Spatially Grounded Explanations in Vision–Language Models for Document Visual Question Answering

Maximiliano Hormazábal[1,2] (✉) ⬤, Héctor Cerezo-Costas[2] ⬤, and Dimosthenis Karatzas[1] ⬤

[1] Computer Vision Center, Universitat Autònoma de Barcelona, Spain
{mhormazabal,dimos}@cvc.uab.es
[2] Gradiant, Vigo, Galicia, Spain
hcerezo@gradiant.org

Abstract. We introduce EaGERS, a fully training-free and model-agnostic pipeline that (1) generates natural language rationales via a vision language model, (2) grounds these rationales to spatial sub-regions by computing multimodal embedding similarities over a configurable grid with majority voting, and (3) restricts the generation of responses only from the relevant regions selected in the masked image. Experiments on the DocVQA dataset demonstrate that our best configuration not only outperforms the base model on exact match accuracy and Average Normalized Levenshtein Similarity metrics but also enhances transparency and reproducibility in DocVQA without additional model fine-tuning. Code available at: https://github.com/maxhormazabal/EaGERS-DVQA

Keywords: Document Intelligence · Visual Question Answering · Multimodal Reasoning · Explainability

1 Introduction

Document Visual Question Answering (DocVQA) has advanced rapidly with Transformer-based methods that integrate OCR, layout modeling, and domain adaptation [1,7,4,14]. Concurrently, general-purpose Vision Language Models (VLMs) [19,11,8] achieve strong document understanding without explicit DocVQA training.

Deploying off-the-shelf vision–language models in enterprise pipelines often involves costly fine-tuning, unstable prompt engineering and a lack of clear grounding between answers and source regions [21]. To address these challenges, we introduce Explanation-Guided Region Selection (EaGERS), a fully model-agnostic, training-free DocVQA pipeline that (i) generates natural language explanations, (ii) selects the top sub-regions over a configurable grid via multimodal embedding similarities and majority voting, and (iii) re-queries the model on a masked image so that answers derive solely from those validated regions ensuring transparency and reproducibility without additional model training.

The problem we address is: how to enforce that the answer can be reconstructed solely from document regions that are explicitly grounded and verbalised, without any additional training of the VLM.

The main contributions of this work are:

1. EaGERS: A fully model-agnostic and training-free DocVQA pipeline capable of generating answers on masked document images using general-purpose multimodal models.
2. Integration of text explanations and visual masking to contribute to the traceability and explainability of inferences.

2   Related Work

2.1   DocVQA

In recent years, DocVQA systems have achieved solid baselines on the DocVQA dataset [12], some approaches combine OCR and QA modules such LayoutLM [20] and TILT, yet still lag behind human performance. Other transformer-based models, such as DocFormer and Donut, adopt end-to-end, OCR-free architectures, and supervised-attention methods like M4C [5] integrate textual, positional, and visual cues to boost retrieval accuracy; however, their interpretability remains stuck to attention-weight analysis. Today, state-of-the-art OCR-free approaches reach near-human accuracy but offer no natural language rationales. Meanwhile, multimodal compression frameworks like mPLUG-DocOwl 2.0 improve scalability but still lack explanations grounded in specific document regions.

2.2   Explainability en DocVQA

In the area of spatial explainability, methods such as DocXplain [17] apply ablation techniques on document segments to measure their impact on predictions, demonstrating higher fidelity than Grad-CAM [18] at the cost of multiple inferences. Hybrid models like DLaVA [13] combine textual answers with bounding boxes, achieving Intersection over Union (IoU) [16] above 0.5 on DocVQA, while MRVQA [10] introduces textual rationales alongside visual highlights and proposes specific visual-text coherence metrics.

2.3   Modal Alignment and Multimodal Embeddings

Modal alignment projects text and image representations into a shared latent space, allowing direct comparison via geometric metrics such as cosine similarity. Pretrained models like CLIP [15], ALIGN [6], and BLIP [9] employ a multimodal contrastive objective to bring semantically related pairs closer together. In our pipeline, we use embeddings from BLIP, CLIP, and ALIGN to vectorize both natural language explanations and document sub-regions, enabling the multimodal similarity measurements that drive masking and focused re-querying.

