MMVQA: A Comprehensive Dataset for Investigating Multipage Multimodal Information Retrieval in PDF-based Visual Question Answering

Yihao Ding\textsuperscript{1,2}, Kaixuan Ren\textsuperscript{2}, Jiabin Huang\textsuperscript{2}, Siwen Luo\textsuperscript{3} and Soyeon Caren Han\textsuperscript{1,2}\textsuperscript{*}

\textsuperscript{1}The University of Melbourne
\textsuperscript{2}The University of Sydney
\textsuperscript{3}The University of Western Australia

\texttt{yihao.ding@sydney.edu.au, kren4925@uni.sydney.edu.au,}
\texttt{jiabin.eta@gmail.com, siwen.luo@uwa.edu.au, caren.han@unimelb.edu.au}

Abstract

Document Question Answering (QA) presents a challenge in understanding visually-rich documents (VRD), particularly with lengthy textual content. Existing studies primarily focus on real-world documents with sparse text, while challenges persist in comprehending the hierarchical semantic relations among multiple pages to locate multimodal components. The paper introduces MMVQA, a dataset tailored for research journal articles, encompassing multiple pages and multimodal retrieval. Our approach aims to retrieve entire paragraphs containing answers or visually rich document entities like tables and figures. The main contribution is introducing a comprehensive PDF Document VQA dataset, allowing the examination of semantically hierarchical layout structures in text-dominant documents. We also present new VRD-QA frameworks to grasp textual contents and relations among document layouts simultaneously, extending page-level understanding to the entire multi-page document. We aim to enhance the capabilities of existing vision-and-language models in handling challenges posed by text-dominant documents in VRD-QA. Code and Appendix are in https://github.com/adlnlp/pdfmvqa.

1 Introduction

The growing demands for visually rich document (VRD) question-answering (QA) areas are becoming increasingly evident, especially in specialised fields such as finance and medicine. VRDs, including forms [Ding \textit{et al.}, 2023a], academic papers [Ding \textit{et al.}, 2023b], and industrial reports [Mathew \textit{et al.}, 2021a], typically comprise text-dense and visually rich components such as titles, paragraphs, tables, and charts. These components, \textit{document semantic entities}, are not only knowledge-intensive but are also organised in a pre-defined layout that maintains a logical and semantic correlation, usually extending across multiple pages. This complexity requires a more grounded and fact-dependent approach to QA. It is essential to comprehend the layout and logical structure of VRDs, especially in multi-page documents, to accurately locate and use these document entities as reliable evidence for answering knowledge-intensive questions. Recent generative models [Ouyang \textit{et al.}, 2022; Touvron \textit{et al.}, 2023; Liu \textit{et al.}, 2023a] have made impressive progress in providing interactive human-like responses by memorising vast knowledge [Zhao \textit{et al.}, 2023]. These models rely on plain text to learn textual content [Touvron \textit{et al.}, 2023] and use image patches to encode visual cues [Yasunaga \textit{et al.}, 2022]. This approach makes understanding document entities’ layout and logical relationships in VRDs difficult. Generative models are suffered from hallucinations [Ye \textit{et al.}, 2023], high costs [Hofst¨atter \textit{et al.}, 2023], and updating knowledge difficulties [Hu \textit{et al.}, 2023]. Retrieval-based QA [Liu \textit{et al.}, 2023b] addresses these limitations when applying generative models to VRD-QA. This approach helps locate answers or supporting evidence precisely, offering more grounded and factually dependent information. While recent retrieval-based applications mainly focus on web-crowded domains like Wikipedia [Hu \textit{et al.}, 2023], VRD-QA requires a deep understanding of domain-specific multimodal knowledge.

A few VRD-QA datasets [Mathew \textit{et al.}, 2021a; Tanaka \textit{et al.}, 2021] have been devised to extract in-line text from input document pages but often overlook prevalent multi-page scenarios. Recent multi-page datasets focus on extracting short phrases or sentences [Tito \textit{et al.}, 2023], causing recently proposed models [Huang \textit{et al.}, 2022; Yu \textit{et al.}, 2022] to excel at retrieving annotated in-line text but disregarding the logical and layout connections among document entities. Moreover, they are limited in handling the entire lengthy document. To address these limitations, entity-level document understanding tasks have been introduced by [Ding \textit{et al.}, 2023a] and [Ding \textit{et al.}, 2023b]. A common issue with these datasets is their text-dense mono-modal information extraction, overlooking visually rich entities such as \textit{tables} and \textit{figures}.

