DocEdit Redefined: In-Context Learning for Multimodal Document Editing

Muhammad Waseem [1] * Sanket Biswas[2] Josep Lladós[2]
[1] Shiv Nadar University Chennai, India
[2] Computer Vision Center, Universitat Autònoma de Barcelona, Spain
muhammad21110007@snuchennai.edu.in,{sbiswas, josep}@cvc.uab.cat

Abstract

Structured document generation relies on mapping logical content to its displayed form, which requires integration of visual, textual, and layout elements. This paper presents a novel approach to structured document editing by leveraging the object recognition capabilities of Visual-Language Models (VLMs) to eliminate the dependency on specialized segmentation modules. By integrating state-of-the-art open-world VLMs, we propose a simple in-context learning framework that enhances the flexibility and efficiency of Language-guided Document Editing (DocEdit) tasks. Our method demonstrates promising performance in addressing complex document modifications, such as spatial alignment, component merging, and regional grouping, while maintaining the coherence and intent of the original document. Through our proposed few-shot evaluation benchmark suite, we highlight the potential of VLMs in this direction. Furthermore, we propose a refined evaluation protocol that incorporates both spatial and semantic reasoning, ensuring a comprehensive assessment of the modified (edited) output for the task. Experimental results underscore the effectiveness of our framework in advancing the capabilities of structured document editing systems.

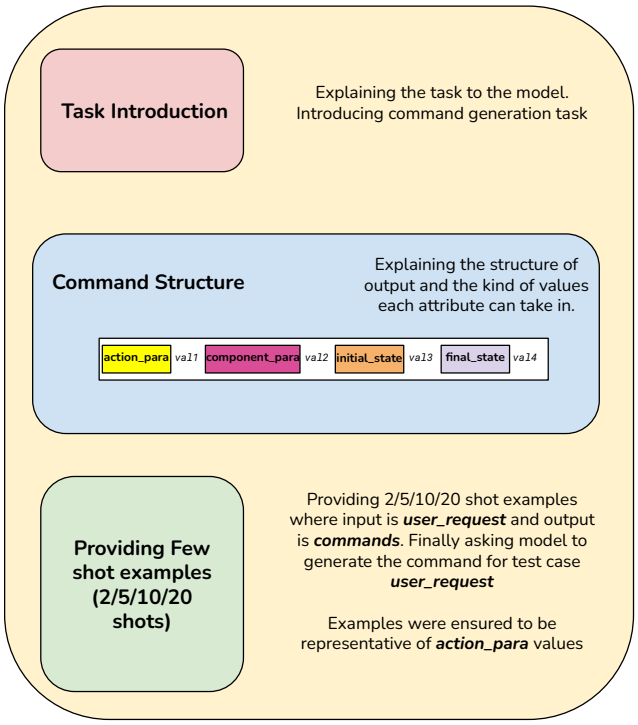

Figure 1. **An Illustration of the In-Context Learning prompt template used for DocEdit**, utilizing a chosen VLM model and an edit request from the user.

1. Introduction

In structured document generation systems, a key component is the mapping of logical content to its displayed form, either on paper or screen. This mapping often links visual specifications with document grammar elements, dictating how each component should appear [8]. Layout can be managed by associating formatting attributes (like margins or spacing) with elements or using action routines to adjust global display settings. Alternatively, a design can be explicitly defined through templates or interactive tools, which allows precise control over where content is placed. In this context, *Language-guided Document Editing* (a.k.a DocEdit) entails modifying the textual, visual and structural components of a document image in response to open-ended user requests related to spatial alignment, component placement, regional grouping, replacement, resizing, splitting, merging, and applying special effects [9, 15]. This editing process is inherently generative, requiring the creation of a new edited output that reflects the "user modifications" while preserving the overall coherence and intent of the original document.

