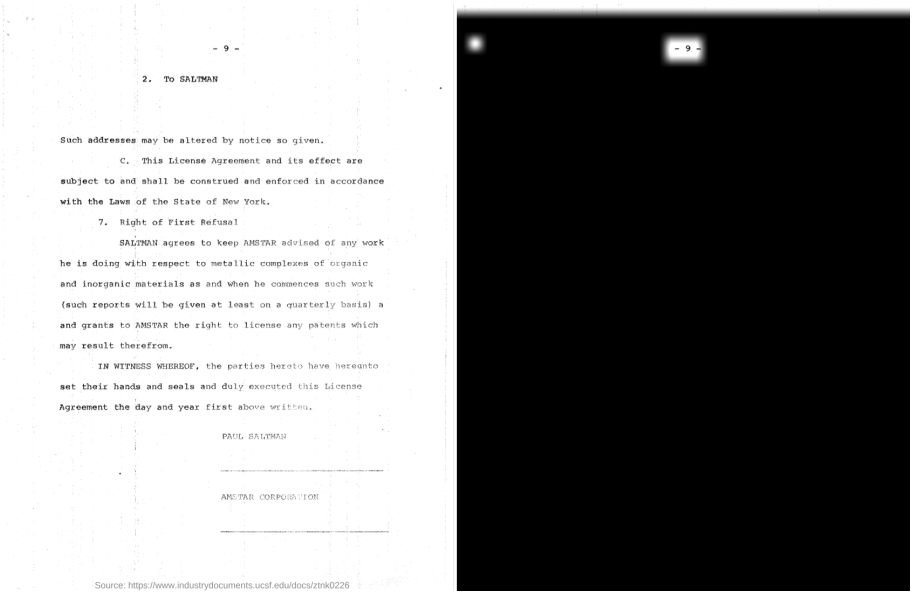


Figure 12. **Question:** What is the page number? **Answer:** 9 (correct).

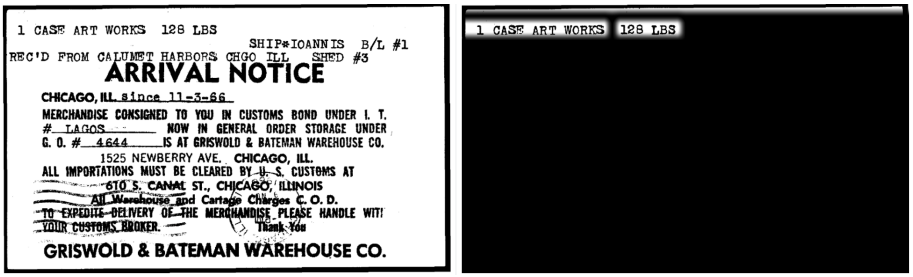


Figure 13. **Question:** How many cases of artworks are there in the shipment? **Answer:** 1 (correct).

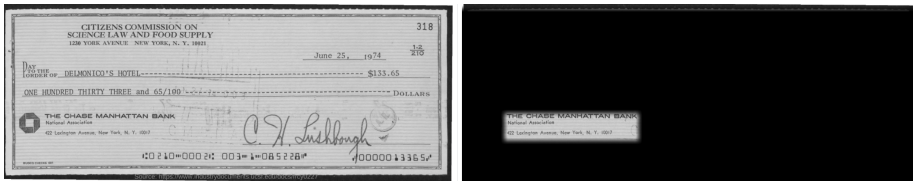


Figure 14. **Question:** Which bank's check is this? **Answer:** The chase manhattan bank (correct).

B.4. Results with a different backbone

In this section, we show the results of DocVXQA using the Donut backbone (Kim et al., 2021), and compare it with the Pix2Struct backbone. The obtained results are given in Figure 15, Figure 16 and Figure 17, demonstrating the model-agnostic capability of our method. These figures highlight the relevant regions produced by DocVXQA using two different backbones while answering the same questions.