3 Methodology

EaGERS-DocVQA leverages the knowledge of general-purpose multimodal models in document understanding without requiring dedicated training. Figure 1 shows the overall architecture of the proposed system, which consists of three main stages: A) Explanation generation, B) Region selection, and C) Answer generation. In our experiments, we use the Qwen2.5VL-3B model as the core component; however, the proposed system is essentially model-agnostic.
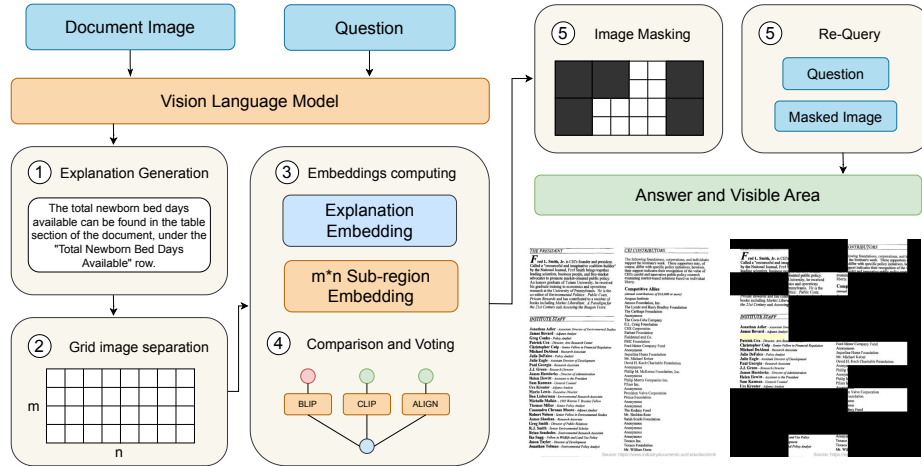

Fig. 1. EaGERS Document VQA pipeline: (1) the multimodal model generates a spatial natural language explanation from the image and the question; (2) the image is segmented into an $m \times n$ grid; (3) embeddings of the explanation and each sub-region are obtained using BLIP, CLIP, and ALIGN; (4) majority voting selects the most relevant regions; (5) the image is masked to retain only those regions, and the model is re-queried with the question to generate the final answer.

3.1 Explanation Generation

The document image and question are passed to a vision–language model, which uses them to generate a natural language explanation of how to obtain the requested answer in visual terms. This explanation is not the final answer but serves as a guide to locate the relevant information in the image. We employ this inference in subsequent steps as a semantic tool for comparing the image sub-regions with the generated explanation. Although these spatial explanations typically exhibit good alignment with relevant regions, there are occasional cases where the generated explanations may inaccurately refer to irrelevant areas, potentially impacting subsequent region selection.

3.2   Region selection via similarity

The document image is divided into an $m \times n$ grid, yielding $m \cdot n$ sub-regions, each of which may or may not be relevant for obtaining the answer.

Each sub-region is converted into a vector representation in order to compare it with the model's spatial explanation by an ensemble of three multimodal models: BLIP, CLIP, and ALIGN to generate embeddings in a complementary way in order to mitigate specific biases and improve robustness across heterogeneous document layouts.

Once we obtain a similarity score based on cosine similarity between reasoning and sub-region embeddings for each embedder, we select the top k for the final answer, where k is 30% (rounded up) of sub-regions based on preliminary experiments, as this provided a good compromise between spatial granularity for evidence localization. Although an adaptive grid might better handle irregular layouts, the fixed grid simplifies reproducibility and reduces complexity for this initial study.

The final ranking is determined by majority voting across the rankings from each embedder. In the event of ties during majority voting, we resolve these by prioritizing regions based on their average cosine similarity scores across all embedders, thus favoring sub-regions with more consistent overall relevance. The resulting list of selected sub-regions $\mathcal{R}$ defines the visible area used in the subsequent answer generation stage.

3.3   Masking and re-query

We create a masked version of the original image in which all grid cells outside $\mathcal{R}$ are filled with black. The question and the masked image are then reintroduced to the same multimodal model, which must generate the answer using only the information within the justified regions, **without access** to the previously generated explanation except implicitly through the defined region mask.

In the following section we evaluate how effectively this approach recovers ground-truth answers under different grid and margin configurations.