Generative approaches, including diffusion models [18] have shown potential in the visual domain but struggled to reproduce multimodal elements (combining visual, textual, and layout modalities) [7, 24]. The emergence of Visual-

*Work done during internship at Computer Vision Center

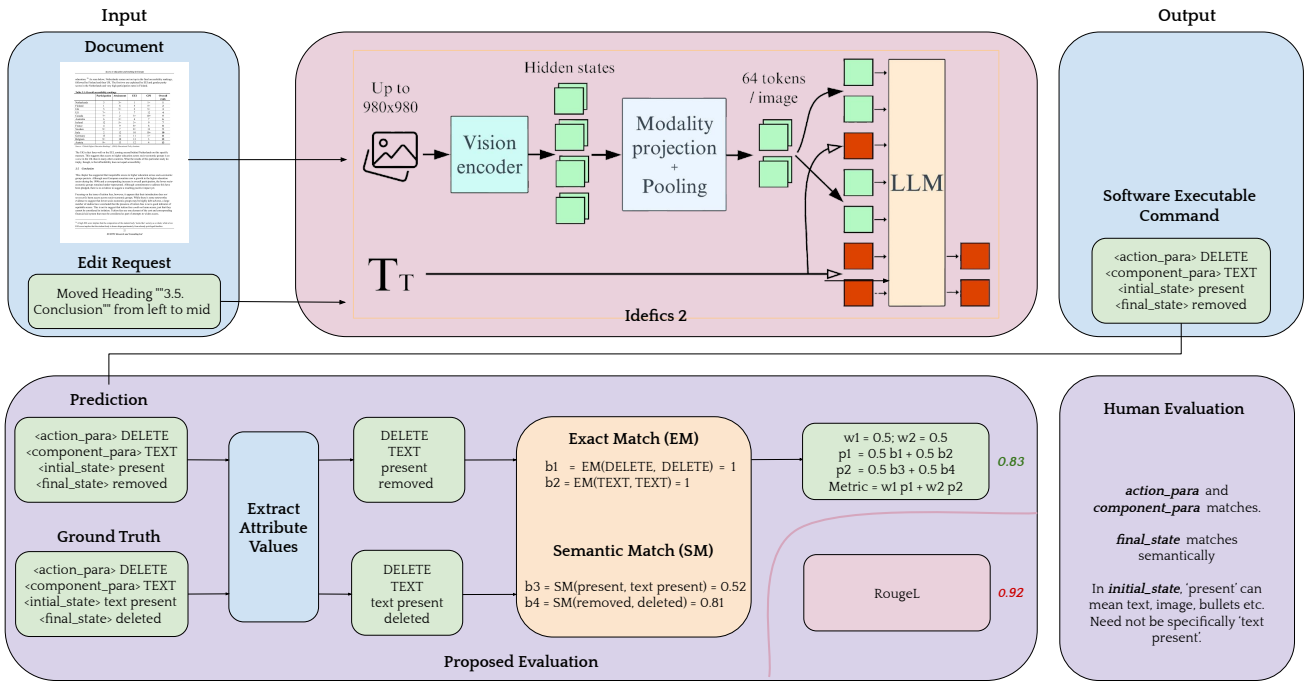

Figure 2. Our proposed framework allows for a plug-and-play approach, enabling the integration of any VLM within the pipeline to extract the corresponding attributes for evaluation.

Language Models (VLMs) with the Flamingo [2] family provided a huge breakthrough in this direction by performing several multimodal tasks (such as captioning, visual dialogue generation, or visual question-answering) from only a few input/output examples. Their ability to handle interleaved text and visual sequences makes it natural to use them for *in-context few-shot learning*, analogously to GPT-3 [4] with few-shot text prompting. Liu et. al. [14] introduced visual instruction tuning to further enhance instruction following capabilities of VLMs. We refer readers to this multimodal large language models (MLLMs) [5] survey for a more up-to-date literature compilation.

