

OURO: A Self-Bootstrapped Framework for Enhancing Multimodal Scene Understanding

Tianrun Xu^{1,2,*}, Guanyu Chen^{1,*}, Ye Li³, Yuxin Xi⁴, Zeyu Mu¹, Ruichen Wang¹, Tianren Zhang¹,
Haichuan Gao^{1,†}, Feng Chen^{1,†}

¹Department of Automation, Tsinghua University, Beijing, China

²Zhongguancun Academy, Beijing, China

³School of Software, Xinjiang University

⁴School of Artificial Intelligence, Beijing Normal University, Beijing, China

{xtr24, chen-gy23}@mails.tsinghua.edu.cn, {ghc2023, chenfeng}@mail.tsinghua.edu.cn

Abstract

Multimodal large models have made significant progress, yet fine-grained understanding of complex scenes remains a challenge. High-quality, large-scale vision-language datasets are essential for addressing this issue. However, existing methods often rely on labor-intensive manual annotations or closed-source models with optimal performance, making large-scale data collection costly. To overcome these limitations, we propose a self-bootstrapped training pipeline that leverages the model’s own multimodal capabilities to recursively refine its understanding. By decomposing existing multimodal data into localized sub-regions and generating hierarchical scene descriptions and multi-faceted question-answer pairs, we construct a dataset based on 1.4M image-task instances. We further utilize this dataset to train the base model, significantly enhancing its ability to interpret complex visual scenes and perform various vision-related tasks. Our OURO model, fine-tuned on Qwen2-VL-7B-Instruct using LoRA, achieves substantial improvements over both the base model and similarly-sized counterparts across multiple multimodal benchmarks. Our self-bootstrapped training pipeline offers a novel paradigm for the continuous improvement of multimodal models. Code and datasets are available at <https://github.com/tinnet123666888/OURO.git>.

1. Introduction

The development of large language models (LLMs) [66] [68] [14], exemplified by GPT [2] and similar architectures, has driven significant advancements

*Equal contribution.

†Corresponding authors.

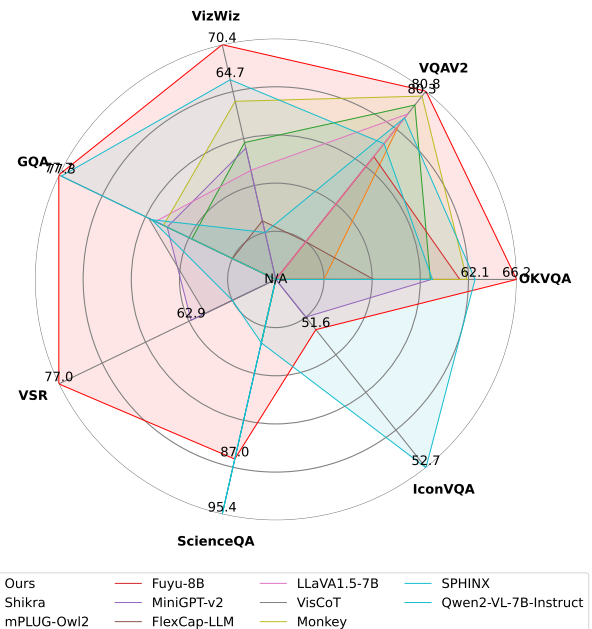


Figure 1. Comparison of our OURO with existing methods across multiple evaluation metrics.

in natural language processing [50], achieving state-of-the-art performance across various linguistic tasks. Building on the success of these models, VLMs [73] [75], such as CLIP [52], Flamingo [3], and BLIP [33], have emerged to address multimodal challenges by jointly modeling textual and visual information. These models benefit from vast amounts of multimodal data, which provides rich visual-contextual information [29, 74], and enables models to learn intricate object relationships, spatial configurations, and context, leading to their ability to reason over complex multimodal tasks.

Despite advancements, current VLMs primarily rely on

whole-image descriptions, overlooking fine-grained object attributes and spatial relationships. Existing annotations often provide broad scene summaries like “many cups and other items on the table,” rather than structured descriptions that differentiate individual objects, their properties, and interactions. This lack of hierarchical organization prevents models from capturing multi-level scene representations, making it difficult to reason about spatial dependencies.

To address these shortcomings, recent efforts have explored various strategies to enhance multimodal understanding, often requiring additional annotation efforts or integrating multiple models to improve performance. Approaches like SPHINX [37] and mPLUG-Owl2 [70] employ joint weight mixing, visual embedding integration, and task-specific tuning to refine representation learning, yet they still struggle to capture fine-grained spatial relationships across objects of different scales. On the other side, methods such as FlexCap [16], Monkey [36], and Visual CoT [57] leverage external models to generate hierarchical descriptions, region-specific annotations, or visual reasoning cues, improving interpretability but relying heavily on large-scale annotated datasets and pre-trained models, making them costly and difficult to scale. Additionally, they often provide annotations at a fixed granularity, either whole-image descriptions or isolated object labels, without a unified framework for modeling hierarchical relationships and spatial dependencies. This inconsistency in abstraction limits structured scene representation and requires extensive manual intervention or pre-processing, reducing feasibility for large-scale data generation. A self-supervised approach is needed to automatically structure fine-grained scene descriptions while ensuring scalability and adaptability.