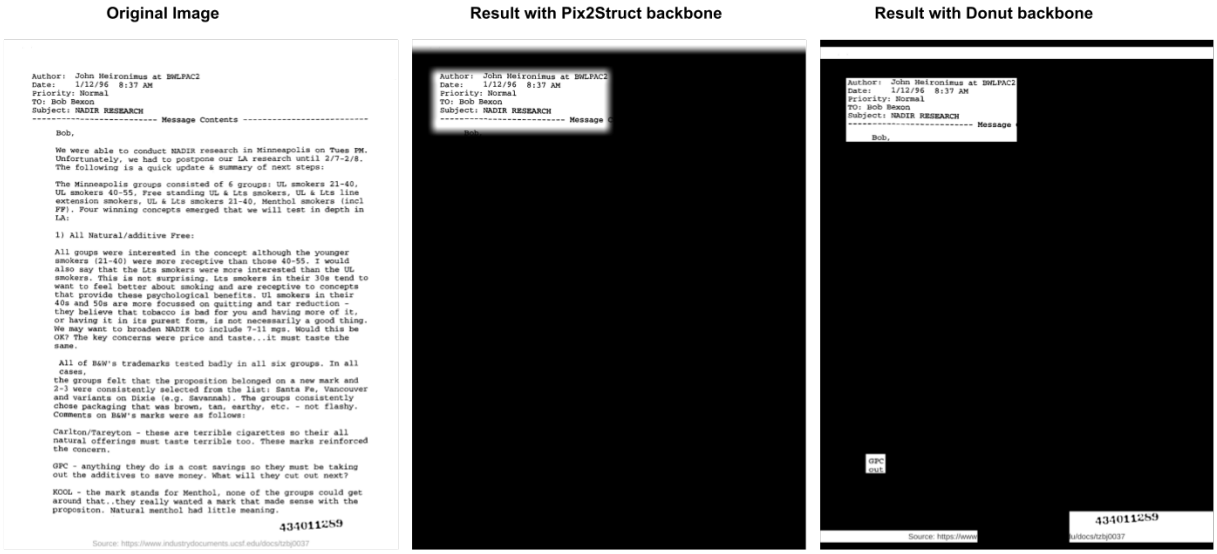


Figure 15. **Question:** Who is this question addressed to? **Answer:** bob bexon (correct).

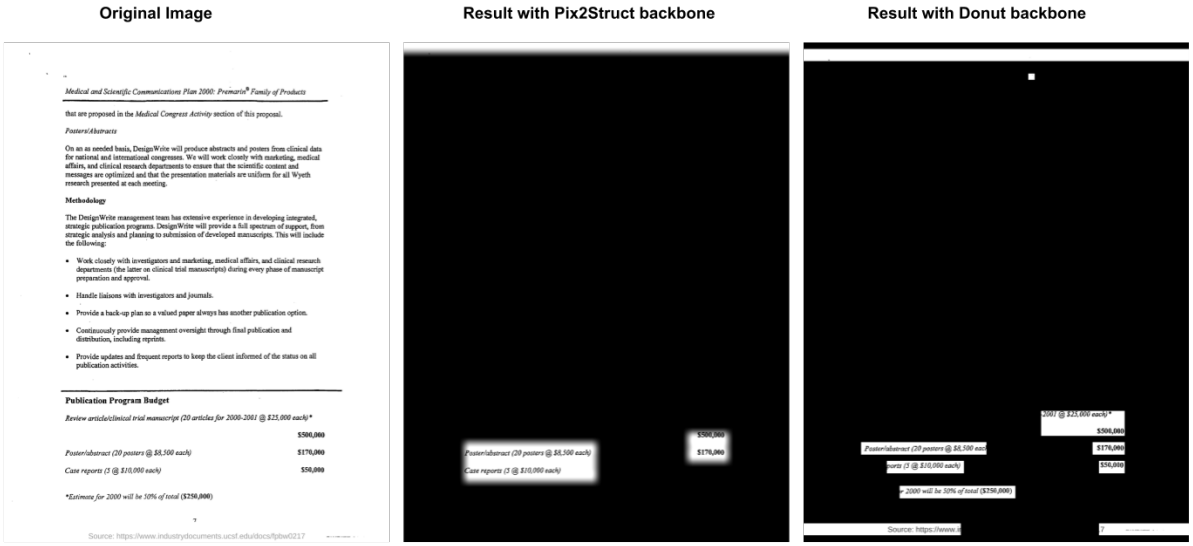


Figure 16. **Question:** What is the amount for publising one poster / abstract? **Answer:** $8,500 (correct).

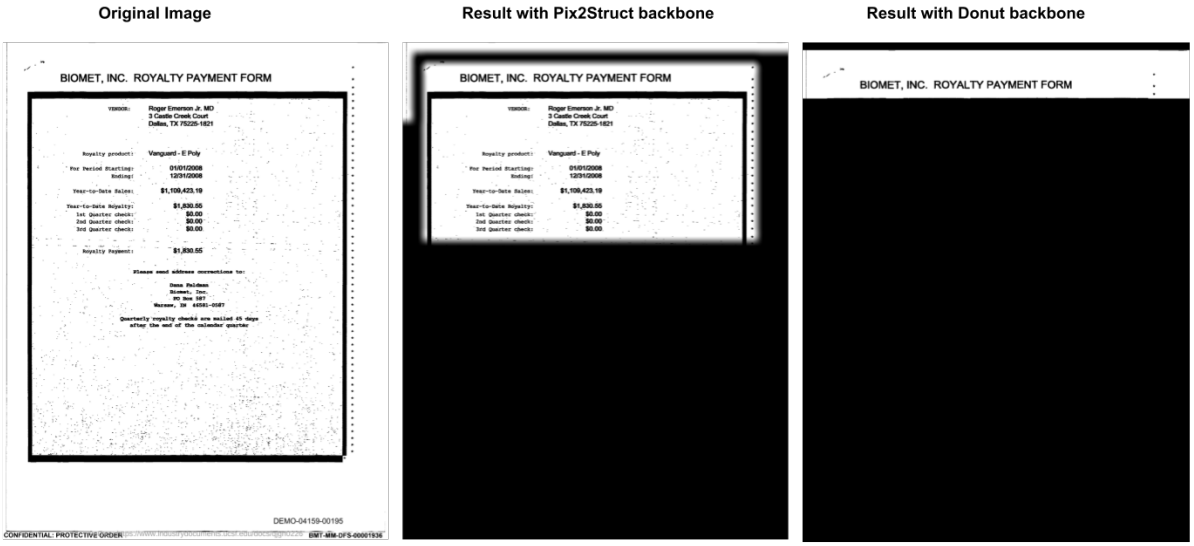


Figure 17. **Question:** what type of communication is issued by Biomet, inc.? **Answer:** royalty payment form (correct).