4   Experimentation

4.1   Datasets and Evaluation Protocols

In our experiments, we use the validation split of the DocVQA Single Page dataset. We applied resizing preprocessing (preserving the aspect ratio) to compress images and optimize inferences. For spatial partitioning, we divide each image into a uniform grid of 5 columns and 5 rows (25 cells) in an initial series of tests, and an alternative configuration of 5 columns and 10 rows (50 cells) to explore the impact of granularity on relevant-region selection. We also have tested a margin expansion of the 15% of the unmasked sub-region for both amount of cells.

We use Exact Match (EM) and Average Normalized Levenshtein Similarity (ANLS) [2] which mitigates the impact of misrecognition errors by thresholding normalized edit distances.

4.2 Main Results

The table 1 presents the performance results (ANLS and EM) of the different pipeline configurations. For comparison we have also run the Qwen2.5-VL-3B model directly in the DocVQA dataset to compare its performance with the pipeline. In addition to this, the unit inference time (model-only and EaGERS pipeline) have been measured to calculate the average inference time for each approach and their coefficients of variation (CV) have been calculated to gain insights of the similarity of times inference between document[3].

Table 1. Overall performance (EM and ANLS) on validation for different grid and margin settings.

Configuration				Performance			
Model	Cols	Rows	Margin	EM (%)	ANLS	Avg Time (s)	CV (%)
$EaGERS_{25|0}$	5	5	0%	64.20	75.25	16.98	37.08
$EaGERS_{50|0}$	5	10	0%	66.67	77.94	17.49	20.78
$EaGERS_{25|15}$	5	5	15%	72.72	82.52	16.72	24.04
$EaGERS_{50|15}$	5	10	15%	**74.50**	**83.31**	17.48	20.72
Qwen2.5-VL-3B	*	*	*	71.17	82.90	**7.21**	30.08

As shown, adding a 15% masking margin consistently improves both EM and ANLS across grid sizes. In particular, the 50 cells grid with a 15% margin yields the best performance, suggesting that finer spatial granularity combined with slight overlap enhances the localization and understanding of regions relevant to the answer. This suggests that using a fixed distribution grid across the entire image can introduce complications when measuring cosine distance between embeddings. Specifically, relevant zones may fall on the borders between sub-regions, which could explain discrepancies (such as a margin of 0 versus 15) when the grid slightly shifts and either includes or excludes the answer location. This indicates clear future steps in improving the system towards a more flexible grid.

It is also relevant to see that the performance of the model "as-is" has gained explainability without experiencing loss in its performance level, but in the best version $EaGERS_{50|15}$ presents a modest improvement pointing out that restricting the viewing space of the model so that it has more clarity of the answer is an interesting research direction.

Although this first study does not aspire to a head-to-head comparison with the state-of-the-art methods yet. It is also relevant to comment that EaGERS

---

[3] The GPU-hardware used for the experiments were 48GB NVIDIA RTX 6000 Ada Generation.

is above the solutions initially proposed in DocVQA [12] which have shown ANLS results in VQA models such as LoRRA=0.110 or M4C=0.385, BERT QA systems around 0.655 and multimodal architectures such as LayoutLMv2-BASE with an ANLS of 0.7421. However, next steps will be to evaluate this model against state-of-the-art solutions, which achieve even higher ANLS values.

5   Limitations and Future Work

Among the limitations of this study that suggest promising directions for future work: relying on fixed grid configurations may not generalize well to documents with irregular layouts or variable aspect ratios; object detectors may be useful for subdividing relevant regions [3]. Another important limitation is the dependence on spatial explanations generated by VLMs, which may occasionally produce inaccurate rationales, leading to incorrect region selections. Our pipeline assumes spatial accuracy of VLM explanations; however, when these do not match ground truth, the final fidelity may degrade which an aspect we will address in future work.

Future work should include systematic evaluations of the frequency and impact of such inaccuracies on the final results, we will evaluate more robust fusion strategies, such as Reciprocal Rank Fusion (RRF), and quantify agreement using Krippendorff alpha (α), in order to analyze in depth the internal consistency of spatial selections.

It would also be valuable to extend experiments to datasets that explicitly involve answer-localization annotations, enabling the use of quantitative fidelity metrics, such as (IoU) or visual–text coherence scores. Moreover, efficiency improvements are needed considering the increase in inference time reported in Table 1. Finally, we plan to explore methods to assess properly the level of explainability in comparison with alternatives.