We introduce *OURO* (derived from *ouroboros*, a serpent eating its own tail), a multimodal model designed to enhance scene understanding through a self-bootstrapped training approach. Unlike conventional VLMs that rely on static datasets with whole-image annotations, *OURO* refines its own training data by decomposing images into structured sub-regions and leveraging hierarchical relationships for a more granular understanding of object attributes and spatial dependencies. As shown in Fig. 1, our model demonstrates outstanding performance across multiple tasks, outperforming 15 other models in a comprehensive evaluation. At the core of *OURO* is an automated dataset construction process that generates structured scene descriptions and diverse VQA question-answer pairs. The model first employs a Region Proposal Network (RPN) [53] to segment images into hierarchical sub-regions. Each sub-region is then processed by a VLM to produce localized descriptions, which are recursively linked back to parent regions based on their relative positions. This ensures that fine-grained details are not only captured at the object level

but also contextualized within their broader scene composition. The model then integrates these structured descriptions into a unified representation, which serves as the foundation for generating diverse VQA question-answer pairs, ensuring comprehensive reasoning over both object-level details and broader scene context.

To maximize the utility of this enriched dataset, *OURO* employs a multi-region joint training strategy that processes both full images and their sub-regions in parallel. Unlike prior approaches that treat sub-regions as independent samples, our method explicitly models hierarchical dependencies, allowing the model to learn fine-grained visual representations while maintaining a coherent global scene understanding. This training paradigm enhances the model’s ability to generate structured image captions, answer complex visual queries, and reason over intricate object relationships with greater accuracy and interoperability.

Our contributions are as follows:

- We propose a novel bootstrapped training pipeline that leverages only a base model along with a simple RPN. Without requiring additional human annotations or closed-source model costs, our method recursively constructs multi-granularity scene annotation data. Through this training process, we significantly enhance the base model’s capabilities in scene understanding and multimodal question answering.
- Building upon publicly available multimodal datasets, we construct an augmented dataset based on 1.4M image-task instances. Each image in this dataset is accompanied by highly detailed descriptions, multi-level sub-image segmentations, and multiple question-answer pairs covering various aspects.
- The model we trained, *OURO*, demonstrates enhanced capabilities in visual understanding and question answering. It achieves significant improvements over the base model across four major multimodal tasks: image captioning, general VQA, scene text-centric VQA, and document-oriented VQA, evaluated on 20 benchmark datasets. Moreover, *OURO* outperforms other open-source models of similar scale on most evaluation metrics. Notably, in document-oriented VQA, our model even surpasses leading closed-source multimodal models.

2. Related Work

RPNs in Multimodal Learning. RPNs [53] are widely used in object detection to generate candidate regions efficiently. They predict object locations directly from feature maps, offering a fast and scalable alternative to traditional region selection methods. Recent multimodal models, including GLIP [35], Grounding DINO [40] [54], and OVDINO [63], incorporate RPNs to automate object region extraction. While segmentation-based methods like SAM [30] focus on precise object boundaries, an RPN retains contex-

tual information around objects, making it better suited for hierarchical scene decomposition. Given its efficiency and ability to preserve spatial relationships, we adopt an RPN in our framework to enhance structured scene understanding and VQA dataset generation.

VLMs for Scene Understanding. VLMs have advanced to handle tasks such as image descriptions, VQA, and multimodal reasoning by integrating visual and linguistic data. Early models like CLIP [52] leveraged contrastive learning for image-text alignment, while subsequent models such as Flamingo [3] and BLIP [33] introduced improved attention mechanisms to enhance cross-modal understanding. More recent architectures, including Qwen [5] and LLaVA [62], integrate vision transformers to refine few-shot learning and reasoning abilities. However, these models struggle with capturing fine-grained spatial relationships and structured scene representations, limiting their ability to understand complex environments.

To address these challenges, several strategies have been explored. Modular multi-modal models such as mPLUG-Owl2 [70] and SPHINX [37] improve vision-language alignment through weight-mixing strategies and task-specific tuning, particularly benefiting high-resolution image interpretation. Meanwhile, Img2Prompt [22] and QAC [43] enhance VQA performance by generating question-guided prompts, yet they rely on predefined reasoning structures and lack adaptability to complex visual dependencies. Another class of approaches, including Flex-Cap [16], FUSECAP [55], CAPSFUSION [72], Visual CoT [57] and Monkey [36], leverages synthetic data or pseudo-annotations to enhance training. However, reliance on external models or additional human annotations can be high-cost, underscoring the need for a scalable approach to enhance the scene understanding ability of VLMs.

3. Methods

This method involves two main stages: data generation and model training. In the first stage, we use an RPN to extract hierarchical sub-regions, and descriptions are generated for each. These sub-region descriptions are sequentially integrated into their parent regions through object name matching, forming a multi-level scene description. Then the generated descriptions, along with predefined questions, are input into the base VLM to produce VQA data. In the second stage, the original image, its sub-regions and the generated question-answer pairs are fed into the VLM for continuous improvement, leveraging both global and local features.