The current state-of-the-art (DocEdit-v2 [20] in the DocEdit benchmark has leveraged the power of Large Multimodal Models (LMMs) like Gemini [22] or GPT4-V [1] by fusing with the Doc2Command [21] segmentation module, which grounds the edit location in the document image to finally generate the edit commands. However, in this work, we utilize the object recognition capabilities of VLM's instead to propose a framework which doesn't require a specialized segmentation module for localizing object elements within the document. Moreover, we also integrate more open-world multimodal VLMs in document understanding [11, 12] and explore their in-context learning abilities by proposing a few-shot evaluation benchmarking. Lastly, we also emphasize the need for a proper evaluation protocol (closer to human evaluation) when evaluating the DocEdit task, which comprises reasoning over both spatial and semantic components of a document to generate the final output. The overall contributions of this work can be summarized in three folds: 1) A simple and flexible In-Context Learning framework for Document Editing Task, that enables the integration of any open world VLM models 2) A Few-Shot Evaluation Benchmark suite for Document Editing task, that could be further utilized for future enhancements 3) A new evaluation metric which assesses both exact and semantic matches for structured document editing.

2. Proposed Methodology

2.1. Dataset: "DocEdit"

The DocEdit dataset addresses the novel task of language-guided localized document editing, pairing user-provided natural language editing requests with executable commands applied to PDF documents. The public version includes 12,704 unique PDF images, split into 8,067 for training, 1,630 for testing, and 3,007 for validation. These splits correspond to 12,464, 1,780, and 3,563 editing request-command pairs, respectively. Unlike the dataset used in prior works [15], [23], [21], [20], public version of this dataset lacks bounding box annotations and ground truth edited images. It supports diverse editing operations—such as adding, deleting, modifying, or rearranging document components like text, images, or layouts—and incorporates multimodal features (textual, visual, spatial). This makes it a benchmark

Model	Rouge-L				Action Match				Component Match			
	2-Shot	5-Shot	10-Shot	20-Shot	2-shot	5-shot	10-shot	20-shot	2-shot	5-shot	10-shot	20-shot
Text-only Models												
T5	0.192	0.136	0.137	0.278	0	0	0	0	0	0	0	0
GPT2	0.331	0.339	0.345	0.344	0.033	0.091	0.079	0.062	0.127	0.145	0.166	0.199
VLMs												
Idefics2	0.678	0.741	0.742	0.753	0.597	0.653	0.706	0.701	0.093	0.649	0.611	0.679
Idefics3	0.652	0.660	0.753	0.769	0.733	0.653	0.727	0.767	0.001	0.320	0.633	0.669
Llava	0.643	0.701	0.724	0.748	0.604	0.646	0.521	0.325	0.047	0.552	0.490	0.299
Llava-next	0.629	0.669	0.705	0.739	0.560	0.692	0.703	0.688	0	0.009	0.273	0.623
Qwen_vl	0.388	0.392	0.399	0.432	0.393	0.478	0.448	0.426	0.502	0.577	0.503	0.554

Table 1. Performance of In-Context Learning with Few-Shot Models across different metrics. Across the columns: dark green highlights the best performance, while light green indicates the next best performance. Detailed explanations of the results are provided in Section 3.1 and Section 3.3.

Model	Finetuned Rouge-L	Finetuned Action Match	Finetuned Component Match
Text-only Models			
T5	0.76	0.81	0.29
GPT2	0.76	0.81	0.29
Explicit OCR Models			
DocEditor	0.86	0.87	0.41
VLMs			
Idefics2	0.86	0.89	0.81
Idefics3	0.85	0.88	0.81
Llava	0.75	0.71	0.72
Llava-next	0.74	0.86	0.80
Qwen_vl	0.76	0.83	0.75
Current SOTA			
DocEdit-v2	0.86	0.85	0.86

Table 2. Comparison of fine-tuned models with few-shot ablations. Across the rows, dark green highlights the best performance, while light green indicates the second-best performance. The performance of Q-LoRA fine-tuned models is comparable to the fully fine-tuned SOTA model (see Section 3.2 and Section 3.3 for details)

for intelligent document editing model development.

