A High-Quality Text-Rich Image Instruction Tuning Dataset via Hybrid Instruction Generation

Shijie Zhou[1], Ruiyi Zhang[2†], Yufan Zhou[2], Changyou Chen[1]
[1]University at Buffalo [2]Adobe Research

Abstract

Large multimodal models still struggle with text-rich images because of inadequate training data. Self-Instruct provides an annotation-free way for generating instruction data, but its quality is poor, as multimodal alignment remains a hurdle even for the largest models. In this work, we propose **LLaVAR-2**[1], to enhance multimodal alignment for text-rich images through hybrid instruction generation between human annotators and large language models. Specifically, it involves detailed image captions from human annotators, followed by the use of these annotations in tailored text prompts for GPT-4o to curate a dataset. It also implements several mechanisms to filter out low-quality data, and the resulting dataset comprises 424k **high-quality** pairs of instructions. Empirical results show that models fine-tuned on this dataset exhibit impressive enhancements over those trained with self-instruct data.

1 Introduction

Instruction tuning is widely used to improve the generalization and controllability of large language models (LLMs) (Wang et al., 2022; Ouyang et al., 2022; Zhang et al., 2023b) by converting unseen tasks into instruction-output pairs. Adopting such a manner, visual instruction fine-tuning (Liu et al., 2023a) enables the visual reasoning capacity of multimodal large language models (MLLMs) by injecting aligned visual tokens leveraging visual encoders such as CLIP-ViT (Alexey, 2020; Radford et al., 2021) and DINO (Caron et al., 2021; Tong et al., 2024b). However, obstacles persist in effectively handling text-centric visual tasks for MLLMs. These challenges likely arise from the underrepresentation of text-rich images in the training dataset, such as COCO (Lin et al., 2014) and Conceptual Captions (Changpinyo et al., 2021). However, the ability to comprehend texts within images is essential for real-world applications. Classical datasets on text-rich images are dedicated to information extraction, such as TextVQA (Singh et al., 2019), TextOCR (Singh et al., 2021), DocVQA (Mathew et al., 2021), and OCR-VQA (Mishra et al., 2019). Recent instruction tuning datasets for text-rich images focus more on classical document images, such as Figure (Kahou et al., 2017), Chart (Masry et al., 2022) , and infographics (Mathew et al., 2022).

To address this issue, LLaVAR (Zhang et al., 2023c) introduces noisy and GPT-4-based instruction-following data of text-rich images, utilizing OCR and caption tools to enhance textual comprehension ability. TRINS (Zhang et al., 2024) creates a text-rich image instruction dataset in a semi-automatic manner with manual annotation effects to guarantee high-quality captions. In this work, we present LLaVAR-2, a text-rich image instruction-following dataset via hybrid instruction generation between human annotations in TRINS and LLMs to improve the effectiveness of visual instruction tuning. Specifically, we enrich the collected manual captions and the QA dataset with fine-grained details and supplementary self-explain instruction data.

LLaVAR-2 consists of two parts: LLaVAR-2-Cap for global descriptive captioning on images and LLaVAR-2-VQA for visual question answering. For LLaVAR-2-Cap data collection, rather than directly applying the human-annotated captions from TRINS as answers for crafted instructions, we rewrite the caption by incorporating necessary text/visual details to get detail-enriched captions. Based on it, we construct LLaVAR-2-Cap for precise summarizing to facilitate global visual text understanding. For instruction tuning data, we introduce a novel approach that combines extractive question answering with supplementary rounds of self-explain conversations. These pairs of self-

†Corresponding Author
[1]Project page: https://github.com/llavar/LLaVAR-2

Proceedings of the 31st International Conference on Computational Linguistics, pages 10091–10110
January 19–24, 2025.

explanations serve to illuminate the rationale behind extractive answers, supported by detailed examinations of relevant image contents. Data with extra pairs of self-explaining bolsters the localized comprehension of text-rich images, offering deeper insights and clearer connections within the visual data. Leveraging the initial high-quality human-annotated captions, VQA and captioning data of LLaVAR-2 are generated via GPT-4o. Compared with TRINS (Zhang et al., 2024), LLaVAR-2 shows better diversity in task categories, the length of instructions, and answers. To filter out low-quality data samples, we propose an automatic filtering mechanism for the multimodal instruction tuning dataset, named multimodal Instruction-following Difficulty (mIFD) score and Fact-Following Difficulty (FFD) score, to filter out incompatible extraction and self-explain pairs in VQA data. Our contributions are as follows.

- We present **LLaVAR-2**, a novel dataset consisting of 42k detail-enriched captions and 382k visual question-answering data pairs, all generated automatically using GPT-4o based on human-annotated text-rich image captions.
- We design mIFD and FFD scores for filtering on LLaVAR-2-VQA that systematically removes irrelevant or redundant data, ensuring the dataset's high quality.
- Beyond demonstrating the dataset's diversity through statistical visualization, we show the superiority of LLaVAR-2 by fine-tuning various base models, showing great improvement on various benchmarks.

2 Related Work

Multimodal Large Language Models Recent significant success in multimodal large language models can be traced back to Flamingo (Alayrac et al., 2022), InstructBLIP (Dai et al., 2023) , and LLaVA (Liu et al., 2023b). While Flamingo and InstructBLIP form the bridge between language and vision via cross-modal attention, LLaVA-style methods transform visual representations from a standalone visual encoder into visual tokens that language models can understand. Some latest examples include Eagle (Shi et al., 2024), Cambrian-1 (Tong et al., 2024a), LLaVA-HR (Luo et al., 2024), InternVL2 (Chen et al., 2023b, 2024b), Idefics2/3 (Laurençon et al., 2024b,a), LLaVA-OneVision (Li et al., 2024a) and Qwen2-VL (Wang et al., 2024). Scaling up the pre-training and SFT data is a crucial part of improving MLLMs. MiniGPT-4 (Zhu et al., 2023) uses ChatGPT to produce data compliant with high-quality instructions, while LLaVA (Liu et al., 2023b) prompts GPT-4 with image captions and boxes to similar ends. Other initiatives (Chen et al., 2023a, 2024a) prompt GPT-4V to generate more than 1 million pieces of quality data for the training of MLLMs. LLaMA-Adapter (Zhang et al., 2023a; Gao et al., 2023) synchronizes text and image features using COCO dataset inputs. InstructBLIP (Dai et al., 2023) has restructured 13 vision-language tasks to fit an instruction-based approach. mPLUG-Owl (Ye et al., 2023a,c) implements multi-task instruction fine-tuning with pre-existing document datasets. VILA (Lin et al., 2024) equips extra interleaved visual language corpus into MLLMs' pretraining. Cambrian-1 (Tong et al., 2024a) and Idefics2 (Laurençon et al., 2024b) create large pools of instruction-tuning data containing a diverse range of visual-language tasks, such as counting, captioning, document understanding, etc. Further research (Liu et al., 2023a, 2024a; Bai et al., 2023; Dong et al., 2024; Xu et al., 2024; Luo et al., 2024) explores enhancing encoder resolution, leading to significant advancements in a variety of downstream applications. For a detailed examination, a comprehensive survey is provided (Li et al., 2023a). Despite these advances, visual-text comprehension remains a challenge for many models (Liu et al., 2023c).

Instruction-Following Data Generation To address the lack of sufficient instruction-following data, data synthesis is commonly employed for the training of LLMs/MLLMs. Recent studies have focused on producing high-quality synthetic instruction-following data by leveraging strong and robust LLMs (Wang et al., 2022; Xu et al., 2023; Chen et al., 2023a; Zhang et al., 2023c; Zhao et al., 2023; Zhang et al., 2024). For instance, Self-Instruct (Wang et al., 2022) boosts LLMs' instruction-following capabilities by allowing them to generate their instructional data. WizardLM (Xu et al., 2023) developed Evol-Instruct to create instruction data with varying levels of complexity. In ShareGPT4V (Chen et al., 2023a), the 1.2M image captioning data is expanded by a superb caption model trained on the initial subset. LLaVAR