3.1. Recursive Multi-Level Scene Annotation

Previous models, such as Monkey [36], CAPSFUSION [72], and Visual CoT [57], have explored sub-region descriptions to enhance scene understanding and VQA interpretability. Additionally, methods like FUSECAP [55]

Algorithm 1 Recursive Scene Annotation with VLM

Require: Image I , Prompt P , Confidence threshold τ

Ensure: Hierarchical descriptions $d^{(0)}$ and QA pairs QA

```

1: function RECURSIVEANNOTATE( $I$ )
2:    $d^{(0)} \leftarrow \text{RecursiveDescribe}(I, 0)$ 
3:    $QA \leftarrow \text{VLM}(I, P, d^{(0)})$   $\triangleright$  Generate QA pairs
   using descriptions
4:   return  $d^{(0)}, QA$ 
5: end function
6: function RECURSIVEDESCRIBE( $r^{(t)}, t$ )
7:    $d^{(t)} \leftarrow \text{VLM}(r^{(t)})$   $\triangleright$  Generate description
8:    $R^{(t+1)} \leftarrow \text{RPN}(r^{(t)})$   $\triangleright$  Generate sub-regions
9:   if  $R^{(t+1)} \neq \emptyset$  then
10:     $D^{(t+1)} \leftarrow \emptyset$ 
11:    for each  $r_i^{(t+1)} \in R^{(t+1)}$  do
12:       $d_i^{(t+1)} \leftarrow \text{RecursiveDescribe}(r_i^{(t+1)}, t + 1)$ 
13:       $D^{(t+1)} \leftarrow D^{(t+1)} \cup \{d_i^{(t+1)}\}$ 
14:    end for
15:     $d^{(t)} \leftarrow \text{Merge}(d^{(t)}, D^{(t+1)})$ 
16:  end if
17:  return  $d^{(t)}$ 
18: end function

```

and others [56] [31] [18] use synthetic data to improve scene comprehension. However, these approaches struggle to fully capture spatial relationships and ensure interpretability between descriptions and question-answer pairs, and also incur additional annotation costs. To address these gaps, we integrate hierarchical scene understanding and multi-level description generation to enhance spatial awareness and interpretability in VQA. Fig. 2 and Algorithm 1 illustrate our pipeline.

The annotation process follows a recursive strategy, beginning with an RPN or an OCR [27] module that collaboratively extract candidate sub-regions from the original image I . The RPN focuses on visual object proposals, while the OCR processes text-containing regions if necessary. Only sub-regions with confidence scores above a predefined threshold are retained. Each retained region is then recursively processed in the same manner until no further subdivisions can be made.

Formally, the recursive segmentation can be defined as follows:

$$R^{(0)} = \text{RPN}(I), \quad (1)$$

where $R^{(0)}$ denotes the initial set of sub-regions obtained from the original image. At recursion step t , each sub-region $r_i^{(t-1)} \in R^{(t-1)}$ is further subdivided:

$$R^{(t)} = \bigcup_{r_i^{(t-1)} \in R^{(t-1)}} \text{RPN}(r_i^{(t-1)}), \quad (2)$$

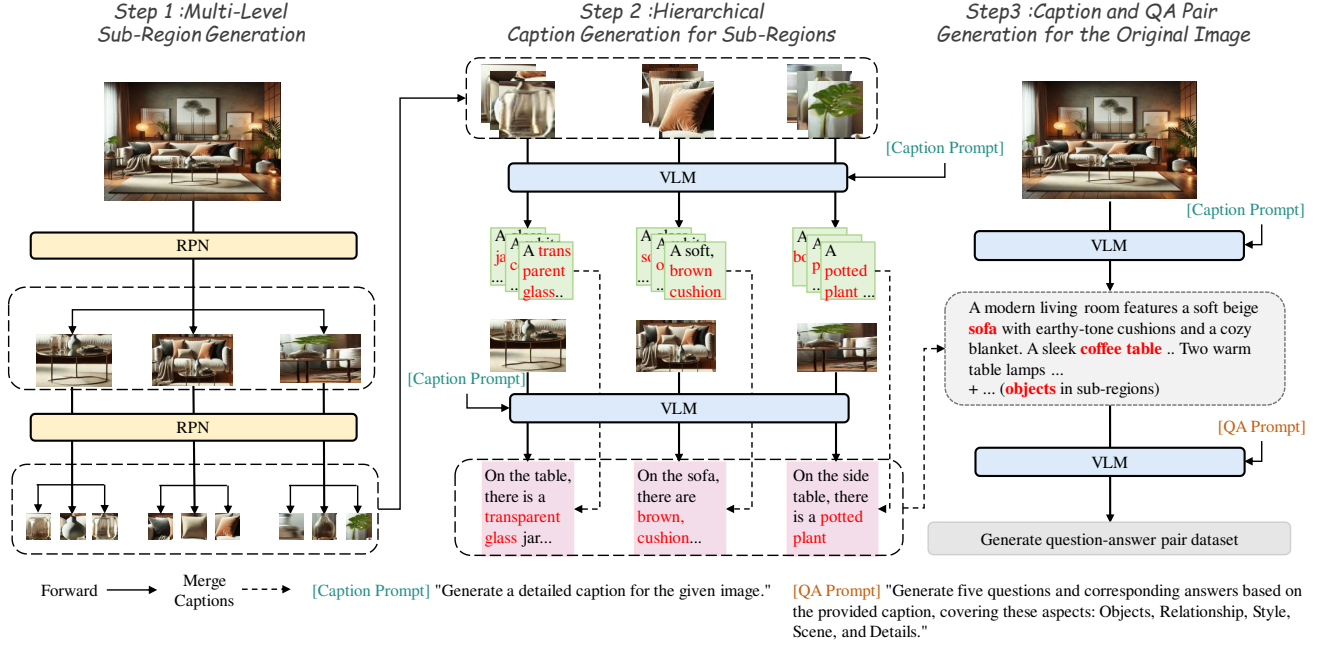


Figure 2. The process of multi-level scene understanding and VQA dataset generation in the OURO framework.

until the termination condition is met, i.e., when no sub-region can be further divided.