Why this dataset? The DocEdit dataset is the first of its kind to combine multimodal features, enabling a robust evaluation of models designed for intelligent and localized document editing tasks. Its diversity and emphasis on real-world editing scenarios make it a suitable benchmark for this domain.

2.2. Overview of Vision-Language Models Used

We evaluated various pre-trained VLMs specifically trained on document images, such as Idefics2 [12], Idefics3 [10], Llava-next [14], and Qwen_vl [3]. These were compared against text-only models like T5 [17] and GPT-2 [16] and VLMs trained on non-document images, such as Llava [13].

2.3. In-Context Learning with Customised Prompts

Our approach leverages in-context learning by constructing tailored prompts that explicitly guide the model on the task, define input-output formats, and outline possible values for each software command attributes (Figure 1). These prompts are designed to include diverse examples that comprehensively represent all potential actions.

2.4. Standard Evaluation Metrics

We used the following established metrics from [15] for command generation: 1) **ROUGE-L**: Measures sequence similarity using the longest common subsequence (LCS). 2) **Action Match (AM)**: Assesses the exact match accuracy of the action_para attribute. 3) **Component Match (CM)**: Assesses the exact match accuracy of the component_para attribute.

2.5. Flow Graph Metric

Global metrics like Rouge-L are effective for most text generation tasks. However, we observe that such metrics are less suitable for structured text generation, where both exact and semantic equivalence are crucial. For instance, in DocEdit, exact match accuracy is required for action_para and component_para, while semantic match equivalence is assessed for initial_state and final_state using a pretrained sentence transformer. To better align with human judgment, we propose a hybrid evaluation metric that combines exact and semantic matches (Figure 2). Specifically, we utilize simple exact match accuracy for action_para and compo-

Model	RougeL Score	Flow Graph Score	Correlation (RougeL, Human)	Correlation (Flow Graph, Human)
idefics2	0.8621	0.7987	0.6375	0.6762
idefics3	0.8507	0.7782	0.5324	0.5559
llava	0.5532	0.6047	0.1988	0.3440
llava_next	0.5563	0.7037	0.2744	0.5429
qwen	0.7982	0.7079	0.3970	0.4702
T5	0.8479	0.7756	0.5504	0.5692
GPT2	0.8092	0.7627	0.4520	0.5576

Table 3. Point-biserial correlation comparison of Rouge-L and Flow Graph scores with human judgment. The scores are averages across all test examples, and the correlation is calculated using the point-biserial function in scipy (Refer Section 2.5)

nent_para, apply a sentence transformer for initial_state and final_state, and aggregate the results. Our proposed metric demonstrates a higher point biserial correlation with human-evaluated binary scores compared to the Rouge-L metric (Table 3).

3. Experiments and Results

3.1. Performance of In-Context Learning

In-context learning demonstrates strong performance with vision-language models (VLMs) pre-trained on diverse document datasets containing interleaved text-image pairs. These models outperform VLMs trained on non-document images. Moreover, as the number of examples (shots) in the prompt increases, the model's task understanding improves, resulting in better command generation performance across the metrics (Table 1). This suggests that VLMs can implicitly capture and utilize document layout information during inference.

3.2. Fine-Tuning Experiments

Due to computational constraints, we fine-tuned all VLMs using Q-LoRA [6]. To ensure a fair comparison, the fine-tuning process was limited to a fixed number of iterations, reflecting the resource limitations commonly faced by the general public. Despite this restricted fine-tuning, idefics2 achieved results comparable to the DocEdit-v2 model. The similar performance of T5 and GPT-2 with models like LLaVA, LLaVA-Next, and Qwen_VL can be attributed to the fact that T5 and GPT-2, being text-only models, were fully fine-tuned. However, the VLMs are expected to exhibit better performance and generalization capability [19] when fully fine-tuned.