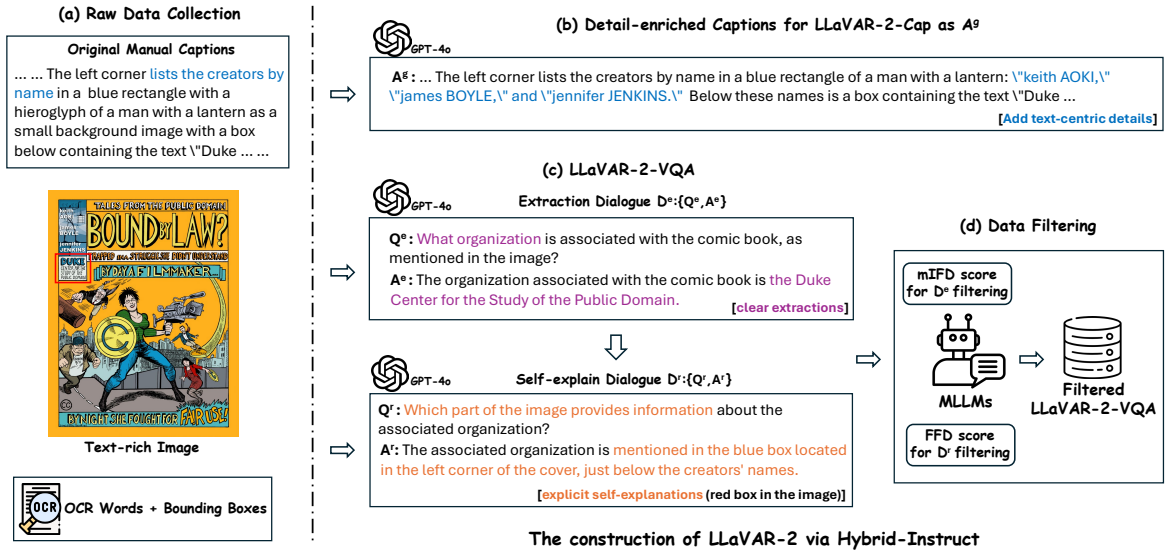


Figure 1: Overview of the data collection pipeline of Hybrid-Instruct for constructing LLaVAR-2. (a) Prompting GPT-4o with the text-rich image, manual caption, and the OCR results; (b) The Original manual caption is rewritten automatically to the detail-enriched caption, which is further used as A^g to form LLaVAR-2-Cap; (c) Generated LLaVAR-2-VQA is composed of Extractive dialogue D^e and Self-explain Dialogue D^r which supplement explicit extraction process for D^e; (d) Besides, mIFD and FFD scores are proposed to filter D^e and D^r respectively.

(Zhang et al., 2023c) targeted instruction-following data generation for text-rich images by additionally prompting LLMs with high-quality demonstrations. TRINS (Zhang et al., 2024) further improved upon LLaVAR by leveraging manually crafted high-quality image captions instead of BLIP-2 generated ones to generate instructions. Our proposed dataset LLaVAR-2 enriches necessary text and visual details into instruction-following data and maintains the data quality via proposed filtering scores, thus improving visual instruction tuning.

3 Hybrid Instruction Generation

In Figure 1, we show the data collection pipeline of **Hybrid Instruction** including the detail-enriched captions LLaVAR-2-Cap in Section 3.1 and the visual question answering data LLaVAR-2-VQA collection in Section 3.2. We conduct data filtering for LLaVAR-2-VQA, with the filtering process described in Section 3.3. Examples of LLaVAR-2-Cap and LLaVAR-2-VQA are given in Appendix C and D respectively.

3.1 Detail-enriched Captions

Manual captions in TRINS provide concrete descriptions of visual components but often offer only general and succinct summaries for text-heavy areas, lacking explicit text labels for these summaries. Thus, clear links between descriptions and corresponding text or image regions are hard to construct for MLLMs. To address this, based on OCR results, we leverage advanced GPT-4o to enrich these captions with missing details and text labels.

Specifically, to enhance global awareness of precise geometric relationships in text-rich images, we combine the text-rich image, OCR results with bounding boxes and manual caption automatically by prompting them to GPT-4o and asking for captions with more text labels while preserving the original style and structure of its manual caption. The system message used is detailed in Appendix B.

Furthermore, although the manual results may contain errors, OCR bounding boxes help GPT-4o better locate text and correct inaccuracies as shown in the example of Figure 1. These detail-enriched captions are further paired with descriptive questions to create single-round caption data D^g:{Q^g, A^g}, named **LLaVAR-2-Cap**, which includes detailed visual text and object descriptions.

3.2 Visual Question Answering with Self-explanations

Visual question-answering data is crucial for visual instruction tuning and typically includes extractive QA data that focuses on the local attributes of an image I. In LLaVAR-2-VQA, each extractive QA pair D^e:{Q^e, A^e} is accompanied by a self-explain pair D^r:{Q^r, A^r} to illuminate the rationale behind the extractive answer A^e. The detailed data collection

Extractive Q1: What does the background of the magazine cover show?
Extractive A1: The background shows a view of a beautiful city with buildings, a bridge over the sea, and a person walking past it.
Self-explain Q1: What visual elements in the background contribute to the description of the cityscape on the magazine cover?
Self-explain A1: The background includes a view of a city with buildings and a bridge over a body of water, along with a person walking past it. These elements collectively create a picturesque representation of a vibrant urban landscape, contributing to the description of a beautiful city.

Extractive Q2: What could be the focus of the article "MIDNIGHT CALLING"?
Extractive A2: The article "MIDNIGHT CALLING" seems to focus on the nightlife and after-dark hotspots in Dublin. It likely highlights the best places to visit, such as bars, clubs, late-night eateries, and other entertainment venues that come alive at night. This article would appeal to readers interested in exploring the city's vibrant night scene.
Self-explain Q2: How does the article title "MIDNIGHT CALLING" and its positioning on the cover suggest its content?
Self-explain A2: The article title "MIDNIGHT CALLING" is positioned in the middle left of the cover in bold yellow text, indicating its importance. The title itself suggests content related to nightlife and after-dark activities, implying that the article will highlight hotspots and entertainment venues that are popular during the night.

LLaVAR-2-Caption: This is the cover of a magazine titled "dylan MAGAZINE (Issue 02)", prominently displayed at the top. The cover highlights various articles featured in the magazine. On the top left, it mentions "DESIGN DUBLIN" with the subtitle "Meet six of the city's most inspiring creatives." Below that, it features "MIDNIGHT CALLING" with the subtitle "After dark hotspots." On the top right, it highlights "12 HOURS IN RANELAGH," describing it as "The capital's hippest hood." The background image showcases a scenic view of a city with buildings and a bridge over water. A person is seen walking along a wooden pathway, adding a dynamic element to the serene setting. At the bottom, the cover encourages readers to "GET OUTSIDE" with a guide to Dublin outdoors, with the text "WITH OUR GUIDE TO DUBLIN OUTDOORS" positioned below it. Additionally, on the right side, it mentions "THE BEST PLACES TO EAT, SHOP AND RELAX IN THE CITY."

Figure 2: An example of LLaVAR-2 dataset: a text-rich image of a magazine cover, its LLaVAR-2-Caption (Section 3.1) and LLaVAR-2-VQA containing extractive and self-explain components (Section 3.2).

pipeline for D^e and D^r is described following.

Extractive Question Answering Data In addition to these globally descriptive instruction-following data, instructions focusing on specific local attributes of text-rich images are also necessary. Thanks to the high-quality human-annotated captions, which clearly indicate each attribute of both text and visual areas in the image, it is convenient to utilize GPT-4o to generate conversational data that focuses on partial attributes in a local area. To achieve this, we prompt GPT-4o with the text-rich image, its human-annotated caption, OCR results from PaddleOCR, and two in-context demonstrations following (Zhang et al., 2023c). Applying the system message in (Liu et al., 2023b), this setup generates multiple single-turn conversations D^e:$\{Q^e, A^e\}$ around an image, each focusing on different components without overlapping. For each single-turn conversation, we formulate the question and answer as the input instruction and the target response, respectively.

Self-explain Conversational Data While the extractive instruction-following data D^e focusing on local attributes will facilitate the understanding of text-rich images, it still presents challenges for MLLMs to answer extractive questions when the model needs to link a visual area with a text area or when the question involves one text area but the answer is found in another.

In these cases, precise associative and implicit reasoning is necessary. Although D^e provides accurate retrieval labels, it lacks clarity on how its extraction is achieved. The answering process behind D^e can supplement the extract result in A^e with text/geometric labels, such as clear indications of where the query object is located in the image, thus improving MLLMs for better recognition and extraction.

To help MLLMs handle challenging scenarios, we make the implicit extraction process explicit by generating a self-explain conversation D^r:$\{Q^r, A^r\}$ as the explanation for each pair D^e. Prompting GPT-4o with the text-rich image, manual caption, OCR results, corresponding D^e, and system messages, the self-explain dialogue D^r is generated to explain how the extractive answer D^e is obtained, highlighting implicit connections between different components. Please refer to Figure 1 for an example of the two dialogues. The self-explain dialogue D^r that is based on D^e will enhance MLLMs' ability to recognize implicit connections in images, improving performance in tasks such as extractive QA and text-rich image analysis, while avoiding verbosity if directly adding reasoning into extractive answers. The system message used for generating D^r is detailed in Appendix B.

3.3 Instruction Data Filtering

Both initial extractive QA D^e and self-explain QA D^r of LLaVAR-2-VQA generated in Section 3.2 may include low-quality samples. Specifically, the low-quality extractive QA D^e generally fails in two cases: (1) the entire D^e is barely related to the input image i but focuses on an unfamiliar or too abstract topic; (2) the extractive answer A^e is anti-intuitive or unrelated to the extractive question A^e.