Once the final set of sub-regions is obtained, a VLM is used to generate descriptions for each sub-region. The description for a region $r_i^{(t)}$ at level t is given by:

$$d_i^{(t)} = \text{VLM}(r_i^{(t)}). \quad (3)$$

If a sub-region $r_i^{(t)}$ contains further subdivisions $R_i^{(t+1)}$, their descriptions are recursively generated and integrated into the parent region's description via object name matching:

$$d_i^{(t)} = \text{Merge} \left(d_i^{(t)}, \{d_{ij}^{(t+1)} \mid r_{ij}^{(t+1)} \in R_i^{(t+1)}\} \right). \quad (4)$$

Here, $r_{ij}^{(t+1)}$ represents a sub-region of $r_i^{(t)}$, ensuring that all finer-level details are recursively embedded into higher-level descriptions, resulting in a comprehensive hierarchical representation of the image.

With the comprehensive multi-level descriptions $d^{(0)}$, along with the original image I and predefined question prompt P , we generate question-answer pairs through the same VLM, focusing on different aspects of the scene, such as object details, spatial relationships, style, scene context, and interactions:

$$QA = \text{VLM}(I, P, d^{(0)}). \quad (5)$$

By structuring the annotation process in this recursive manner, we ensure that both fine-grained and high-level

contextual information is captured, leading to a more detailed and context-aware dataset.

3.2. Joint Bootstrapping Training

In the second stage, we refine the VLM using the self-bootstrapped dataset, enhancing its ability to generate structured scene descriptions and answer complex visual questions, as shown in Fig. 3. During training, the model processes a full image along with k randomly selected sub-regions. This multi-granularity input encourages the model to attend to both global layouts and localized details, promoting better contextual alignment across scales.

For scene description generation, the model is trained to predict a description that includes its own details along with the combined descriptions of all its child regions, optimized by:

$$\mathcal{L}_{\text{desc}} = - \sum_{t=1}^T \log P(c_t \mid I, c_{<t}). \quad (6)$$

For VQA, the model takes the full image, its sub-regions, and a set of up to five diverse questions as input, predicting answers using:

$$\mathcal{L}_{\text{qa}} = - \sum_{i=1}^k y_i \log(\hat{y}_i). \quad (7)$$

To encourage mutual enhancement, both tasks are trained jointly. The final loss function is $\mathcal{L}_{\text{desc}} + \mathcal{L}_{\text{qa}}$,

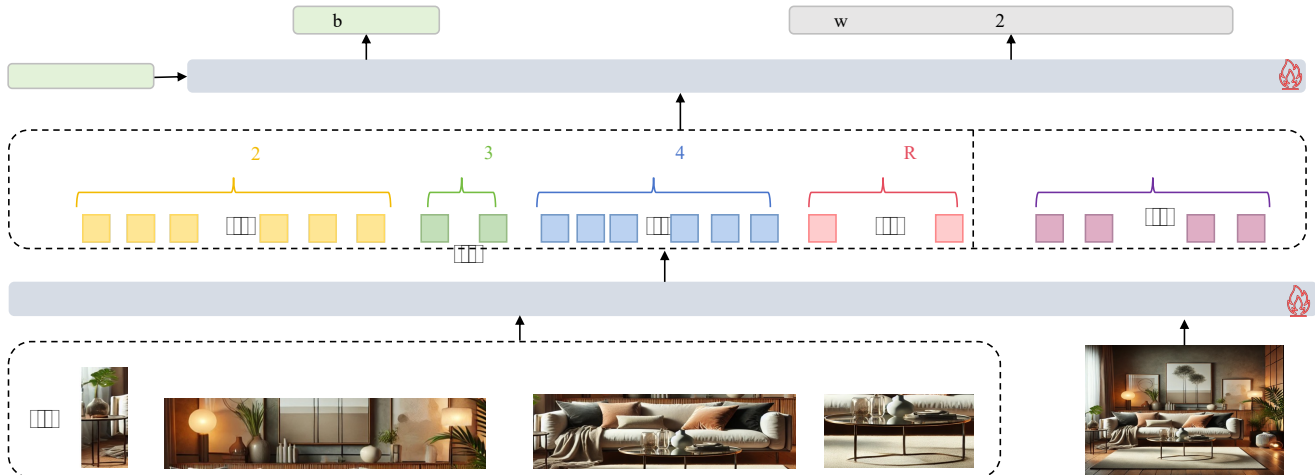


Figure 3. The overall architecture of the OURO model training process. The original image and its sub-regions are input into the Visual Encoder, which extracts both region and global image features. These features are then passed to the LLM Decoder to generate answers to the given questions, while simultaneously training descriptions.

enabling shared visual-language representations to benefit from both generation and reasoning objectives.

