Marten: Visual Question Answering with Mask Generation for Multi-modal Document Understanding (Supplementary Material)

Zining Wang[1*], Tongkun Guan[2*], Pei Fu[1(✉)], Chen Duan[1], Qianyi Jiang[1], Zhentao Guo[3], Shan Guo[1], Junfeng Luo[1], Wei Shen[2(✉)], Xiaokang Yang[2]

[1] Meituan [2] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University

[3] Beijing Institute of Technology

{wangzining03,fupei}@meituan.com, {gtk0615,wei.shen}@sjtu.edu.cn

1. More visualizations about VQAMask

In this section, we show more visualization examples in Figure 1 and 2. Each example includes (a) Input image, (b) Attention w/o MGM, (c) Attention with MGM, (d) Prediction Mask, and (e) Our generated label. Specifically, these attention maps in the "Attention w/o MGM" column (b) are obtained from the version without our proposed mask generation module (MGM). These attention maps in the "Attention with MGM" column (c) are obtained from the version using our proposed mask generation module (MGM). The "Predicted Mask" column (d) exhibits the final predicted mask, which delineates all text locations in the document, with spatially-aware supervision by our generated labels (e).

Example A:

Figure 1 exhibits the visualizations from the task: **Reading Full Text.** Given an image, the model needs to predict all visual texts sequentially. Specifically, the image, question, and answer are embedded into a question-answer template like:

QUESTION: `Recognize all texts.|Convert the image into Markdown format.`
ANSWER: `BRAND R6D ... SALEM LTS 85.`

In this task, our model combines the question and answer to activate the visual text regions of the input image. When comparing the attention maps from the (b) and (c) columns, we observed MGM promotes the alignment between visual tokens and language tokens. In other words, visual tokens corresponding to the visual text regions are further highlighted. The highlighted attentions allow our model to capture more important information for subsequent visual question answering.

Example B:

Figure 2 exhibits the examples from the task: **Reading Partial Text within Localization.** Similarly, the question-answer template is formulated:

QUESTION: `Identify the text within the bounding box <bbox> 109, 85, 595, 389 </bbox>.`
ANSWER: `9 Nov.22 Morehead State Win 40 6 8-1.`

In this task, the model needs to understand the significance of the number within the `<bbox>, </bbox>` tags. The number represents a box and its specific location in the image. Only by understanding this can the model accurately predict the text in the box. Obviously, this task is more challenging. As shown in the second column, the version without our proposed MGM is difficult to find the specific location of the given box. If the location is incorrect, the prediction result will also be wrong. In the version with MGM, with explicit position supervision (presented in the last column), the interaction between language and image can effectively promote the model's understanding of these tokens. As a result, the obtained attention maps are more accurate.

Example C:

In Figure 3, we further exhibit the qualitative comparison results of using and not using MGM. Without spatially-aware supervision, the outputs from the version without MGM may disproportionately rely on the powerful semantic context capabilities of large language models (LLMs) rather than optimizing image features from visual encoders, potentially leading to model hallucinations. As discussed above, our proposed VQAMask optimises two tasks simultaneously: VQA-based text parsing and mask generation. The former allows the model to implicitly align images and text at the semantic level. The latter introduces an additional

[*] These authors contributed equally. [✉] Corresponding Author.

mask generator (discarded during inference) to explicitly ensure alignment between visual texts within images and their corresponding image regions at a spatially-aware level. Together, they can prevent model hallucinations when parsing visual text and effectively promote spatially-aware feature representation learning.

2. More examples compared to other MLLMs

As shown in Figure 4, we present more qualitative visualization results to demonstrate Marten's capabilities in various VQA tasks. Marten analyzes the question, identifies the key elements in the image relevant to answering the question, and exhibits the impressive localization ability to perceive even minute text within the image.