This paper proposes a new multi-page, multimodal document entity retrieval dataset, MMVQA, for knowledge-intensive domain. MMVQA addresses the limitations of generative models and expands upon the benefits of retrieval-based models by incorporating multimodal document entities like paragraphs, tables and figures and exploring the cross-page layout and logical correlation between them. This expansion supports the models to navigate and interpret real-
world documents at a multi-page or entire document level by leveraging joint-grained and multimodal information. The proposed models demonstrate how to effectively use existing VLPMs and pretrained language models with long sequence support to locate target entities from MMVQA.

The contributions are summarised as follows: We introduce MMVQA, a new VQA dataset for retrieving multimodal document semantic entities in multi-page VRDs, accompanied by versatile metrics for diverse scenarios. A set of frameworks for multi-page document entity retrieval is proposed by leveraging the implicit knowledge from VLPMs and fine-grained level information. A series of experiments are performed to provide deeper insights into MMVQA and demonstrate the effectiveness of our proposed techniques for multimodal multi-page document entity retrieval.

2 Related Work

The first document image-based QA dataset, DocVQA [Mathew et al., 2021b], includes scanned industrial documents. Questions in the DocVQA dataset are designed as in-line questions where the single-span answers and the keywords in questions are in the same line of text. Based on the DocVQA dataset document images, CS-DVQA [Du et al., 2022] proposed new questions requiring commonsense knowledge. Unlike extracting in-line answers on document pages, answers to CS-DVQA dataset questions could be the node of ConceptNet. RDVQA dataset [Wu et al., 2022], on the other hand, focuses on the question answering over coupon and promotion vouchers. Unlike the in-line questions, the RDVQA dataset proposed the in-region questions, which require the answer to be inferences from the information in the related region. In contrast to the single document page processing, DocCVQA [Tito et al., 2021] and SlideVQA [Tanaka et al., 2023] datasets proposed the question answering over the document collections. DocCVQA specifically focuses on a single document source, the US Candidate Registration Form. Due to the similar form layout and form fields, this dataset only proposed a limited number of in-line questions. However, multiple answer values could be extracted from multiple independent document images for answering one question. SlideVQA collects the set of slides, and there will be multiple answers to one question from different slide pages. Although DocCVQA and SlideVQA improve document VQA tasks to a multi-page level from the ordinary single page, their documents are not consecutive pages with dense texts. On the other hand, VisualMRC [Tanaka et al., 2021] collected the text-dense webpage screenshots, and questions are formed like in the machine reading comprehension task that requires the contextual understanding of textual paragraphs. However, VisualMRC limits the task scope to the single-page level. Existing datasets primarily extract text on MRC style and overlook visually rich elements like tables and figures. Current multi-page datasets mainly use sparse text sources, such as slides, while the demand is growing for text-dense documents. Our proposed MMVQA dataset aims to bridge these gaps by creating a multi-modal VRD-QA dataset that retrieves target document entities across multiple pages.学术文献是一个多篇文档的示例，其中每个文档都包含不同的研究领域、研究方法和研究结果。
4 Dataset Analysis

Document Components Statistics Our dataset includes only documents that contain multiple pages with numbers of tables/figures or includes complex structures of contents with multiple different sections and subsections. Based on the statistics\(^7\), we found the number of document components is quite consistent. Most documents contain around ten pages and have 10-20 different sections with around 20-40 paragraphs of 2000-4000 tokens. Hence, with the analysis, we can ensure that the collected documents are mostly lengthy and have a complex structure enough to evaluate the model’s feasibility to contextualise understanding over multiple consecutive pages. In addition to this, each document contains enough tables and figures to ensure the possible questions asked over these components. Most documents have around five tables and figures or more.

Super-Section Component Analysis We refer first-level section of each document as Super-Section, where the sections under the same Super-Section play similar structural roles in a medical domain academic paper, including Introduction (Intro), Material and Method (M&M), Result and Discussion (R&D), Conclusion (Concl) and Other \(^8\). Sections are categorised into Other Super-Section in documents, like Conflict of Interest, Funding, Ethical Approval, and Supplementary are less common but contain critical information.