Extractive Q1: What type of music and dance style is being celebrated in the event?
Extractive A1: The event celebrates a style of Jamaican reggae called "Rub-a-Dub," which emerged in the 70s and 80s. The term Rub-a-Dub comes from a dance style where a man and woman rub up very close together."

Extractive Q2: When and where is the event taking place?
Extractive A2: The event is taking place on February 12th, 2017, at 255 Trinity Ave, Atlanta, GA.
Self-explain Q2: What information in the image specifies the date and location of the event?
Self-explain A2: The date and location of the event are specified in the red box in the lower center of the image. It states "February 12th, 2017" and "255 Trinity Ave, Atlanta, GA ".

Extractive Q3: Who is the event a tribute to?
Extractive A3: The event is a tribute to Bob Marley.
Self-explain Q3: How is it indicated that the event is a tribute to Bob Marley?
Self-explain A3: The tribute to Bob Marley is indicated by the large text in the middle of the image that reads "3rd Annual Bob Marley Tribute".

LLaVAR-2-Caption: This is an advertisement for the 3rd Annual Bob Marley Tribute, presented by Highlanda.net and The Honorary Citizen. The cover features a stylized image of Bob Marley. The main title in the middle of the cover reads "3RD ANNUAL BOB MARLEY TRIBUTE". At the bottom center, in a red box, it mentions the event "WILD PITCH" on February 12th, 2017, located at 255 Trinity Ave, Atlanta, GA. The event timing is from 6 PM to 1 AM, and tickets are available at Highlanda.net for $15 online and $20 at the door. Additional details include an early live reggae mixer from 6-8 PM featuring Ijahknowah and the B.R.A.P Band. The Rub-A-Dub Session from 8 PM to 1 AM will feature Highlanda Sound, Agard, DJ Passport, Black Magic, Natural Vibes, and Innocent Sound. There will be eats by Webba's Jerk Hut.

Figure 3: A poster example of LLaVAR-2 dataset.

While both cases may happen together, we develop an effective filtering strategy based on Instruction Following Difficulty(IFD) (Li et al., 2023c). Given the instruction, the corresponding answer, and the model θ, IFD is calculated as:

$$\mathrm{IFD}_\theta(Q, A) = \frac{s_\theta(A \mid Q)}{s_\theta(A)}, \tag{1}$$

where s_θ is the sum of the next token prediction loss of the input computing by the model θ. $s_\theta(A \mid Q)$ inputs model θ with $\{Q, A\}$ and sums the loss of each A's token. $s_\theta(A)$ is computed without the instruction A input. IFD can measure how helpful the instruction is towards the generation of the answer, and smaller IFD reflects better correspondence and compatibility between them. Inspired by IFD, we propose **M**ultimodal **I**nstruction-**F**ollowing **D**ifficulty (mIFD) to filter D^e as:

$$\begin{aligned}&\mathrm{mIFD}_\theta(Q_i^e, A_i^e, I_i)\\ &= \frac{s_\theta(Q_i^e, A_i^e \mid I_i)}{s_\theta(Q_i^e, A_i^e)} \times \frac{s_\theta(A_i^e \mid Q_i^e)}{s_\theta(A_i^e)}\\ &= \mathrm{VFD}(D_i^e, I_i) \times \mathrm{IFD}(Q_i^e, A_i^e)\end{aligned} \tag{2}$$

In mIFD score, the first term Visual-Following Difficulty (VFD) takes the entire extractive dialogue D^e and the image as the input to evaluate the correspondence between them. The second term directly utilizes the IFD score to detect samples in case (2). Computing mIFD scores on D^e, we excluded the highest 70% of samples. Examples filtered out are shown in Figure 15. We can find that many unrelated answers A^e are generated based on MLLMs' pre-knowledge of the object.

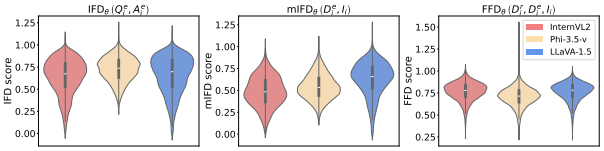


Figure 4: The distributions of $\mathrm{IFD}(Q_i^e, A_i^e)$, $\mathrm{mIFD}_\theta(Q_i^e, A_i^e, I_i)$ and $\mathrm{FFD}_\theta(D_i^e, D_i^r, I_i)$ computing by different MLLMs.

Case 1 : FFD score = 0.9484

Extractive

Q: What additional information is included in the title?
A: The title includes "KEY TAKEAWAYS, SUMMARY & ANALYSIS INCLUDED.

Self-explain

Q: Why is the additional information "KEY TAKEAWAYS, SUMMARY & ANALYSIS INCLUDED" significant?
A: The additional information "KEY TAKEAWAYS, SUMMARY & ANALYSIS INCLUDED" is significant because it indicates that this edition of the book provides a comprehensive overview, including key points, summaries, and analyses. This makes the content more accessible and understandable for readers. **[irrelevant and nonspecific contents]**

Case 2 : FFD score = 0.1796

Extractive

Q: How does the cover image look?
A: The cover image features a small figure approaching a blue sci-fi pyramid with floating rocks around it. There is an ornate sci-fi pattern graphic with three full moons under and around the title.

Self-explain

Q: What elements are depicted in the cover image?
A: The cover image depicts a small figure approaching a blue sci-fi pyramid with floating rocks around it, and there is an ornate sci-fi pattern graphic with three full moons under and around the title. **[repetitive text]**

Figure 5: Example of FFD-based filtering. In case 1, the self-explain pair is composed by not specific contents barely related to the extractive pair and image; In case 2, the self-explain pair repeats the similar question-answering content.

As for the data filtering of self-explain data D^r, we find abnormal D^r sometimes falls into two cases: (1) D^r focuses on a new concept unrelated to the image and D^e; (2) D^r is over-corresponded

with D^e leading to A^r very similar with A^e. Examples of 2 cases are shown in Figure 5.

Extending IFD to the dialogue level, we propose the Fact-Following Difficulty (FFD) score to reflect the closeness between the extractive data D^e and the self-explain data D^r, which is formed as:

$$\text{FFD}_\theta(D_i^e, D_i^r, I_i) = \frac{s_\theta(D_i^r \mid D_i^e, I_i)}{s_\theta(D_i^r \mid I_i)}. \quad (3)$$

From Figure 5, high FFD reflects unrelated D^r and D^e, while low FFD score indicates repetitive answers. Utilizing Phi-3.5-vision to compute FFD, we filter out 1.5k poorly correlated and 5.6k over-related self-explain pairs D^r. Examples of filtered-out data are shown in Appendix E.

Furthermore, we verify the consistency of mIFD and FFD score computing by different MLLMs. Their distributions are shown in Figure 4. We can observe that their FFD scores share a very similar FFD pattern. While their IFD scores have slight differences, their mIFD scores computed using the IFD term keep these differences consistent, e.g. IFD computed by LLaVA-1.5 shows a narrower lower tail and this difference remains in mIFD.

4 Enabling LLaVA to better Read

To verify the benefits of LLaVAR-2 for visual text understanding, we propose LLaVAR-2-3.8B following the similar architecture of LLaVA (Liu et al., 2023b). LLaVAR-2-3.8B is an efficient multimodal language model leveraging the small-scale but effective `microsoft/Phi-3-mini` (Abdin et al., 2024) as the language decoder D and adapting the Mixture-of-Resolution Adaptation (MRA) in LLaVA-HR (Luo et al., 2024) to build the vision encoder V. Specifically, `CLIP-ViT-L` and `CLIP-ConvNeXt-L` are utilized to encode low- and high-resolution images. High-resolution features are embedded into low-resolution paths by applying MRA. Integrating high-resolution embeddings in this manner enhances the MLLM's image comprehension ability, especially for text-rich images, while maintaining efficiency without extra visual tokens. The grid embeddings before the last layer of the transformer are aligned to the word embedding space via a two-layer MLP projector.

We conduct two-stage training following LLaVA-HR that optimizes the projector only without the MRA module in stage 1 and fully optimizes the entire MLLM during stage 2. Besides the 158K instruction-following data from LLaVA, our 40k LLaVAR-2-Cap data and 182k LLaVAR-2-VQA data are utilized in stage 1 and stage 2 training and will benefit the image text understanding capacity of LLaVAR-2-3.8B.

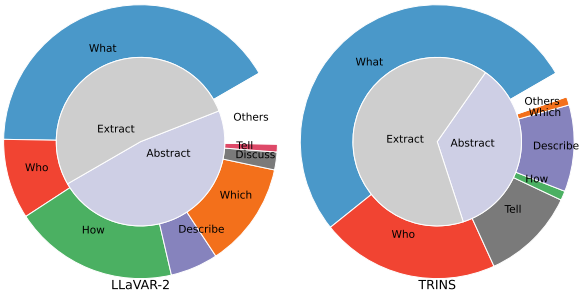


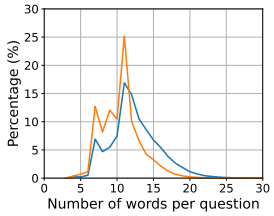
Figure 6: Instruction type statistics based on questioning words and keywords.