6   Conclusions

We have proposed a pipeline that unifies natural language explanations with quantitative region selection and masked re-querying to ensure answers derive only from validated document regions. Our approach yields significant improvements in Exact Match and ANLS over standard baselines, demonstrating enhanced transparency and reproducibility. In future work, we will investigate adaptive grid partitioning to better handle structural variability, conduct comprehensive ablation studies to optimize embedder selection, and directly compare performance with more advanced state-of-the-art models. We also plan to integrate quantitative explainability metrics and carry out user studies to assess the clarity and reliability of the generated explanations.

References

1. Appalaraju, S., Jasani, B., Kota, B.U., Xie, Y., Manmatha, R.: Docformer: End-to-end transformer for document understanding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 993–1003 (2021)

2. Biten, A., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Jawahar, C., Valveny, E., Karatzas, D.: Scene text visual question answering, pp. 4290–4300. Proceedings of the IEEE International Conference on Computer Vision, Institute of Electrical and Electronics Engineers Inc., United States (Oct 2019). https://doi.org/10.1109/ICCV.2019.00439, funding Information: This work has been supported by projects TIN2017-89779-P, Marie-Curie (712949 TECNIOspring PLUS), aBSINTHE (Fundacion BBVA 2017), the CERCA Programme / Generalitat de Catalunya, a European Social Fund grant (CCI: 2014ES05SFOP007), NVIDIA Corporation and PhD scholarships from AGAUR (2019-FIB01233) and the UAB. Publisher Copyright: © 2019 IEEE.

3. Gómez, L., Biten, A.F., Tito, R., Mafla, A., Rusiñol, M., Valveny, E., Karatzas, D.: Multimodal grid features and cell pointers for scene text visual question answering. Pattern Recognition Letters **150**, 242–249 (2021). https://doi.org/https://doi.org/10.1016/j.patrec.2021.06.026, https://www.sciencedirect.com/science/article/pii/S0167865521002336

4. Hu, A., Xu, H., Zhang, L., Ye, J., Yan, M., Zhang, J., Jin, Q., Huang, F., Zhou, J.: mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding (2024), https://arxiv.org/abs/2409.03420

5. Hu, R., Singh, A., Darrell, T., Rohrbach, M.: Iterative Answer Prediction With Pointer-Augmented Multimodal Transformers for TextVQA . In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9989–9999. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2020). https://doi.org/10.1109/CVPR42600.2020.01001, https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.01001

6. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 4904–4916. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/jia21b.html

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

8. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 19730–19742. PMLR (23–29 Jul 2023), https://proceedings.mlr.press/v202/li23q.html

9. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 12888–12900. PMLR (17–23 Jul 2022), https://proceedings.mlr.press/v162/li22n.html

10. Li, K., Vosselman, G., Yang, M.Y.: Convincing rationales for visual question answering reasoning (2025), `https://arxiv.org/abs/2402.03896`
11. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 34892–34916. Curran Associates, Inc. (2023), `https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-`
12. Mathew, M., Karatzas, D., 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)
13. Mohammadshirazi, A., Neogi, P.P.G., Lim, S.N., Ramnath, R.: Dlava: Document language and vision assistant for answer localization with enhanced interpretability and trustworthiness (2024), `https://arxiv.org/abs/2412.00151`
14. Powalski, R., Borchmann, Ł., Jurkiewicz, D., Dwojak, T., Pietruszka, M., 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)
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021), `https://proceedings.mlr.press/v139/radford21a.html`
16. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression . In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 658–666. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2019). `https://doi.org/10.1109/CVPR.2019.00075`, `https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.00075`
17. Saifullah, S., Agne, S., Dengel, A., Ahmed, S.: Docxplain: A novel model-agnostic explainability method for document image classification. In: International Conference on Document Analysis and Recognition. pp. 103–123. Springer (2024)
18. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Gradcam: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (2017). `https://doi.org/10.1109/ICCV.2017.74`
19. Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., Kiela, D.: Flava: A foundational language and vision alignment model. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15617–15629 (2022). `https://doi.org/10.1109/CVPR52688.2022.01519`
20. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., 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. p. 1192–1200. KDD '20, ACM (Aug 2020). `https://doi.org/10.1145/3394486.3403172`, `http://dx.doi.org/10.1145/3394486.3403172`
21. Zhou, Y., Muresanu, A.I., Han, Z., Paster, K., Pitis, S., Chan, H., Ba, J.: Large language models are human-level prompt engineers. In: The Eleventh International Conference on Learning Representations (2022)