Training Data. Several widely recognized datasets have been used to train and validate our model. These datasets span a variety of tasks, including image captioning, general VQA, scene text-centric VQA, and doc-oriented VQA. The Image Caption task utilizes datasets such as COCO Caption [71], TextCaps [58], and Detailed Caption with 404k samples. For General VQA, we make use of VQAv2 [20], OKVQA [44], GQA [26], ScienceQA [42], and VizWiz [24], collectively adding up to 306k samples. The Scene Text-centric VQA task is supported by datasets like TextVQA [59], OCRVQA [28], and AI2D [7], which provide a total of 308k samples. For Doc-oriented VQA, datasets such as DocVQA [47], ChartQA [45], InfoVQA [48], and others, with 423k samples, are employed. The distribution of these datasets across tasks is illustrated in Fig. 4. Based on them, we utilize our pipeline to construct a multimodal dataset with 1.4 million samples, each enriched with detailed annotations and question-answer pairs.

4. Experiment

To evaluate the effectiveness of our model, we conduct extensive experiments on a series of tasks, including image caption and VQA. These tasks are chosen to assess the model’s ability to generate detailed, contextually rich descriptions and accurately answer questions that require reasoning over both global and local visual features.

4.1. Model details

Model Configuration. We use the Qwen2-VL [64] model as the vision encoder and the LLM decoder. This configuration enables efficient processing of both global

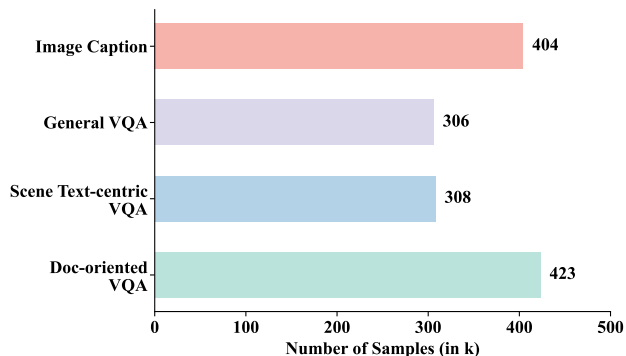


Figure 4. Training dataset distribution across different tasks

image features and fine-grained sub-region features. During training, the original image and its selected sub-regions are processed together through the shared encoder-decoder framework. The model leverages a robust attention mechanism with 28 attention heads, distributed across 28 hidden layers. Key-value head sharing is used with 4 key-value heads, and the root mean square normalization epsilon is set to 1×10^{-6} .

Training Parameters. The model is trained across 8 NVIDIA H100 GPUs, with each device using a batch size of 4, resulting in a total batch size of 32. Gradient accumulation steps are set to 8. The learning rate is set to 1.0×10^{-4} , and a cosine learning rate scheduler is employed, accompanied by a warmup ratio of 0.1. The training process is conducted for 3 epochs using mixed precision (BF16 enabled) to enhance efficiency. Additionally, we incorporate Low-Rank Adaptation (LoRA) [25] with a rank of 8, significantly reducing the number of trainable parameters while maintaining model performance.

Model	OKVQA	VQAV2	VizWiz	GQA	VSR	ScienceQA	IconVQA
BLIP-2-7B [34]	45.9	-	19.6	41.0	50.9	61.0	40.6
InstructBLIP-7B [12]	-	-	33.4	49.5	52.1	-	44.8
LLaMA-AdapterV2-7B [19]	49.6	70.7	39.8	45.1	-	-	-
Shikra-13B [9]	47.2	77.4	-	-	-	-	-
mPLUG-Owl2-7B [70]	57.7	79.4	54.5	56.1	-	68.7	-
Fuyu-8B [6]	60.6	74.2	-	-	-	-	-
MiniGPT-v2-7B [8]	57.8	-	53.6	60.1	62.9	-	51.5
FlexCap-LLM [17]	52.1	65.6	41.8	49.5	-	-	-
Qwen-VL-7B [5]	58.6	79.5	35.2	59.3	<u>63.8</u>	67.1	-
Qwen-VL-7B-Chat [5]	56.6	78.2	38.9	57.5	61.5	68.2	-
LLaVA1.5-7B [39]	-	78.5	50.0	62.0	-	66.8	-
LLaVA1.5-13B [39]	-	80.0	53.6	63.3	-	71.6	-
VisCoT-7B [57]	-	-	-	63.1	61.4	-	-
Monkey-7B [36]	61.3	<u>80.3</u>	61.2	60.7	-	69.4	-
SPHINX-7B [37]	<u>62.1</u>	78.1	39.9	62.6	58.5	69.3	52.7
Qwen2-VL-7B [64]	<u>57.9</u>	75.5	<u>64.7</u>	<u>77.3</u>	-	95.4	-
OURO-7B	66.2	80.8	70.4	77.7	77.0	<u>87.0</u>	<u>51.6</u>

Table 1. Results on General VQA and other related tasks.