5 Dataset Analysis

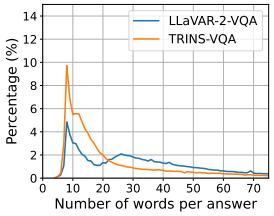
Besides scaling dataset size, improving data diversity is essential for enhancing end-to-end visual instruction tuning in MLLMs. The variety of instructions across concepts and tasks determines a model's ability to understand and navigate text-rich images with complex semantics. In this section, we evaluate the instruction diversity of LLaVAR-2-VQA to assess its quality.

Statistics and Analysis In Figure 6, we show the visualization of instruction distribution of LLaVAR-2-VQA and TRINS-VQA (Zhang et al., 2024), a high-quality text visual instruction-following dataset, based on questioning words and keywords statistics following (Wang et al., 2022). The inner circle illustrates the Abstract and Extract instruction's distribution decided by keywords in questions. The outer circle reflects the distribution of questioning words in **LLaVAR-2-VQA**. It shows overall more diverse patterns of **LLaVAR-2-VQA** than **TRINS-VQA** and the emergence of self-explain data prominently enriches the diversity, e.g. self-explain questions beginning with "Which" and "How" ask for the exact image sources for the extractive answer Q_e. In Figure 7, LLaVAR-2-VQA (averages 12.4 words per question, 38.9 words per answer) shows a more balanced question and answer length distribution than TRINS-VQA (averages 10.6 words per question, 24.3 words per answer), indicating a greater variety of complexity levels in its QAs. Additional statistics are shown in Appendix A.

Diversity Evaluation via Embedding Distance Measurement The statistics and visualization in



(a) Question statistics



(b) Answer statistics

Figure 7: Statistics for LLaVAR-2-VQA: Questions and Answers' lengths.

Embedding Cat.	Task2Vec↑	S-BERT↑
Ours	0.1444	0.6334
TRINS	0.1156	0.5410

Table 1: Instruction Diversity Coefficient of **LLaVAR-2-VQA** and TRINS-VQA based on Task2Vec and Sentence-BERT

Section 5 provide clear sights of how the **LLaVAR-2-VQA** is diversely composed. However, this evaluation based on single-word extraction using manual or neural parsers is not representative enough to reflect the precise instruction coverage. To study a more accurate diversity evaluation of **LLaVAR-2-VQA**, we use the intra-dataset diversity coefficient (Lee et al., 2023) to compute the dataset diversity indicator based on the task-specific TASK2VEC (Achille et al., 2019) embedding and the general sentence-level Sentence-BERT (Reimers and Gurevych, 2019) embedding. Specifically, we formed the intra-dataset instruction diversity coefficient following (Lee et al., 2023) as:

$$\mathrm{div}(D) = \mathbb{E}_{B_1,B_2\sim D} d\left(f_{B_1}, f_{B_2}\right), \qquad (4)$$

where B_1 and B_2 are batches sampled from the target dataset D and f_{B_i} denotes the batch-level embedding which is the diagonal of Fisher Information Matrix for TASK2VEC or the mean of the instructions' Sentence-BERT embeddings of the batch. We apply Cosine distance d to measure the semantic gap between sampled batches within the target dataset. Thus, larger $\mathrm{div}(D)$ indicates higher instruction diversity. While TASK2VEC provides fine-grained task-level diversity measurement, Sentence-BERT indicates general semantic diversity.

We compare the instruction diversity coefficient **LLaVAR-2-VQA** with **TRINS-VQA** in Table 1, with batch size $\|B\| = 512$ and 200 batches for each dataset. As shown, **LLaVAR-2-VQA** exhibits a higher diversity level under both embedding settings, indicating the enriched instruction category and semantics discussed in Section 5.

6 Experimental Results

We aim to illustrate the benefits of integrating LLaVAR-2 in visual instruction tuning. LLaVAR-2 models are evaluated on LLaVAR-2-Cap, LLaVAR-2-VQA and classical text-rich benchmarks to demonstrate the effect of LLaVAR-2 to MLLMs. All experiments are implemented with PyTorch and performed on Nvidia A100 GPUs.

6.1 Visual Question Answering

We first evaluate LLaVAR-2 models and baselines LLaVAR-2 in LLaVAR-2-VQA shown in Table 2. We report Extract Accuracy following (Wu et al., 2023), text similarity metrics BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), and CIDEr (Vedantam et al., 2015). LLaVAR-2 with Phi-3-Mini as the backbone outperforms other methods in zero-shot VQA on LLaVAR-2-VQA. Thanks to the enriched details and the implicit answer process in LLaVAR-2, LLaVAR-2 models can have a better comprehension of visual texts and generate precise extractive answers reflecting from the metric of Extract Accuracy. While sentence similarity metric's results can reflect how well the model can deal with complex extract questions and to what extent the model understands the reasoning process behind the extractive result regarding the self-explain QA set, LLaVAR-2-3.8B's better results on these metrics demonstrate the benefits of LLaVAR-2 on these two perspectives and also indicates the interplay of extractive and self-explain pairs in LLaVAR-2-VQA. Comparing the performance of LLaVAR-2 and LLaVAR-2 w/o D^r, the supplement of the self-explain pair D^r after D^e can enhance the model's capacity toward text-rich images. As for other methods, MiniCPM-V (Yao et al., 2024), Cambrian-1 (Tong et al., 2024a) , and Phi-3/3.5-vision (Abdin et al., 2024) perform well on LLaVAR-2-VQA, highlighting the importance of including diverse visual text data during training.

In addition, we also conduct evaluations on the classical visual text understanding benchmarks (Liu et al., 2023b) and OCR-Bench (Liu et al., 2023c) shown in Table 3 and Table 4. For benchmarks involving images that include both textual and abstract visual elements, such as DocVQA (Mathew et al., 2021), ChartQA (Masry et al., 2022), and InfoVQA (Mathew et al., 2022), LLaVAR-2 models significantly outperform other

Method	Extract Acc.	B@1	B@2	B@3	B@4	METEOR	ROUGE	CIDEr
Phi-3-vision-128k (Abdin et al., 2024)	22.3	47.2	35.8	29.1	24.4	51.5	63.5	293.8
Phi-3.5-vision (Abdin et al., 2024)	18.7	42.6	32.1	26.0	21.7	47.7	61.1	280.8
LLaVA-1.5-7B (Liu et al., 2023b)	7.8	34.5	23.9	17.9	14.1	44.0	46.1	123.0
LLaVA-NeXT-7B (Liu et al., 2024a)	9.3	36.9	26.5	20.6	16.7	49.4	50.4	142.2
Idefics3-8B (Laurençon et al., 2024a)	32.1	28.2	22.4	18.9	16.4	44.5	55.1	222.4
InternVL2-8B (Chen et al., 2024b)	25.3	31.1	22.3	17.2	13.9	55.1	52.1	174.6
MiniCPM-V-2.6-8B (Yao et al., 2024)	42.7	40.4	30.3	24.5	20.6	61.0	60.4	259.6
Cambrian-1-8B (Tong et al., 2024a)	43.8	40.4	30.5	24.7	20.8	56.4	57.7	222.3
LLaVAR-2 (Llama-3.1 8B) ‡	59.0	53.6	43.1	36.6	31.9	64.5	68.2	355.5
LLaVAR-2 (Vicuna-1.5 13B)‡	**62.1**	55.2	**45.0**	**38.4**	**33.7**	65.2	**69.6**	**371.0**
LLaVAR-2 (Phi-3-Mini) w/o D^r ‡	53.2	52.5	42.1	35.5	30.8	64.8	67.9	349.1
LLaVAR-2 (Phi-3-Mini) ‡	59.6	**55.4**	44.8	38.1	33.3	**65.4**	69.1	362.5

Table 2: Results of LLaVAR-2 models trained w/wo self-explain data D^r for text-rich image question-answering tasks. We use ‡ to refer to models fine-tuned on the proposed LLaVAR-2 dataset. We use **bold** and underline to indicate the best and second best results, respectively. Row in the gray is the ablation result without D^r.