3.3. Analysis of Results

In Table 1, we observe a trend where increasing the number of examples (shots) in in-context learning generally leads to improved performance metrics for RougeL, Action Match, and Component Match, respectively. However, in the case of Action Match, we note that performance peaks at a certain number of shots, after which it declines as the shot count continues to increase. For fine-tuned models on this specific task, state-of-the-art performance is achieved using quantized fine-tuning, which outperforms full training for the SOTA Docedit-V2 model [20] except in component match score as shown in Table 2.

4. Conclusion and Future Scope

In this work, we demonstrated the potential of Visual Language Models (VLMs) for language-guided document editing, leveraging their multimodal grounding to overcome the limitations of traditional OCR-based methods. VLMs exhibited strong performance by balancing semantic coherence and structural precision, significantly outperforming text-only models and non-document-trained VLMs. Our proposed hybrid evaluation metric aligned better with human judgment, addressing the nuanced demands of structured command generation. We also highlighted the efficiency of in-context learning as a resource-friendly alternative to fine-tuning, enabling VLMs to adapt effectively to diverse tasks with minimal computational overhead. Lightweight fine-tuning methods like Q-LoRA showed competitive performance, suggesting a path forward for accessible and scalable solutions.

Acknowledgement

We thank the support of Spanish projects GRAIL PID2021-126808OB-I00, DocAI 2021-SGR-01559, the CERCA Program / Generalitat de Catalunya, and PhD Scholarship from AGAUR 2023 FI-3-00223.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 2

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou.

Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023. 3

[4] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 2

[5] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The (r) evolution of multimodal large language models: A survey. *arXiv preprint arXiv:2402.12451*, 2024. 2

[6] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024. 4

[7] Liu He, Yijuan Lu, John Corring, Dinei Florencio, and Cha Zhang. Diffusion-based document layout generation. In *International Conference on Document Analysis and Recognition*, pages 361–378. Springer, 2023. 1

[8] Gunther Kress and Theo Van Leeuwen. *Reading images: The grammar of visual design*. Routledge, 2020. 1

[9] Katya Kudashkina, Patrick M Pilarski, and Richard S Sutton. Document-editing assistants and model-based reinforcement learning as a path to conversational ai. *arXiv preprint arXiv:2008.12095*, 2020. 1

[10] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024. 3

[11] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[12] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024. 2, 3

[13] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 3

[14] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2, 3

[15] Puneet Mathur, Rajiv Jain, Jiuxiang Gu, Franck Dernoncourt, Dinesh Manocha, and Vlad I Morariu. Docedit: language-guided document editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1914–1922, 2023. 1, 2, 3

[16] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

[17] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 3

[18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[19] Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. Lora vs full fine-tuning: An illusion of equivalence. *arXiv preprint arXiv:2410.21228*, 2024. 4

[20] Manan Suri, Puneet Mathur, Franck Dernoncourt, Rajiv Jain, Vlad I Morariu, Ramit Sawhney, Preslav Nakov, and Dinesh Manocha. Docedit-v2: Document structure editing via multimodal llm grounding. *arXiv preprint arXiv:2410.16472*, 2024. 2, 4

[21] Manan Suri, Puneet Mathur, Ramit Sawhney, Preslav Nakov, and Dinesh Manocha. Doc2command: Furthering language guided document editing. In *The Second Tiny Papers Track at ICLR 2024*. 2

[22] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2

[23] Te-Lin Wu, Rajiv Jain, Yufan Zhou, Puneet Mathur, and Vlad I Morariu. Agent-docedit: Language-instructed llm agent for content-rich document editing. In *First Conference on Language Modeling*. 2

[24] Zongyuan Yang, Baolin Liu, Yongping Xxiong, Lan Yi, Guibin Wu, Xiaojun Tang, Ziqi Liu, Junjie Zhou, and Xing Zhang. Docdiff: Document enhancement via residual diffusion models. In *Proceedings of the 31st ACM international conference on multimedia*, pages 2795–2806, 2023. 1