Model	DocVQA	ChartQA	InfoVQA	DeepForm	KLC	WTQ
Closed-source Models						
GPT-4o [49]	92.8	85.7	66.4	38.4	29.9	46.6
GeminiPro-1.5 [13]	91.2	34.7	73.9	32.2	24.1	50.3
Claude-3.5 [4]	88.5	51.8	59.1	31.4	24.8	47.1
Open-source Models						
InternVL-2.5-2B [11]	87.7	75.0	61.9	13.1	16.6	36.3
DeepSeek-VL2-Tiny [67]	88.6	81.0	63.9	25.1	19.0	35.1
Phi3.5-Vision [1]	86.0	82.2	56.2	10.5	7.5	17.2
LLaVA-NeXT-7B [23]	63.5	52.1	30.9	1.3	5.4	20.1
Llama3.2-11B [21]	82.7	23.8	36.6	1.8	3.5	23.0
ALIGNVLM-8B [46]	81.2	75.0	53.8	63.3	<u>35.5</u>	45.3
Qwen-VL-7B [5]	65.1	65.7	35.4	4.1	15.9	21.6
Monkey [36]	66.5	65.1	36.1	40.6	32.8	25.3
Qwen2-VL-7B [64]	91.4	73.5	<u>76.8</u>	42.6	30.6	<u>57.9</u>
OURO-7B	93.5	<u>84.1</u>	79.1	<u>52.5</u>	56.2	72.0

Table 2. Results on Doc-oriented VQA.

Model	TextVQA	AI2D	STVQA	ESTVQA
Pix2Struct-Large [32]	-	42.1	-	-
BLIP-2 [34]	42.4	-	-	-
InstructBLIP [12]	50.7	-	-	-
mPLUG-DocOwl-7B [69]	52.6	-	-	-
mPLUG-Owl2-7B [70]	54.3	-	-	-
Qwen-VL-7B [5]	63.8	62.3	59.1	77.8
Qwen-VL-Chat-7B [5]	61.5	57.7	-	-
LLaVA-1.5 [39]	58.2	-	-	-
Monkey-7B [36]	67.6	62.6	<u>67.7</u>	82.6
Qwen2-VL-7B [64]	<u>82.2</u>	<u>77.6</u>	61.3	<u>83.7</u>
OURO-7B	85.3	80.2	77.0	90.2

Table 3. Results on Scene Text-centric VQA.

4.2. Results

Our evaluation covers general VQA, document-oriented VQA, and scene text-centric VQA, as well as fine-grained

Model	CoCo Caption	Flickr30K	TextCaps
Flamingo-80B [3]	-	67.2	-
Palm-E-12B [15]	135.0	-	-
BLIP-2 [34]	-	71.6	-
InstructBLIP (Vicuna-13B) [12]	102.2	82.8	-
Shikra (Vicuna-13B) [9]	117.5	73.9	-
mPLUG-Owl2-7B [70]	137.3	85.1	-
LLaVA1.5 (Vicuna-7B) [39]	-	-	-
Qwen-VL (Qwen-7B) [5]	-	85.8	65.1
Qwen-VL-Chat-7B [5]	131.9	81.0	-
Monkey-7B[36]	-	<u>86.1</u>	93.2
OURO-7B	<u>136.2</u>	88.0	<u>89.7</u>

Table 4. Results on Image Captions Tasks.

captioning to demonstrate the model’s capacity for detailed and spatially-aware scene interpretation.

VQA Results. We evaluate OURO across multiple VQA benchmarks, covering general VQA, scene text-centric VQA, and document-oriented VQA. The general VQA tasks include OKVQA [44], VQAv2 [20], VizWiz [24], GQA [26], VSR [38], ScienceQA [42], and IconVQA[41], focusing on diverse real-world scenarios. The scene text-centric VQA tasks, such as TextVQA [59], AI2D [28], STVQA [7], and ESTVQA [65], assess the model’s ability to process embedded text. Additionally, document-oriented VQA, evaluated on DocVQA [47], ChartQA [45], InfoVQA[48], DeepForm [61], KLC [60], and WTQ [51], examines structured document reasoning. As shown in Table 1, OURO surpasses previous similarly-sized models such as Monkey and SPHINX. Notably, in scene text-centric VQA, OURO attains 85.3 on TextVQA and 80.2 on AI2D, significantly outperforming previous approaches (Table 3). Additionally, in document-oriented VQA, OURO achieves top-tier performance, outperforming or matching

k	OKVQA	VQAv2	VizWiz	GQA	ChartQA	ScienceQA	TextVQA	AI2D	STVQA	ESTVQA	Average
0	61.3	77.7	73.3	69.0	76.8	<u>86.7</u>	83.0	83.5	72.3	89.4	77.3
1	62.5	75.3	76.4	66.0	62.4	<u>85.8</u>	77.3	76.3	69.8	87.7	74.0
2	66.0	79.7	75.9	<u>73.6</u>	<u>80.5</u>	80.8	<u>85.1</u>	86.0	72.4	88.4	<u>78.8</u>
3	65.9	79.9	77.5	71.1	79.1	79.5	84.6	<u>86.5</u>	<u>73.5</u>	89.2	78.7
5	<u>66.2</u>	80.8	<u>70.4</u>	77.7	84.1	87.0	85.3	80.2	77.0	90.2	80.0
7	70.3	<u>78.1</u>	79.2	71.2	63.5	79.7	77.1	76.6	72.2	87.7	75.6

Table 5. Ablation study on the number of sub-regions used during training.

k	Annot.	OKVQA	VQAv2	ChartQA	TextVQA	AI2D	STVQA
3	Orig.	58.4	72.3	72.1	76.7	66.0	67.9
3	Ours	<u>65.9</u>	79.9	79.1	84.6	<u>86.5</u>	<u>73.5</u>
5	Orig.	66.2	79.6	80.2	<u>85.1</u>	85.9	72.8
5	Ours	66.2	80.8	84.1	85.3	80.2	77.0

Table 6. Ablation study on enhanced annotations.