	ST-VQA	TextVQA	DocVQA	ChartQA	InfoVQA	FUNSD	SROIE
BLIP-2 (Li et al., 2023b) †	21.7	32.2	4.9	3.4	11.3	0.20	0.14
OpenFlamingo (Awadalla et al., 2023) †	19.3	29.1	5.1	9.1	15.0	0.85	0.12
MiniGPT4 (Zhu et al., 2023) †	14.0	18.7	3.0	4.3	13.3	1.19	0.04
mPLUG-Owl (Ye et al., 2023c) †	29.3	40.3	6.9	9.5	16.5	1.02	0.60
LLaVA (Liu et al., 2023b) †	28.9	36.7	6.9	28.9	13.8	1.02	0.12
LLaVA1.5 (Liu et al., 2023b) †	38.1	38.7	8.5	9.3	14.7	0.20	1.70
LLaVAR (Zhang et al., 2023c)†	39.2	48.5	11.6	12.2	16.5	0.50	5.20
mPLUG-Owl2 (Ye et al., 2023b) †	29.3	40.3	6.9	19.4	18.9	1.40	3.20
Monkey (Li et al., 2024b)†	54.7	64.4	50.1	54.0	25.8	24.1	41.9
TextMonkey (Liu et al., 2024b) †	**61.8**	**71.3**	64.3	58.2	28.2	32.3	47.0
LaRA-13B (Zhang et al., 2024) †	47.2	59.9	50.8	25.6	28.4	23.2	36.6
LLaVAR-2 (Llama-3.1 8B)	51.6	61.0	69.3	69.9	32.0	32.3	58.4
LLaVAR-2 (Phi-3-Mini)	52.3	60.2	66.1	**78.5**	30.3	36.0	51.9
LLaVAR-2 (Vicuna-1.5 13B)	59.5	66.0	**71.3**	76.3	**40.2**	**36.7**	**61.9**

Table 3: Zero-shot performance (accuracy %) on text-based VQA. We use † to refer to the results obtained from previous work (Liu et al., 2023c), and use **bold** and underline to indicate the best and second best results, respectively.

Method	Recog.	VQA^S	VQA^D	KIE	Total
Gemini	215	174	128	134	651
GPT-4v	167	163	146	160	636
Text-Monkey	169	**164**	115	116	561
Monkey	174	161	91	88	514
mPLUG-Owl2	153	153	41	19	366
LLaVAR	186	122	25	13	346
LLaVA1.5-13B	176	129	19	7	331
MiniGPT-V2	124	29	4	0	157
LaRA-13B	206	151	101	145	603
LLaVAR-2 (Phi-3-Mini)	225	152	114	143	634
LLaVAR-2 (Vicuna-1.5)	**241**	162	**121**	**156**	**680**

Table 4: Results of MLLMs on OCRBench. Recog. represents text recognition, VQA^S represents Scene Text-Centric VQA, VQA^D represents Document-Oriented VQA. We use **bold** and underline to indicate the best and second best results, respectively.

methods, highlighting the importance of high proportion of abstract QAs within LLaVAR-2-VQA to improve abstract document understanding. Additionally, while these three datasets are involved in the training of Monkey (Li et al., 2024b) and TextMonkey (Liu et al., 2024b), LLaVAR-2 performs better than these finetuned models on these three datasets, thanks to high-quality text-centric VQAs in LLaVAR-2-VQA/Cap. For scene text-centric ST-VQA (Biten et al., 2019) and TextVQA (Singh et al., 2019) which are dominated by natural scene/object images with text, LLaVAR-2 models have the comparable performance with Monkey and TextMonkey, which includes TextVQA during their training. Surprisingly, we find LLaVAR-2 model is good at reading on scanned pure text images even noisy ones, such as those in FUNSD (Jaume et al., 2019) and SROIE (Huang et al., 2019). Combined with the result of OCRBench, LLaVAR-2-VQA shows comprehensive improvement in fine-tuning for Visual Question Answering of text-rich images. Furthermore, Monkey and TextMonkey show comparable results on benchmarks such as ST-VQA and TextVQA, due to the rich text labels involved in their multi-task training. LLaVAR-2 shares similar ideas: the enriched detail in LLaVAR-2-Cap and the self-explain answers in VQA supplement considerable text labels.

Method	B@1	B@2	B@3	B@4	METEOR	ROUGE	CIDEr
Phi-3-vision-128k	43.6	27.7	18.5	13.0	34.3	51.6	35.9
Phi-3.5-vision	40.5	25.5	16.8	11.7	32.2	49.8	30.9
LLaVA-1.5-7B	28.3	14.7	7.9	4.9	21.8	36.6	10.2
LLaVA-NeXT-7B	30.8	19.8	13.6	10.0	28.1	45.0	27.7
Idefics3-8B	21.8	15.4	11.0	8.2	36.2	31.6	1.4
InternVL2-8B	21.9	14.8	10.2	7.5	37.7	35.9	4.9
MiniCPM-V-2.6-8B	41.0	27.6	20.1	15.4	34.8	48.9	28.9
Cambrian-1-8B	36.6	23.2	15.6	11.1	31.0	47.8	28.8
LLaVAR-2 (Llama-3.1 8B) ‡	58.7	43.6	33.8	27.1	47.1	60.5	61.4
LLaVAR-2 (Vicuna-1.5 13B)‡	**58.9**	**44.4**	**34.9**	**28.3**	**48.0**	**61.7**	**66.9**
LLaVAR-2 (Phi-3-Mini) w/o D^g ‡	31.6	20.1	14.2	10.4	32.1	52.2	39.7
LLaVAR-2 (Phi-3-Mini) ‡	58.1	43.4	33.8	27.2	47.2	60.9	60.8

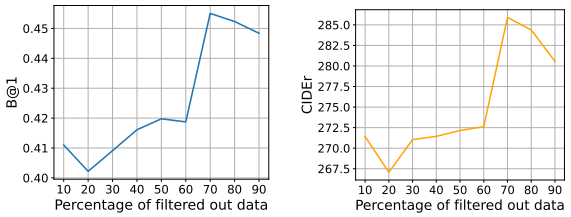
Table 5: Results of LLaVAR-2 models trained w/wo global instruction-following data D^g based on detailed captions for text-rich image captioning tasks. We use ‡ to refer to models fine-tuned on LLaVAR-2 dataset, use **bold** and underline to indicate the best and second best results, respectively. Row in the gray is the ablation result without D^g.

6.2 Text-rich Image Captioning

We evaluate the captioning capacity of LLaVAR-2 models and baselines on LLaVAR-2-Cap, which requires MLLMs to generate summaries and keep necessary text labels at the same time. We compare the performance on this challenging task in Table 5. We can observe that MiniCPM-V, Cambrian-1, and Phi-3/3.5-vision achieved remarkable improvement upon classic MLLMs such as LLaVA-1.5. It is due to the considerable attention of these models on text-heavy visual tasks, for example, MiniCPM-V (Yao et al., 2024) includes a large amount of text-rich captioning data in its stage-1 and 2 training, Phi-3.5-vision's post-training is involved with diverse text image tasks. Fine-tuned on LLaVAR-2-Cap, LLaVAR-2 models outperform other methods on all the text similarity metrics, indicating the necessity of the comprehensive captions in LLaVAR-2-Cap. Besides, the Mixture-of Resolution Adaptation used in the visual encoder of LLaVAR-2 models adopted from LLaVA-HR plays a crucial role in making accurate summaries in captioning tasks. Among LLaVAR-2 models with different backbones, LLaVAR-2 (Vicuna-1.5 13B) achieved the best result, while smaller-sized backbones such as Phi-3-Mini and Llama-3.1 8B still obtained comparable performance. LLaVAR-2 (Phi-3-Mini) w/o D^g is only fine-tuned on the VQA data without adaptation on captioning data. Its not ideal result on captioning tasks further demonstrates the essential of global descriptive data for summarizing tasks.

6.3 Additional Experiments on Data Filtering

In this section, we aim to verify the effectiveness of the proposed mIFD and FFD scores. We apply the mIFD score to select the extractive data D^e first and then filter D^r via FFD score. Filtering



(a) BLEU-1 (b) CIDEr

Figure 8: LLaVAR-2's performance on LLaVAR-2-VQA with different percentages of filtered-out data.

out data ranging from 10% to 90%, We apply the different filtered LLaVAR-2-VQA train set as the only fine-tuning dataset for different checkpoints to verify the effectiveness of the proposed filtering method. In Figure 8, we present the BLEU-1 and CIDEr results for different filtering percentages. It is evident that 70% is the sweet spot for LLaVAR-2-VQA that our filtering scores are efficient before 70% and after 70% performance drops due to the limited size of data for fine-tuning. These results demonstrate that the proposed filtering score can select high-quality samples enabling the model to reach better performance with a smaller data size.

7 Conclusions

We propose Hybrid-Instruct, an automatic multi-modal instruction generation framework based on manual captions for visual instruction tuning tailored to text-rich images, to construct LLaVAR-2 data. The detail-enriched captions and self-explain dialogues in LLaVAR-2 enhance the performance of MLLMs on different benchmarks. In addition, the proposed filtering score mIFD and FFD are shown to be effective in filtering out unqualified dialogues in the LLaVAR-2 given the image and extraction contexts. In general, LLaVAR-2 introduces novel approaches to generate and select diverse and high-quality instruction-following data for MLLMs.