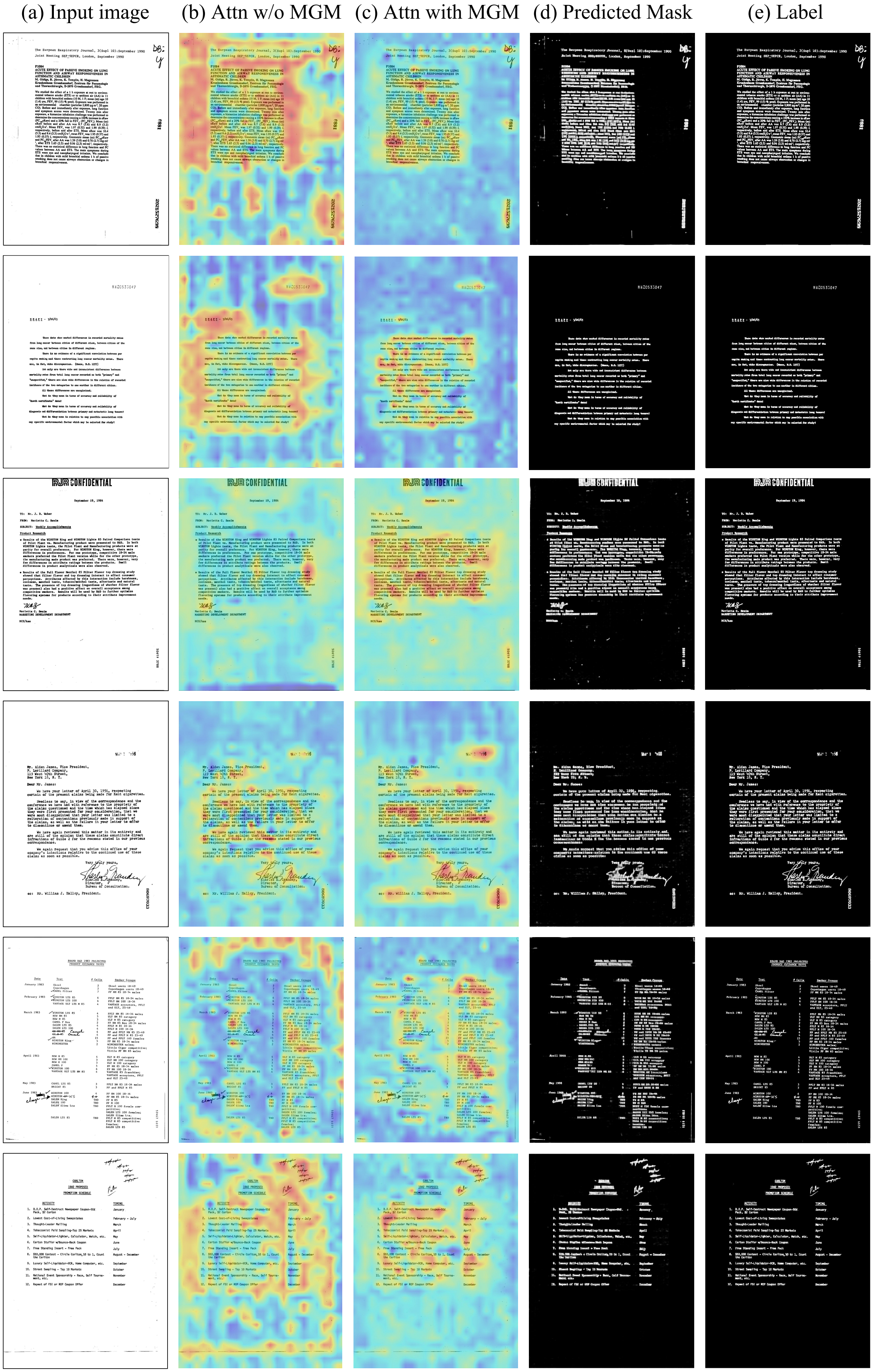
3. Explanation on mask generation

We only enable mask generation in stage 1 due to all answers of text parsing tasks appearing directly in the image. In contrast, stage 2 involves reasoning tasks (*e.g., How much taller is the red bar than the blue bar?*), where answers do not appear in the image. Thus, mask generation is reasonably omitted due to unfeasible experiments.

For tasks like text grounding (**boxes in answers**) and reading partial text within localization (**boxes in questions**), we conduct the mask generation pipeline on the **local image** within provided boxes to create local masks. For other parsing tasks where all visual texts are required, we provide global masks. Thus, both local and global masks are provided in our dataset.