The document layout statistics across Super-Sections are in Figure 2. The Materials and Methods (M&M) and Results and Discussion (R&D) sections are normally more complex, with multiple subsections, paragraphs, and most tables and figures. In contrast, the Introduction (Intro) and Conclusion (Concl) sections are simpler, with fewer subsections. The Other Super-Section, encompassing diverse contents like Supplementary or Fundings, has a larger interquartile range and more outliers, reflecting its varied nature.

Number of Question Distribution MMVQA contains 3,146 documents, which are a total of 30,239 pages. Each document is averagely associated with 84 questions, resulting in 262,928 question-answer pairs in MMVQA. The detailed Training/Validation/Test set size and the question number of each document Super-Section can be found in Table 1.

Super-Section-oriented Question-Answer Distribution The distribution of questions over each Super-Section is shown in Figure 3a. Most questions are asked over M&M and R&D sections, each having an average of around 17 questions. The average question length is in Figure 3b. Table/figure-related questions are longer, and the average question length of M&M sections is the shortest. For table/figure-related questions, answers to questions can be recognised from one document entity (segmented by a bounding box). For other Super-Section questions, answers may located in more than one document entity.

5 Task Definitions and Metrics

We introduce our main task as Multimodal Document Information Retrieval (DIR) aimed at retrieving semantic entities, such as paragraphs, tables, and figures, from the input entity sequence across multiple pages. As demonstrated by [Ding et al., 2023b; Gu et al., 2021], the document entity-level task encourages the exploration of logical and spatial relationships between semantic entities, and it is more straightforward to extend to the multi-page level compared to fine-grained token-level inputs. For instance, as shown in Figure 4, utilising document-entity sequences as input enhances both logical aspects (e.g., linking Table \(E_t\) with its corresponding Table Caption) and semantic understanding (e.g., handling split Paragraph entries \(E_{p1}\) and \(E_{p2}\)).\(^9\) Additionally, to address diverse application scenarios and effectively meet specific requirements, we introduced a set of distinct evaluation metrics for more adaptive performance assessment, including Exact Matching (EM), Partial Matching (PM), and Multi-Label Recall (MR). More details can be articulated in Section 5.2.

\(^7\)Please refer to Appendix C.2 to check the statistics chart.

\(^8\)Please check Appendix C.3 for more Super-Section analysis.

\(^9\)Token-level models struggle to capture entity-level correlations.
5.1 Task Definition

How is our multimodal DIR task conducted? Assuming \( Q \) is a natural language question and \( S_E = \{ E_1, E_2, \ldots, E_m \} \) is a set of document entities comprising \( m \) semantic entities of the target multiple document pages, \( S_{E_\text{gt}} = \{ E_1, \ldots, E_j \} \) represents the ground truth entity set for \( Q \). If a paragraph is divided into several regions, \( S_{E_\text{gt}} \) may include more than one entity (as in Figure 4). The task involves proposing a model \( F_{ir} \) with inputs \( Q \) and \( S_E \) to predict an entity set \( S_{E_{\text{pred}}} \). As in Figure 4, for a paragraph-based question \( Q_1 \), the ground truth set \( S_{E_{\text{gt}}_1} = \{ E_{p_1}, E_{p_2} \} \), where \( E_{p_1}, E_{p_2} \) belong to the same paragraph but are split into two regions. For a table/figure-based question \( Q_2 \) in Figure 4, the ground truth set only contains the table entity \( E_t \).

5.2 Evaluation Metrics

Distinct evaluation metrics cater to the varied application scenarios of retrieved entities. These metrics encompass stringent exact-match accuracy to more lenient measures, allowing partial retrieval and multi-label recall and providing a comprehensive performance assessment. **Exact Matching Accuracy (EM)** is a stringent metric suitable for scenarios requiring precise, unambiguous information retrieval, particularly when used as supporting evidence or reliable references. We also introduced **Partial Matching Accuracy (PM)** with tolerance for partial matches. It is especially beneficial when capturing every relevant entity is less crucial than ensuring the correctness of the predicted entities, such as ensuring the correct identification of the primary entity \( E_{p_1} \) in a target paragraph. **Multi-Label Recall (MR)** is applied to assess the proportion of correctly identified actual positives in situations where identifying all positive instances is critical. We provide the detailed definitions of each metric in Appendix D.