8 Limitations

While the LLaVAR-2 dataset offers significant improvements in MLLMs' tuning, It still leaves us limitations to improve: (1)the quality of OCR results we relied on to augment manual captions and to create self-explain dialogues may include errors. (2) LLaVAR-2's data generation is dependent on strong MLLMs such as GPT-4o, which possibly introduces biases and is expensive to use. (3) Our proposed filtering process only occurs during post-filtering. The filtered-out samples are wasted and should still be leveraged.

Thus, for future work, we will focus on designing self-correction data collection pipelines. The potential solutions could be adding an extra checking module to ensure the quality of inputs and utilizing the filtered data to fine-tune a small rating model for better data selection.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona. 2019. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6430–6439.

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

Dosovitskiy Alexey. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024a. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*.

Lin Chen et al. 2023a. Sharegpt4v: Improving large multi-modal models with better captions. *Preprint*, arXiv:2311.12793.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension

in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

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

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

Alycia Lee, Brando Miranda, Sudharsan Sundar, and Sanmi Koyejo. 2023. Beyond scale: the diversity coefficient as a data quality metric demonstrates llms are pre-trained on formally diverse data. *arXiv preprint arXiv:2306.13840*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Chunyuan Li et al. 2023a. Multimodal foundation models: From specialists to general-purpose assistants. *Preprint*, arXiv:2309.10020.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2023c. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *arXiv preprint arXiv:2308.12032*.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024b. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26763–26773.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pretraining for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. 2023c. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*.

Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. 2024b. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*.

Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. 2024. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *arXiv preprint arXiv:2403.03003*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

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

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, et al. 2024. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv preprint arXiv:2408.15998*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. 2021. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8802–8812.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. 2024a. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024b. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Zizhang Wu, Xinyuan Chen, Jizheng Wang, Xiaoquan Wang, Yuanzhu Gan, Muqing Fang, and Tianhao Xu. 2023. Ocr-rtps: an ocr-based real-time positioning system for the valet parking. *Applied Intelligence*, 53(14):17920–17934.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, Maosong Sun, and Gao Huang. 2024. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. *arXiv preprint arXiv:2403.11703*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. 2023a. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023b. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Qinghao Ye et al. 2023c. mplug-owl: Modularization empowers large language models with multimodality. *Preprint*, arXiv:2304.14178.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2023a. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

Ruiyi Zhang, Yanzhe Zhang, Jian Chen, Yufan Zhou, Jiuxiang Gu, Changyou Chen, and Tong Sun. 2024. Trins: Towards multimodal language models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22584–22594.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023b. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023c. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

Henry Hengyuan Zhao, Pan Zhou, and Mike Zheng Shou. 2023. Genixer: Empowering multimodal large language models as a powerful data generator. *arXiv preprint arXiv:2312.06731*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A LLaVAR-2 Statistics

Number of images	42870
Number of detail-enriched captions	42870
Number of LLaVAR-2-Cap pairs	42870
Average # of words per detail-enriched captions	114.5
Number of LLaVAR-2-VQA pairs	382406
Average # of words per VQA question	12.4
Average # of words per VQA answer	38.9

Table 6: LLaVAR-2 Dataset Statistics

B Details for LLaVAR-2 Data Collection

System message for the generation of detail-enriched captions A^g in LLaVAR-2-Cap:

You are an AI visual assistant, and you are seeing a single image. What you see is provided with an image, one corresponding OCR result, and one corresponding image caption describing the information within the same image you are looking at. Image captions and OCR results might include hallucinations, while OCR results are more accurate. OCR results are constructed with [[4 bounding-box coordinates], text, OCR confidence]. Enrich the image caption with detailed support text and location information from the OCR result.
The enriched image caption should be in a tone that a visual AI assistant is seeing the image and describing the image. Maintain the content in the original image caption while adding the support information from the OCR result to its corresponding part in the original image caption for a detailed description:
(1) for the support text added to the image caption, its bounding box coordinates should indicate the similar location described in the corresponding part of the original image caption;
(2) DO NOT mention OCR bounding box coordinates in your caption. Provide the location information with the same style as the original image caption;
(3) DO NOT add information that looks unrelated to or contradicts OCR results;
(4) When OCR results include new information, enrich this content into the original caption;
(5) When OCR results and image caption has large difference in describing the same area of the image, maintain the description in the original image caption;
(6) Do not include garbled characters in the generation.

System message for the generation of extractive conversation D^e in LLaVAR-2-VQA following (Liu et al., 2023b):

You are an AI visual assistant, and you are seeing a single image. What you see is provided with two OCR results and one image caption describing the information within the same image you are looking at. Image captions might include hallucinations, while OCR results are more accurate. Answer all questions with definite answers as you are seeing the image.
Design a conversation between you and a person asking about this photo. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question. Ask diverse questions and give corresponding answers.
Include questions asking about the visual content of the image (e.g., the man, the sunset, the ocean.) and the texts contained in the image. Only include questions that have definite answers:
(1) one can see the content in the image that the question asks about and can answer confidently;
(2) one can determine confidently from the image that it is not in the image. Do not ask any questions that cannot be answered confidently;
(3) DO NOT mention OCR or image caption in your questions and answers;
(4) DO NOT ask about information from captions while it looks unrelated to or contradicts OCR results.
Also include complex questions that are relevant to the content in the image, for example, asking about background knowledge of the texts in the image, asking to discuss the design of the image, etc. Again, do not ask about uncertain details. Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. You can include multiple paragraphs if necessary.

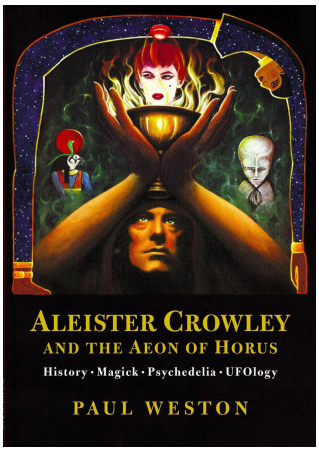
System message for the generation of self-explain conversation D^r in LLaVAR-2-VQA:

You are an AI visual assistant, and you are seeing a single image. What you see is provided with an image, one corresponding OCR result, one corresponding image caption describing the information within the same image you are looking at, and a set of reference QAs on this image. OCR results

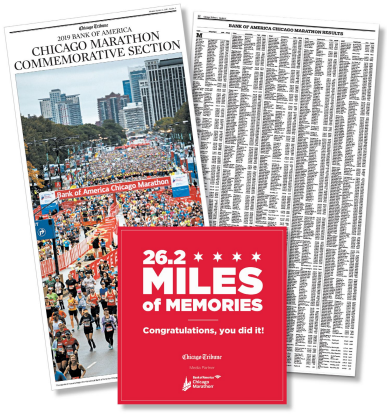
contain text description with location information which is constructed with [[4 bounding-box coordinates], text, OCR confidence]. The image caption contains a visual description from the human. A set of reference questions and answers around this image are provided after the image, OCR result, and image caption. Based on them, generate a pair of reasoning QA that asks why/how/where each provided reference answer looks like it and explains it in the generated reasoning answer
The generated reasoning QA should be in a tone that a visual AI assistant is seeing the image and answering the question:
(1) The generated question can ask the reason behind the provided reference QA's answer or ask how to get the provided QA's answer according to the provided context;
(2) If the reasoning process for the reference QA is very explicit, the generated reasoning QA should focus on the source of the reference answer, e.g. for simple extractive question answering, the reasoning QA should focus on where the extraction of the target object of the reference question from in the image and its content in the image;
(3) If the reasoning process for the reference QA is complex, the generated reasoning QA should reason on the relationship/connection between the involved objects (text object, visual object if needed) in the image;
(4) The generated questions should have diverse styles, do not only use a single style format;
(5) Find supportive information from the provided image/caption/OCR results to answer the generated reasoning question, but only assuming you can see the image, do not mention that you can see the caption/OCR results;
(6) Make sure the generated reasoning QA corresponds/very related to the reference QA and the image, avoid only extending new concepts in the reasoning QA;
(7) Generated reasoning QA should be in the same format as the provided QA:'Reasoning Question: {generated reasoning question}\\n Reasoning Answer: {generated reasoning answer}'.

C LLaVAR-2-Cap Examples

Figure 9 and Figure 10 present examples in LLaVAR-2-Cap.



This is the title page of the book titled \"ALEISTER CROWLEY AND THE AEON OF HORUS.\" The book discusses themes such as History, Magick, Psychedelia, and UFOlogy. The author's name, Paul Weston, is prominently displayed at the bottom center of the page. The cover art features an illustration of hands holding a flaming chalice, with mystical figures and symbols surrounding it. On the left side, there's a figure with a bird's head and a red disk, while on the right side, there's another mystical figure. The background is adorned with cosmic and spiritual imagery, enhancing the esoteric theme of the book.