4. Mask construction differences from SIGA

SIGA is initially designed to support standard text instance images cropped from the whole image based on **GT boxes**. This faces challenges when dealing with incomplete text instance images cropped by **pseudo boxes** generated from PaddleOCR, due to the scarcity of box annotations. Thus, we propose global judgment conditions to replace edge conditions of SIGA.



(a) Input image (b) Attn w/o MGM (c) Attn with MGM (d) Predicted Mask (e) Label

Figure 1. Visualizations of some key items in Reading Full Text task, including (a) Input image (b) Attention without MGM (c) Attention with MGM (d) Prediction Mask and (e) Our generated label.

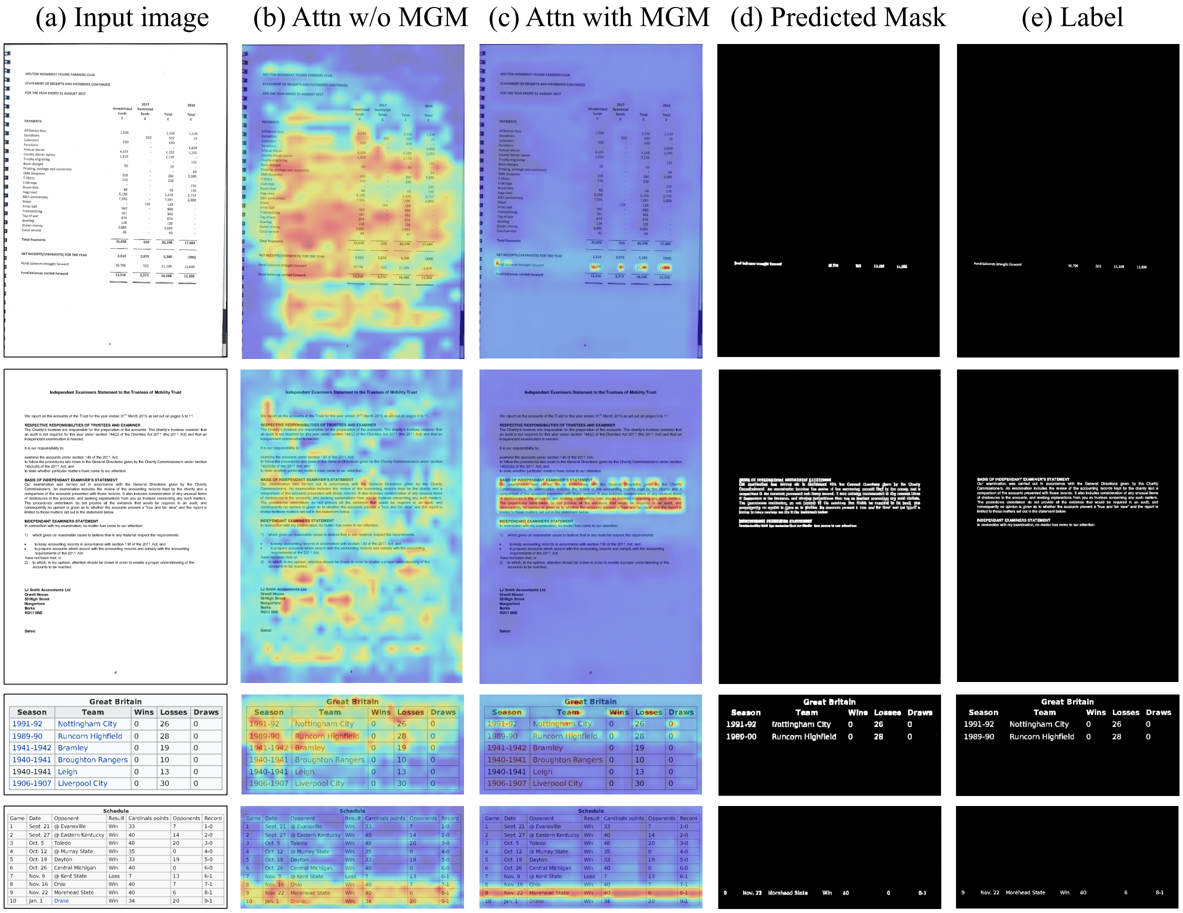
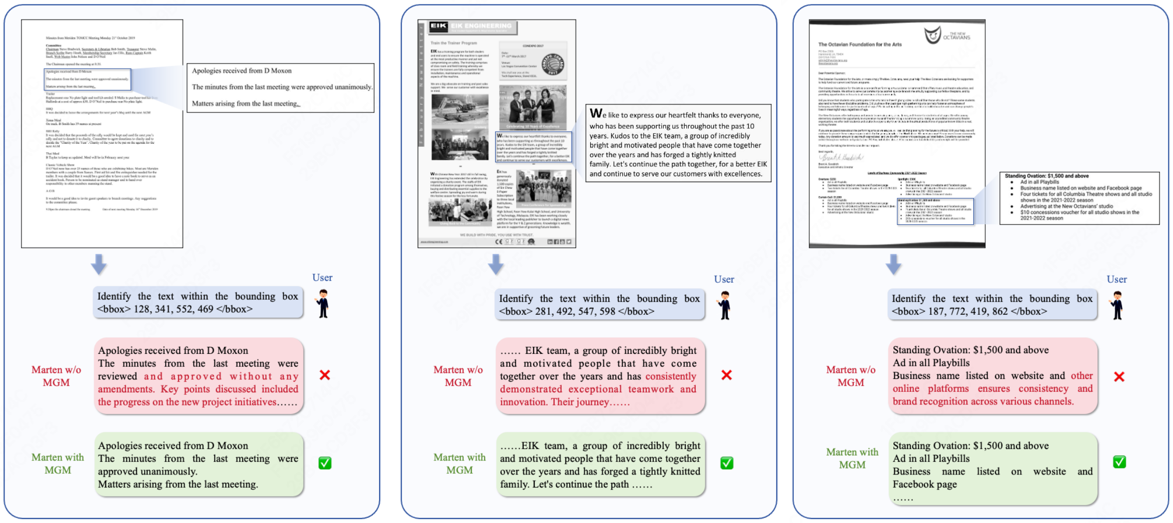