12 state-of-the-art models across six datasets, as shown in Table 2. These results demonstrate OURO’s ability to capture fine-grained text details, spatial structure, and document layout, leading to more accurate and interpretable VQA responses. Furthermore, Fig. 5 (b-e) presents qualitative comparisons across different VQA aspects using OURO, ChatGPT-4o [49], Qwen2-VL [64] and DeepSeek-VL [67]. Specifically, (a) corresponds to scene reasoning, (b) to structured table understanding, (c) to spatial reasoning, and (d) to knowledge-based question answering. OURO successfully answers all questions, demonstrating its strong ability to integrate spatial, structural, and contextual information for accurate and interpretable responses.

Image Captioning Results. We evaluate OURO on image captioning tasks using the CoCo Caption [10], Flickr30K [71], and TextCaps [58] datasets. As shown in Table 4, OURO achieves the best performance on Flickr30K while maintaining competitive results on CoCo Caption and TextCaps. This highlights its ability to generate structured and contextually rich descriptions by leveraging hierarchical scene understanding and spatially aware captioning mechanisms. Fig. 5 (e) compares responses from OURO, ChatGPT-4o [49], Qwen2-VL [64] and DeepSeek-VL [67] to the prompt, “Please describe the sculpture in the image in details.” OURO provides a more structured and precise description, accurately identifying each of the seven sculptures, detailing their posture, movement, and abstract forms, while other models offer more generic responses.

4.3. Ablation Study

We conduct ablation experiments to verify the effectiveness of our method.

Impact of sub-region number. We analyze the effect of sub-region selection on performance across multiple datasets. As shown in Tab. 5, incorporating sub-regions

consistently improves results, with the best performance achieved when using $k = 5$ sub-regions, yielding the highest average score. Using fewer than five sub-regions may limit fine-grained detail capture, as the model has less localized information to refine its understanding. Meanwhile, using more than five sub-regions, particularly seven, results in performance degradation, likely due to an excessive focus on small regions, reducing the model’s ability to maintain global context. Interestingly, using only one sub-region performs worse than using none at all (74.0 vs. 77.3), possibly because randomly selecting a single region risks excluding crucial information, leading to inconsistencies in scene interpretation. These results emphasize the importance of balancing local and global context through an optimal number of sub-regions, with five providing the most effective trade-off.

Impact of enhanced annotations. To assess our recursive annotation pipeline, we compare models trained on original (i.e., dataset-provided human-written annotations) vs. enhanced annotations using $k = 3$ and $k = 5$ sub-images per sample (Tab. 6). Note that all models are trained with multi-subregion inputs; the only difference lies in the supervision: original annotations use individual human-written captions and QA, while enhanced annotations employ structured merged descriptions and self-generated QA.

Results show consistent improvements from enhanced annotations across benchmarks. For example, $k = 3$ (Ours) surpasses $k = 3$ (Orig.) by +7.6 on VQAv2 and +20.5 on AI2D, highlighting the value of hierarchical and QA-augmented supervision. Moreover, $k = 3$ (Ours) performs comparably to $k = 5$ (Orig.), indicating that annotation quality outweighs mere increases in subregion count. These results validate the effectiveness of self-bootstrapped annotation in enhancing scene understanding and reasoning.

4.4. Limitations

Our approach to multi-level scene understanding introduces challenges related to description length and consistency. As the hierarchical generation process expands across multiple levels, captions may become excessively long, leading to redundancy or potential misalignment between different levels of descriptions. Future research should explore strategies to generate concise yet comprehensive scene descrip-



Figure 5. Qualitative comparison of scene descriptions and VQA responses across different datasets, illustrating outputs from our model, ChatGPT-4o, Qwen2-VL and DeepSeek-VL .

tions while ensuring accurate alignment between hierarchical representations and their corresponding VQA pairs. Additionally, our current method selects sub-regions randomly, which lacks interpretability and may not always focus on the most relevant areas for question answering. This randomness in region selection can impact the model’s ability to prioritize informative regions effectively. Future work could explore more structured selection mechanisms that enhance interpretability, allowing for a more deliberate and explainable sub-region selection process.

5. Conclusion

We introduce a self-bootstrapped approach for enhancing scene understanding in multimodal models, and develop a multimodal model, OURO. By generating hierarchical scene descriptions and structured VQA data, our method captures detailed object attributes and spatial relationships without relying on costly manual annotations. Additionally, our joint training strategy integrates both full-image and sub-region features, improving the model’s ability to balance local and global context.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under STI 2030—Major Projects (No. 2021ZD0200300), in part by the National Key Research and Development Program of China (No. 2024YDLN0006), in part by the National Natural Science Foundation of China (Grant No. 62176133), in part by the Tsinghua-Meituan Joint Institute for Digital Life under Agreement No. C0210322000380, in part by the Tsinghua-Fuzhou Data Technology Joint Research Institute (Project No. JIDT2024013), and in part by Qualcomm Technologies, Inc. under Statement of Work No. TSI-617560.