6 Methodology

6.1 Multimodal Multi-Page Retriever

Existing document understanding models [Huang et al., 2022; Kim et al., 2022; Wang et al., 2022; Li et al., 2021] and datasets [Mathew et al., 2021a; Tanaka et al., 2021] are designed for single-page document comprehension, relying on token-level representations. However, the fine-grained token-level information suffers from the limited length. It neglects the correlations between document entities, particularly in capturing long contextual dependencies in more prevalent multi-page scenarios. Instead of employing sequences of tokens that lead to significant memory consumption, we introduce a multimodal entity-level retrieval framework \( \mathcal{R} \) to identify the target entity set \( S_Q \) from the cross-page entity sequence in a given question \( Q \), as illustrated in Figure 5.

The input, comprising multiple pages, consists of a set of document entity embeddings \( \mathbb{E} = \{ E_1, E_2, \ldots, E_n \} \). These embeddings, elaborated in Section 6.2, are combined with 1D positional encoding \( \mathbb{P} \), bounding box embedding \( \mathbb{B} \), and label embedding \( \mathbb{L} \). The combined representation, \( \mathbb{E} + \mathbb{P} + \mathbb{B} + \mathbb{L} \), is fed into the **multimodal Entity Encoder** \( \mathcal{E} \), alongside the question token embeddings \( \mathbb{Q} = \{ q_1, q_2, \ldots, q_m \} \) and additional context elements like image patch embeddings \( \mathbb{P} \). The encoder \( \mathcal{E} \) models the correlations among these entities, the question, and other contexts. The enhanced entity representation \( \mathbb{E}' \) from \( \mathcal{E} \), along with \( \mathbb{Q} \), serves as input for a transformer-based **Multimodal Entity Decoder** \( \mathcal{D} \), producing the final representation \( \mathbb{E}'' \). Each entity in \( \mathbb{E}'' \) is linearly projected by a **Entity Recogniser** \( \mathcal{L}_{er} \) for binary classification, distinguishing target entities (label 1) from non-target entities (label 0) in the context of the question \( Q \) and Entity Set \( \mathbb{E} \).

6.2 VLPM Augmented Retriever

Existing Vision Language Pre-training Models (VLPMs) can be classified into two categories based on their focus on visual cues: Region-of-Interest (RoI)-based and Image Patch-based [Long et al., 2022]. RoI-based models utilise features from ground truth or predicted regions, while Patch-based models process segmented image patches. Even though these VLPMs are initially pretrained on general photo-like image-related tasks rather than visually-rich documents, previous studies have illustrated the feasibility of employing VLPMs such as [Li et al., 2019; Tan and Bansal, 2019; Kim et al., 2021] in tasks related to understanding documents. Thus, we propose methods to harness the implicit information embedded in pretrained VLPMs for obtaining more comprehensive and robust representations of multimodal entities.

**RoI-Based Frameworks**

RoI-based VLPMs focus on learning the contextual entity relationships and correlation between textual content and as-
associated visual cues of each RoI, which in our scenario are document-semantic entities (e.g., section, paragraph, table, etc.). $F_{roil}$ donates a RoI-based VLPM backbone. This backbone takes a question token sequence $Q$ and a set of visual representations $V$ as input, where $V = \{V_1, V_2, \ldots, V_n\}$ signifies the initial visual representations of each entity in the document $D$. Our objective is to generate an improved visual embedding set $V'$, capturing the contextual relationships among entities and their correlation with the question. Then, $V'$ is concatenated with textual embedding $T$ and fed into a linear Vision-Textual Projector $L_{vt}$ to produce the entity representation set $E$ for input into the retriever $R$. We employ vanilla Transformer as a foundational benchmark for evaluating the impact of various pretrained techniques in comparative studies [Ding et al., 2023a]. Additionally, we introduce VisualBERT [Li et al., 2019] and LXMERT [Tan and Bansal, 2019] to enhance the initial visual embedding of each document entity. The improved visual embeddings are concatenated with $T$ to obtain $E$.

Image Patch-Based Frameworks

Recently emerged VLPMs commonly employ image patches without prior RoI bounding box information, a practice also observed in document understanding frameworks designed for single-page scenarios [Xu et al., 2021; Huang et al., 2022]. Despite these advancements, the demands of cross-page document understanding remain insufficiently addressed. Consequently, our research investigates the effectiveness of image-patch-based VLPMs in the general domain in cross-page information retrieval tasks. Extensive experiments and analyses are conducted to evaluate the effectiveness of patch-based methods in enhancing entity representation in cross-page document information.