Figure 2. Visualizations of some key items in Reading Partial Text within Localization task, including (a) Input image (b) Attention without MGM (c) Attention with MGM (d) Prediction Mask and (e) Our generated label.


Apologies received from D Moxon
The minutes from the last meeting were approved unanimously.
Matters arising from the last meeting_
We like to express our heartfelt thanks to everyone, who has been supporting us throughout the past 10 years. Kudos to the EIK team, a group of incredibly bright and motivated people that have come together over the years and has forged a tightly knitted family. Let's continue the path together, for a better EIK and continue to serve our customers with excellences.
Standing Ovation: $1,500 and above
Ad in all Playbills
Business name listed on website and Facebook page
Four tickets for all Columbia Theatre shows and all studio shows in the 2021-2022 season
Advertising at the New Octavians' studio
$10 concessions voucher for all studio shows in the 2021-2022 season
User
Marten w/o MGM
Marten with MGM

User
Identify the text within the bounding box <bbox> 128, 341, 552, 469 </bbox>

Marten w/o MGM
Apologies received from D Moxon
The minutes from the last meeting were reviewed and approved without any amendments. Key points discussed included the progress on the new project initiatives…… ✗

Marten with MGM
Apologies received from D Moxon
The minutes from the last meeting were approved unanimously.
Matters arising from the last meeting. ✓

User
Identify the text within the bounding box <bbox> 281, 492, 547, 598 </bbox>

Marten w/o MGM
…… EIK team, a group of incredibly bright and motivated people that have come together over the years and has consistently demonstrated exceptional teamwork and innovation. Their journey…… ✗

Marten with MGM
……EIK team, a group of incredibly bright and motivated people that have come together over the years and has forged a tightly knitted family. Let's continue the path …… ✓

User
Identify the text within the bounding box <bbox> 187, 772, 419, 862 </bbox>

Marten w/o MGM
Standing Ovation: $1,500 and above
Ad in all Playbills
Business name listed on website and other online platforms ensures consistency and brand recognition across various channels. ✗

Marten with MGM
Standing Ovation: $1,500 and above
Ad in all Playbills
Business name listed on website and Facebook page
…… ✓

Figure 3. Qualitative comparison results of using and not using MGM.