References

- [1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 6
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1
- [3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1, 3, 6
- [4] Anthropic. Model card addendum: Claude 3.5 haiku and upgraded claude 3.5 sonnet, 2024. Accessed: 2025-03-07. 6
- [5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023. 3, 6
- [6] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sagnak Ta. s1rlar. *Introducing our multimodal models*, 2, 2023. 6
- [7] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluís Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimos-thenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301, 2019. 5, 6
- [8] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 6
- [9] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 6
- [10] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 7
- [11] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 6
- [12] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven CH Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arxiv abs/2305.06500* (2023), 2023. 6
- [13] Google DeepMind. Gemini pro-1.5: Technical report, 2024. Accessed: 2025-03-07. 6
- [14] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1
- [15] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palme: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 6
- [16] Debidatta Dwibedi, Vidhi Jain, Jonathan Tompson, Andrew Zisserman, and Yusuf Aytar. Flexcap: Generating rich, localized, and flexible captions in images. *arXiv preprint arXiv:2403.12026*, 2024. 2, 3
- [17] Debidatta Dwibedi, Vidhi Jain, Jonathan J Tompson, Andrew Zisserman, and Yusuf Aytar. Flexcap: Describe anything in images in controllable detail. *Advances in Neural Information Processing Systems*, 37:111172–111198, 2025. 6
- [18] Roy Ganz and Michael Elad. Clipag: Towards generator-free text-to-image generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3843–3853, 2024. 3
- [19] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xianguyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 6
- [20] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 5, 6
- [21] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 6
- [22] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10867–10877, 2023. 3

- [23] Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*, pages 390–406. Springer, 2024. 6
- [24] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 5, 6
- [25] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 5
- [26] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 5, 6
- [27] Noman Islam, Zeeshan Islam, and Nazia Noor. A survey on optical character recognition system. *arXiv preprint arXiv:1710.05703*, 2017. 3
- [28] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016. 5, 6
- [29] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122, 2023. 1
- [30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2
- [31] Simon Kornblith, Lala Li, Zirui Wang, and Thao Nguyen. Guiding image captioning models toward more specific captions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15259–15269, 2023. 3
- [32] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR, 2023. 6
- [33] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1, 3
- [34] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 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, 2023. 6
- [35] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022. 2
- [36] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 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, 2024. 2, 3, 6
- [37] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 2, 3, 6
- [38] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 6
- [39] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 6
- [40] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 2
- [41] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021. 6
- [42] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 5, 6
- [43] Duc-Tuan Luu, Viet-Tuan Le, and Duc Minh Vo. Questioning, answering, and captioning for zero-shot detailed image caption. In *Proceedings of the Asian Conference on Computer Vision*, pages 242–259, 2024. 3
- [44] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 5, 6
- [45] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 5, 6
- [46] Ahmed Masry, Juan A Rodriguez, Tianyu Zhang, Suyuchen Wang, Chao Wang, Aarash Feizi, Akshay Kalkunte Suresh,

- Abhay Puri, Xiangru Jian, Pierre-André Noël, et al. Align-*v*lm: Bridging vision and language latent spaces for multimodal understanding. *arXiv preprint arXiv:2502.01341*, 2025. 6
- [47] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 5, 6
- [48] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 5, 6
- [49] Microsoft. Introducing gpt-4o-2024-08-06 api with structured outputs on azure. <https://techcommunity.microsoft.com/blog/azure-ai-services-blog/introducing-gpt-4o-2024-08-06-api-with-structured-outputs-on-azure/4232684>, 2024. Accessed: 2025-03-07. 6, 7
- [50] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, 2011. 1
- [51] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*, 2015. 6
- [52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3
- [53] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 2
- [54] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaohe Jiang, Yihao Chen, et al. Grounding dino 1.5: Advance the “edge” of open-set object detection. *arXiv preprint arXiv:2405.10300*, 2024. 2
- [55] Noam Rotstein, David Bensaïd, Shaked Brody, Roy Ganz, and Ron Kimmel. Fusecap: Leveraging large language models for enriched fused image captions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5689–5700, 2024. 3
- [56] Sara Sarto, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Positive-augmented contrastive learning for vision-and-language evaluation and training. *arXiv preprint arXiv:2410.07336*, 2024. 3
- [57] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2025. 2, 3, 6
- [58] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020. 5, 7
- [59] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 5, 6
- [60] Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. Kleister: key information extraction datasets involving long documents with complex layouts. In *International Conference on Document Analysis and Recognition*, pages 564–579. Springer, 2021. 6
- [61] S Svetlichnaya. Deepform: Understand structured documents at scale. 2020. 6
- [62] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3
- [63] Hao Wang, Pengzhen Ren, Zequn Jie, Xiao Dong, Chengjian Feng, Yinlong Qian, Lin Ma, Dongmei Jiang, Yaowei Wang, Xiangyuan Lan, et al. Ov-dino: Unified open-vocabulary detection with language-aware selective fusion. *arXiv preprint arXiv:2407.07844*, 2024. 2
- [64] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 5, 6, 7
- [65] Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10126–10135, 2020. 6
- [66] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023. 1
- [67] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 6, 7
- [68] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211, 2024. 1
- [69] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 6

- [70] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051, 2024. [2](#), [3](#), [6](#)
- [71] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. [5](#), [7](#)
- [72] Qiyang Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. Capsfusion: Rethinking image-text data at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14022–14032, 2024. [3](#)
- [73] Jingyi Zhang, Jiaying Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. [1](#)
- [74] Chujie Zhao, Tianren Zhang, Guanyu Chen, Yizhou Jiang, and Feng Chen. M3pl: Identifying and exploiting view bias of prompt learning. *Transactions on Machine Learning Research*, 2024. [1](#)
- [75] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. [1](#)