To apply a vision-language model for cross-page document understanding, we first merge multiple document pages $I = \{I_1, I_2, ..., I_m\}$ into a composite image $I$. After that, the resized image and question are fed into VLPM processors to produce image patch pixel and question token sequences, which are the inputs of corresponding Patch-based VLPM encoders. The generated patch embedding $P = \{p_1, p_2, ..., p_l\}$ and the question token embedding $Q$ are combined with the entity embedding $E$ and fed into a Multi-modal Entity Encoder $E$ within the retriever $R$, facilitating contextual learning between them. Then, we can get $Q', P', E' = \mathcal{E}(Q, P, E)$, where $E = L_{vt}(V \oplus T)$. $E'$ and $Q$ are fed into the Multimodal Decoder Entity Decoder $D$ within $R$ as target embedding and memory embedding for the retrieval process. We introduce patch-based VLPMs to obtain contextual patch embedding $P$, including models such as CLIP [Radford et al., 2021], ViLT [Kim et al., 2021], BridgeTower [Xu et al., 2023].

6.3 Joint-Grained Retriever

Entity-level document understanding models can gain advantages by incorporating logical and layout relationships to improve entity representations. However, overlooking fine-grained details, such as crucial phrases and sentences within

![Figure 6: Joint-grained(coarse-and-fine grained) Retriever](image)

Table 2: Overall performance under various evaluation metrics.

<table>
<thead>
<tr>
<th>Type</th>
<th>Model</th>
<th>EM</th>
<th>PM</th>
<th>MR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rol-based</td>
<td>Transformer</td>
<td>17.92</td>
<td>19.46</td>
<td>22.48</td>
</tr>
<tr>
<td></td>
<td>VisualBERT</td>
<td>15.39</td>
<td>17.80</td>
<td>21.92</td>
</tr>
<tr>
<td></td>
<td>LXMERT</td>
<td>17.81</td>
<td>19.77</td>
<td>23.37</td>
</tr>
<tr>
<td>Patch-based</td>
<td>CLIP</td>
<td>20.71</td>
<td>22.55</td>
<td>25.70</td>
</tr>
<tr>
<td></td>
<td>ViLT</td>
<td>21.71</td>
<td>23.47</td>
<td>27.56</td>
</tr>
<tr>
<td></td>
<td>BridgeTower</td>
<td>19.88</td>
<td>22.37</td>
<td>23.99</td>
</tr>
<tr>
<td>Joint-grained</td>
<td>BridgeTower w/ OCR</td>
<td>21.62</td>
<td>22.56</td>
<td>26.63</td>
</tr>
<tr>
<td></td>
<td>BridgeTower w/ PDF Miner</td>
<td>21.53</td>
<td>23.25</td>
<td>26.90</td>
</tr>
</tbody>
</table>

11 For detailed model configurations, please refer to Appendix E.1.
12 For further configuration details, please refer to Appendix E.2.

To assess the effectiveness of Rol-based and Patch-based frameworks in retrieving entities from multi-page documents under different scenarios, performance metrics (EM, PM and MR)
**7.2 Joint-Grained Framework Results**

**Overall and Super-Section Breakdown Performance**

To illustrate the effectiveness of the proposed Joint-grained framework (Figure 6), we conducted a performance comparison between the top two vanilla frameworks on paragraph-based questions from both the **RoI-based** (Transformer and LXMERT) and **Patch-based** (ViLT and BridgeTower) groups and their respective Joint-grained architectures by feeding the provided context attribute of each question. Overall, Joint-grained models consistently improve performance, with LXMERT and BridgeTower showing more than a 2% increase. Regarding Super-Sections, complex Super-Sections like M&M and R&D benefit notably, especially BridgeTower, which improves by around 4% in M&M and 3.5% in R&D. Super-Sections with simple complexity (Intro and Concl) see less improvement, and the **Conclusion** (Concl) even performance decreases, especially in Patch-based frameworks (around 6% decrease). These trends suggest that fine-grained information enhances the understanding of text-dense entity textual representations by capturing important words or phrases missed at the entity level.