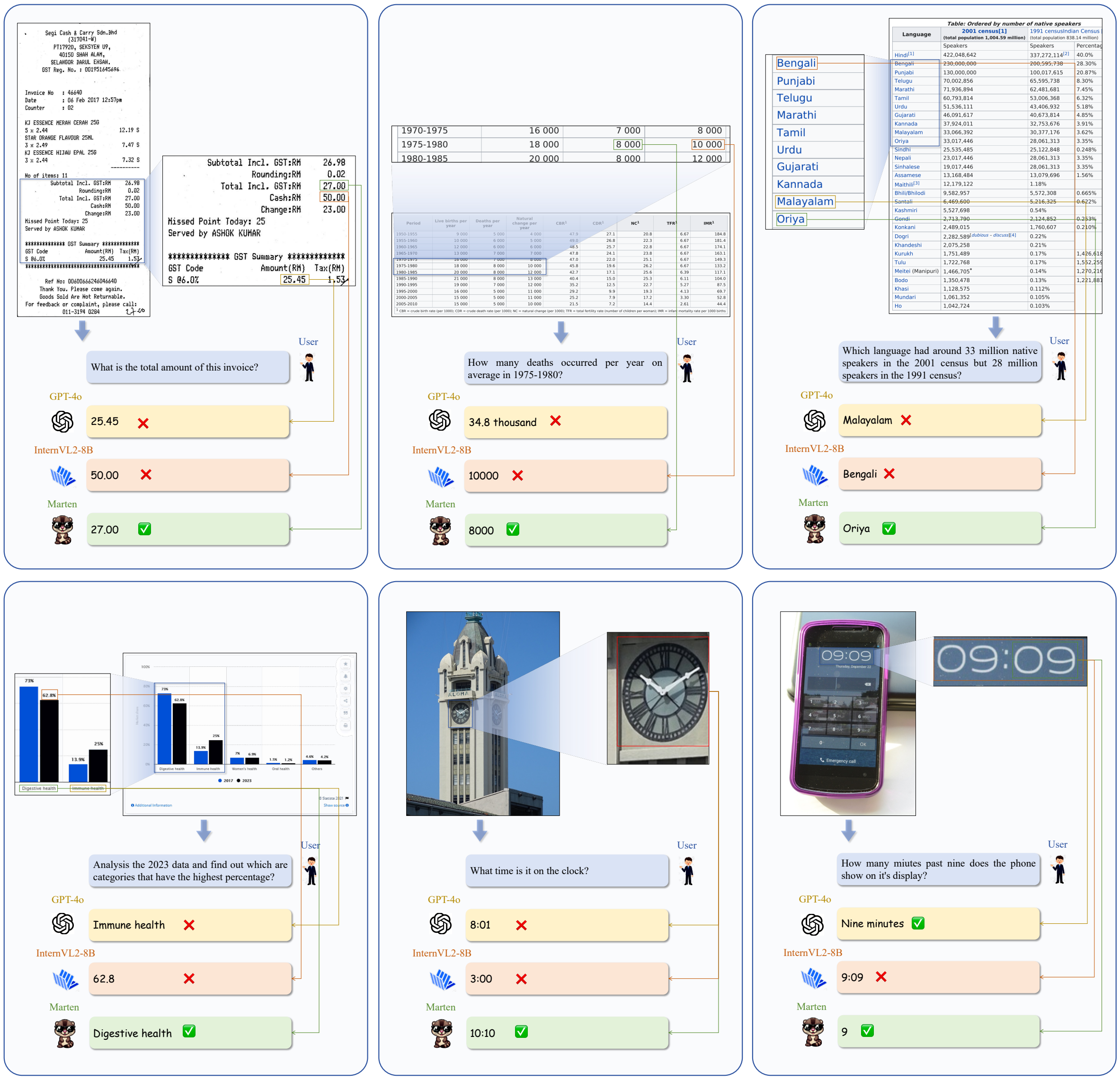


Figure 4. Visualization of Marten's comparison with GPT-4o, internvl2-8B on VQA tasks.