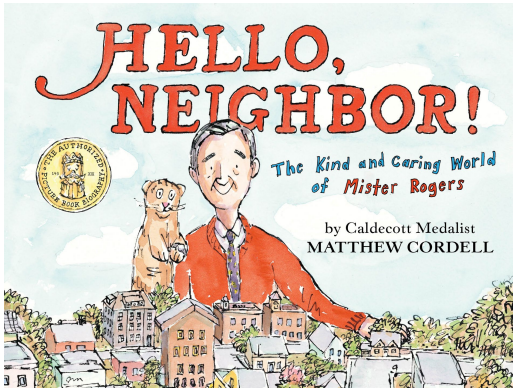


This is a promotional ad celebrating the 2019 Bank of America Chicago Marathon. In the lower center, the bold title \"26.2 Miles of Memories\" is prominently displayed, with the encouraging subtext \"Congratulations, you did it!\" just below. Below this, the logos for the Chicago Tribune and the Chicago Marathon are also visible.\n\nThe ad background mimics a newspaper layout split into two pages. The left page features the headline \"Chicago Tribune 2019 Bank of America Chicago Marathon Commemorative Section\" at the top. Below this, there is a vibrant photo capturing a large crowd of runners beneath a banner that reads \"Bank of America Chicago Marathon\" set against an urban backdrop with tall buildings.\n\nThe right page is titled \"Bank of America Chicago Marathon Results\" and is densely filled with the detailed race results, listing numerous participants' names and times. The overall background of the ad is plain white, enhancing the prominence of the colorful and detailed newspaper sections

Figure 9: Caption examples in LLaVAR-2-Cap

This is a movie poster for \"Gentleman's Agreement.\" The poster prominently features the names of the leading actors: Gregory Peck, Dorothy McGuire, and John Garfield, all in large blue text on the right side. The phrase \"DARRYL F. ZANUCK presents\" is displayed at the top. Below the cast names, it is noted that the movie is based on Laura Z. Hobson's work.\n\nThe title \"Gentleman's Agreement\" is displayed prominently in large yellow and red text towards the bottom center of the poster. The poster also includes additional cast names at the bottom: Celeste Holm, Anne Revere, June Havoc, Albert Dekker, Jane Wyatt, and Dean Stockwell. Production credits include Darryl F. Zanuck as the producer, Moss Hart for the screenplay, and Elia Kazan as the director.\n\nIn the background, there is an illustration of a man sitting and appearing pensive, while a woman stands beside him. Additionally, the poster features a \"20th CENTURY-FOX\" logo at the bottom right corner, highlighting the studio behind the film.

This is a blog image titled \"Capricorn Rituals.\" The title is prominently displayed in large white and green text at the upper center of the image. Below the title, various rituals are listed, each accompanied by an illustration. The subtext reads:\n\n- \"meditate,\" illustrated with a gesturing hand on the left side of the image.\n- \"write down long-term intentions,\" depicted with a list of intentions in the center-right part of the image.\n- \"eat plant-based for a day, buy a plant, or plant a sapling for your home!\" shown with hands cupping a plant in the center-right section.\n- \"go for a silent walk or hike listening to the sounds of nature,\" illustrated with a girl in an orange jacket walking next to an orange leaf tree at the bottom left.\n- \"get rid of one distraction for a week,\" depicted with a cell phone and likes and heart emojis coming out of it in the bottom right.\n\nEach ritual is visually represented, creating a cohesive and engaging guide for Capricorn rituals.



This is a book cover featuring the title \"HELLO, NEIGHBOR!\" in large red text in the upper center. Below the title, in smaller blue and red text, it reads \"The Kind and Caring World of Mister Rogers.\" On the lower right side, black text states, \"by Caldecott Medalist MATTHEW CORDELL.\" An illustrated gold seal on the left side reads \"The Authorized Picture Book Biography.\" The background illustration showcases an elderly man with gray hair and a red sweater standing behind a miniature town with brown buildings and green trees, accompanied by a cat.

This is a sound circus cover post. The title reads \"JOANNA MACGREGOR GERSHWIN: Rhapsody in Blue, Piano Concerto in F, The Gershwin Songbook, Broadway Arrangements.\" The background cover image features a lady with blonde hair wearing a black and gray dress, standing in front of old brick buildings.

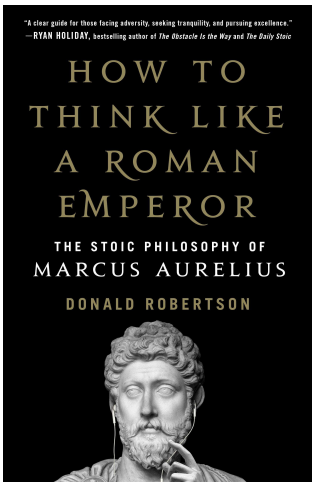
Figure 10: Caption examples in LLaVAR-2-Cap

Extractive Question: What is the price for a VIP meet and greet with both Lifehouse and Switchfoot?
Extractive Answer: The price for a bundle VIP meet and greet with both Lifehouse and Switchfoot is $135.

Self-explain Question: What details are provided in the ad to determine the price of a VIP meet and greet bundle with both Lifehouse and Switchfoot?
Self-explain Answer: The price for a VIP meet and greet bundle with both Lifehouse and Switchfoot is listed under \"OPTION 3\" in the VIP Experience Options section. It states \"BUNDLE: LIFEHOUSE & SWITCHFOOT VIP MEET & GREET - $135,\" providing the specific cost.

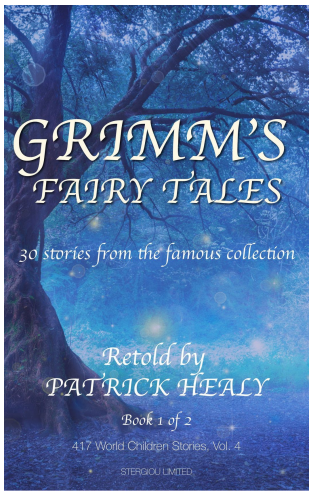
Extractive Question: Who awarded Nilanjan Saha the title of \"The World's Best Brother\"?
Extractive Answer: The title of \"The World's Best Brother\" was awarded to Nilanjan Saha by Sayantani.

Self-explain Question: What indicates that Sayantani awarded the title to Nilanjan Saha?
Self-explain Answer: The award plaque explicitly states \"Awarded by: Sayantani\" at the bottom right corner, confirming that Sayantani is the one who awarded the title to Nilanjan Saha.



Extractive Question: What is the main subject of the book \"How to Think Like a Roman Emperor\"?
Extractive Answer: The main subject of the book \"How to Think Like a Roman Emperor\" is the Stoic philosophy of Marcus Aurelius.

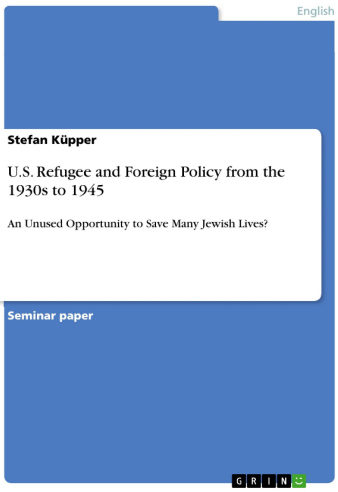
Self-explain Question: What elements on the book cover indicate the main subject of \"How to Think Like a Roman Emperor\"?
Self-explain Answer: The main subject is indicated by the subtitle \"The Stoic Philosophy of Marcus Aurelius\" which clearly states the focus on Stoic philosophy through the figure of Marcus Aurelius.



Extractive Question: How many volumes are there in this series, and which volume is this book?
Extractive Answer: This book is the first of two volumes in the series.

Self-explain Question: What information on the cover indicates the number of volumes in the series and the specific volume of this book?
Self-explain Answer: The information \"Book 1 of 2\" is located towards the bottom center of the cover, indicating that this is the first of two volumes in the series.

Figure 11: QA examples in LLaVAR-2-VQA



Extractive Question: What organization or publisher logo is visible on the book cover?
Extractive Answer: The logo visible on the book cover is \"GRIN.\"

Self-explain Question: Where is the logo of the organization or publisher located on the book cover?
Self-explain Answer: The logo \"GRIN\" is located at the bottom right corner of the book cover, as indicated by the small green and white logo.

Extractive Question: What is the main theme suggested by the short context on the cover?
Extractive Answer: The main theme suggested by the short context on the cover is about a character attempting to break through the emotional barriers of someone who is perceived as cold-hearted. It hints at themes of love, perseverance, and possibly redemption, as one character strives to reach the heart of another.

Self-explain Question: What elements of the short context hint at the main theme of the book?
Self-explain Answer: The elements \"cold-hearted bastards\" and \"working her way into his heart\" suggest themes of love, perseverance, and emotional breakthrough, indicating a narrative focused on overcoming emotional barriers and possibly redemption.