**Page Range-Based Breakdown Analysis**

To assess the Joint-grained framework’s robustness across different input page numbers, we conducted a comparative analysis, shown in Figure 7. Figure 7a indicates that the Joint-grained framework enhances performance with smaller page gaps but experiences a decrease in performance with larger input page numbers. This suggests that fine-grained information may improve document entity representations. But, with the number of input pages increasing, textual tokens may introduce more noise that adversely affects document entity representations. Exploring additional Joint-grained mechanisms may help enhance entity representations. However, as shown in Figure 7b, Joint-grained frameworks notably enhance robustness in MR-oriented scenarios, from smaller to larger numbers of pages. This highlights that incorporating fine-grained textual information can aid the model in locating target entities even in long, visually rich document scenarios.

**7.3 Real-World Scenarios**

Table 4: Comprehensive Breakdown Performance: BridgeTower Joint-grained frameworks based on various sourced textual token sequences, overall and super-Section based breakdown.

To demonstrate the real-world efficacy of our proposed Joint-grained framework, we evaluated its performance using text extracted from off-the-shelf tools. Because BridgeTower, highlighted in Table 3, exhibits significant improvements, we present the performance of BridgeTower-based Joint-grained frameworks on various text token sequences from the MMVQA dataset (Jg-BridgeTower), PDF parser (Jg-BridgeTower-PDFMiner), and OCR tools (Jg-BridgeTower-OCR). As shown in Table 4, incorporating fine-grained text-
tual information results in performance enhancements, increasing from 22.37% to 23.56% (PDFMiner) and 23.25% (OCR) in overall. In addition, high structural complexity sections (e.g., M&M, R&D) show notable improvements, particularly in MMVQA, reaching around 4.5% in M&M and 3.5% in R&D. This may be attributed to the “context” provided by the MMVQA dataset, extracted from XML nodes containing prior information. Despite inherent noise raised by off-the-shelf tools, they still yield substantial improvements. Notably, OCR, while facing challenges with mis-detected characters, demonstrates considerable increases in retrieving Table (about 5%) and Figure (7%) based questions. However, Introduction (Intro) shows a decreasing trend after the incorporation of fine-grained information. This could be due to the introduction covering the entire document content, making learning the relations between tokens and entities more challenging. Future work may explore more refined Joint-grained aligning methods.

7.4 Category-Oriented Entity Representation

Figure 8: Category-oriented entity representation T-SNE analysis of various frameworks including (a) Transformer, (b) VisualBERT, (c) LXMERT, (d) CLIP, (e) ViLT, (f) BridgeTower, (g) Jg-BridgeTower-PDFMiner, (h) Jg-BridgeTower-OCR.

To understand the insight of document entity representations of each framework, two-dimensional T-SNE analysis is performed on final entity embeddings extracted from decoder $D$, as shown in Figure 8. In general, RoI-based frameworks tend to have more representative feature embedding in understanding the semantic roles of each document entity. Especially compared with unclear boundaries between various text-dense entities such as Abstract, Title, Paragraph, RoI-based models can effectively distinguish them. However, RoI-based models underperform compared to Patch-based models, as shown in Table 2. The possible reason is although they benefit from pre-trained backbones and are good at learning visual cues within document entity RoIs, they lack in addressing the broader document layout and the relationships between question and target entities, crucial for understanding multi-page documents.

14We conducted an additional question-answering embedding correlation analysis in Appendix G.2.

8 Conclusion

This paper presents a contribution by introducing the MMVQA dataset and a novel joint-grained architecture. The MMVQA from PubMed Central showcases diverse document types, complex structures, and extensive content-related questions in multi-page documents. We also introduce the strong benchmark, Joint-grained retrieval architecture, which consistently enhances model performance, particularly in complex document sections. We hope this research could not only advance the understanding of multi-page document comprehension but also set a foundation for future exploration and refinement of models in this domain, marking a significant step forward in document understanding research.

15Please refer to Appendix H to check more qualitative samples.
Acknowledgments

We express our profound gratitude to all the authors—Yihao Ding, Kaixuan Ren, Jiabin Huang, Siwen Luo, and Soyeon Caren Han—for their critical contributions to this project. Their combined expertise, associated with The University of Melbourne, The University of Sydney, and The University of Western Australia, has been essential in advancing this research. We are grateful for the unwavering support from these institutions, which provided the necessary resources and conducive environments for our studies. Furthermore, we appreciate the insightful feedback from our peers and reviewers, which has greatly enhanced the quality of our work. We hope that our research will make a meaningful impact in the field.

References


[Ouyang et al., 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin,