Figure 12: QA examples in LLaVAR-2-VQA

FFD score = 0.064267

Extractive

Q: What is the content of the Amharic text on the book cover?
A: The Amharic text on the book cover translates to English as follows: \"Kill the righteous religion. Abune Tekle was a religious man and a farmer. When he saw them, they cut down their pine trees, put them in the blood and hid. When Balege was traveling from the countryside to the city, he was approached by Abune Takle Haimanot from the road. Because he knows the nature of the man, he is the master of what he teaches. How are you, my brother, where is the prayer of Hajj? I am going to the city forever. Abune Tekleahimano and also a Hajj to Zyaw. When it's time to sleep, bro. We sat at the edge of this place, tasted grain and drank water.\"

Self-explain

Q: Where is the Amharic text located on the book cover, and what does it translate to in English?
A: The Amharic text is located in the middle section of the book cover. It translates to English as follows: \"Kill the righteous religion. Abune Tekle was a religious man and a farmer. When he saw them, they cut down their pine trees, put them in the blood and hid. When Balege was traveling from the countryside to the city, he was approached by Abune Takle Haimanot from the road. Because he knows the nature of the man, he is the master of what he teaches. How are you, my brother, where is the prayer of Hajj? I am going to the city forever. Abune Tekleahimano and also a Hajj to Zyaw. When it's time to sleep, bro. We sat at the edge of this place, tasted grain and drank water.\"" **[repetitive text]**

FFD score = 1.070361

Extractive

Q: Can you describe the design of the book cover?
A: Yes, the book cover has the title \"Cars and How to Drive Them\u2014Primary Source Edition\" written in a large, gold font at the center. The background is dark blue with a textured pattern. Below the title, the word \"Anonymous\" is written in a smaller gold font.

Self-explain

Q: How does the title \"Cars and How to Drive Them: Primary Source Edition\" indicate the nature of the book's content?
A: The title \"Cars and How to Drive Them: Primary Source Edition\" indicates that the book contains information on driving cars, and the label \"Primary Source Edition\" suggests that this information might be original and unaltered, providing firsthand accounts or instructions from the time it was first published. This implies the content could be historically significant and authentic. **[irrelevant Self-explain]**

Figure 13: Filtered out Self-explain Examples based on FFD score

D LLaVAR-2-VQA Examples

Figure 11 and Figure 12 present examples in LLaVAR-2-VQA.

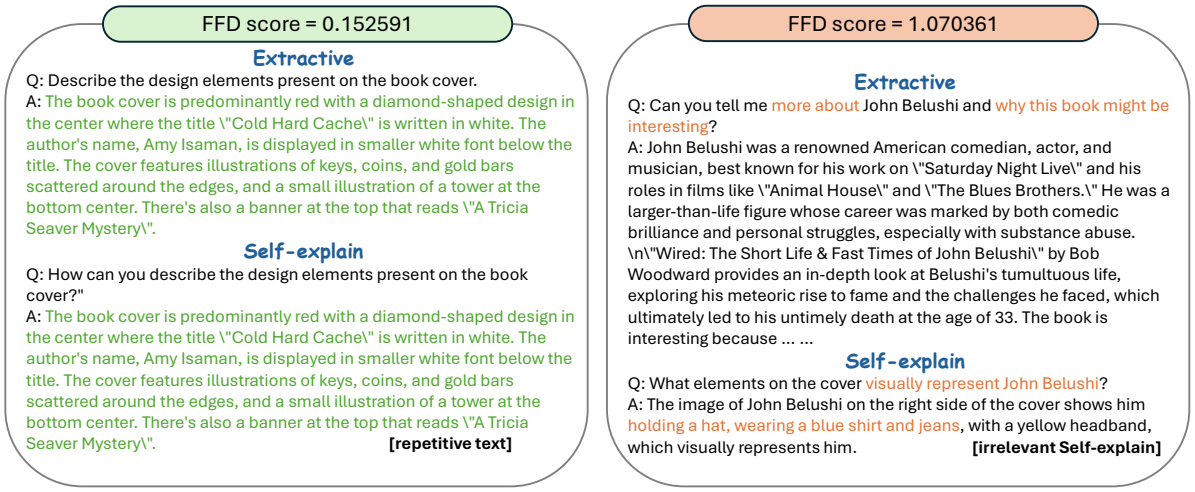
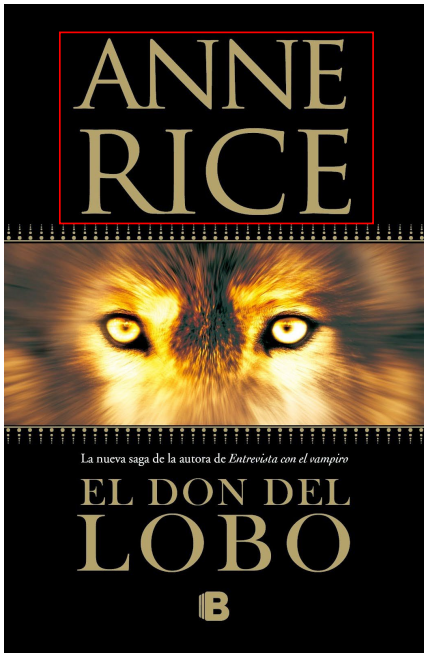


Figure 14: Self-explain Examples filtered out based on FFD score

E Filtered out Examples

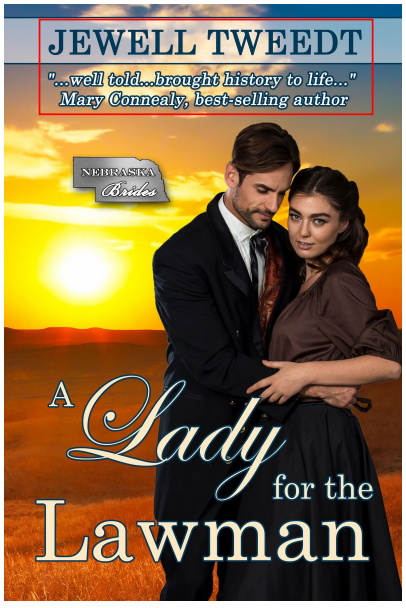
Extractive data filtered out by mIFD score Figure 15 presents extractive examples filtered out using mIFD score. The analysis is included in each example.

VQA examples filtered out by FFD score Figure 13 and Figure 14 present examples filtered out using FFD score including 2 cases: FFD score is closed to 0 and FFD score is closed to 1.

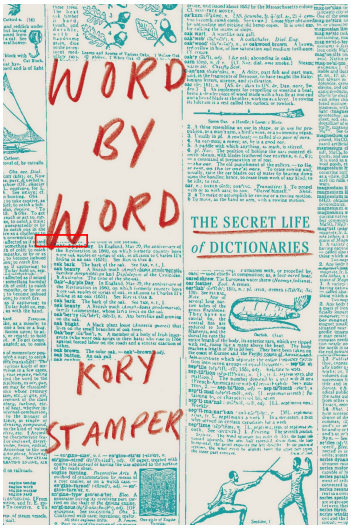


mIFD score: 1.1470636980980542
Question: What is notable about the author Anne Rice?
Answer: Anne Rice is known for her novel \"Interview with the Vampire,\" which is a part of her famous \"The Vampire Chronicles\" series.
[target related but content unrelated with the image, the content is generated by MLLM's pre-knowledge]

mIFD score: 1.1438712435292755
Question: Can you tell me more about the West Memphis Three mentioned on the cover?
Answer: The West Memphis Three refers to {a case involving three young men\u2014Damien Echols, Jason Baldwin, and Jessie Misskelley Jr.\u2014who were convicted as teenagers in 1994 for the murders of three boys in West Memphis, Arkansas. The case garnered national and international attention due to concerns about the fairness of the trial and the evidence used to convict the three men. Over the years, new evidence and re-examinations of the case led to calls for their release, and in 2011, the three men entered Alford pleas...}
[target related but content unrelated with the image, the content is generated by MLLM's pre-knowledge]



mIFD score: 1.1389945569978437
Question: What effect do the black and white fonts have on the book cover design?
Answer: The use of black and white fonts on the book cover adds a classic and timeless touch to the design. This choice of font color contrasts well with the sunset backdrop, making the title and author\u2019s name stand out clearly, while also evoking a sense of nostalgia and elegance.
[The front is blue and white, focusing on a very abstract concept and the answer is slightly related with the image]



mIFD score: 1.1151096519137953
Question: Discuss the historical context of the phrase \"oak apple day\" mentioned in the text.
Answer: "\"Oak Apple Day\" refers to an annual celebration in England on May 29th, commemorating the restoration of the monarchy in 1660. It marks the day Charles II was restored to the throne. The term \"oak apple\" comes from the story of Charles II hiding in an oak tree to escape capture by Parliamentary forces after the Battle of Worcester in 1651. The oak tree became a symbol of loyalty to the monarchy, and the day was celebrated with various festivities, including the wearing of oak leaves.
[very small concept in the image, the content is generated by MLLM's pre-knowledge]

Figure 15: Extractive Examples filtered out based on mIFD score