

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

Tianrun Xu\*, Guanyu Chen\*, Ye Li, Yuxin Xi, Zeyu Mu, Ruichen  
Wang, Tianren Zhang, Haichuan Gao†, Feng Chen†

# Motivation: Why Fine-Grained Understanding?

Global captions miss object-level attributes and spatial relations.  
High-quality fine-grained labels are costly; scaling is difficult.  
We need hierarchical, multi-granularity scene representations.

# Key Contributions

Self-bootstrapped pipeline (base VLM + RPN) for hierarchical annotations; no extra human labels.

Large-scale multi-granularity corpus (captions + QA) built automatically.

Improved performance across 20+ benchmarks and multiple task families.

# OURO at a Glance

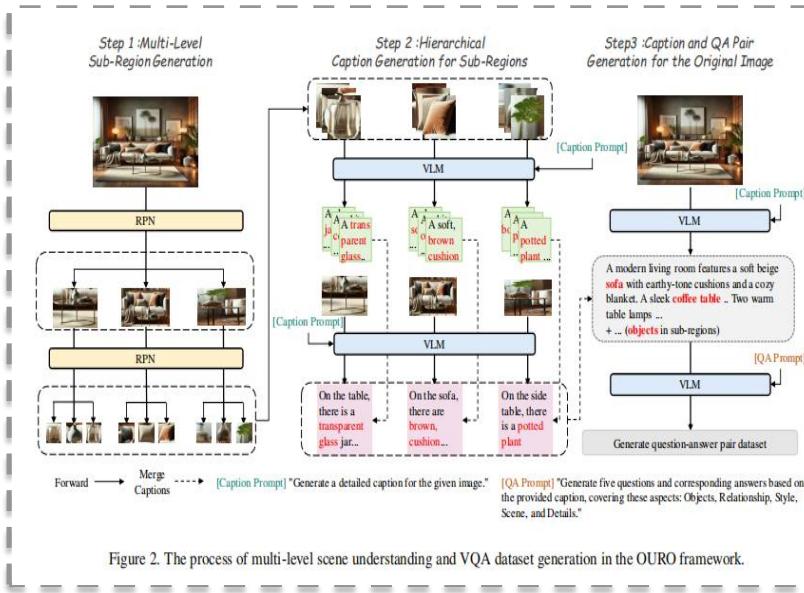


Figure 2. The process of multi-level scene understanding and VQA dataset generation in the OIRO framework.

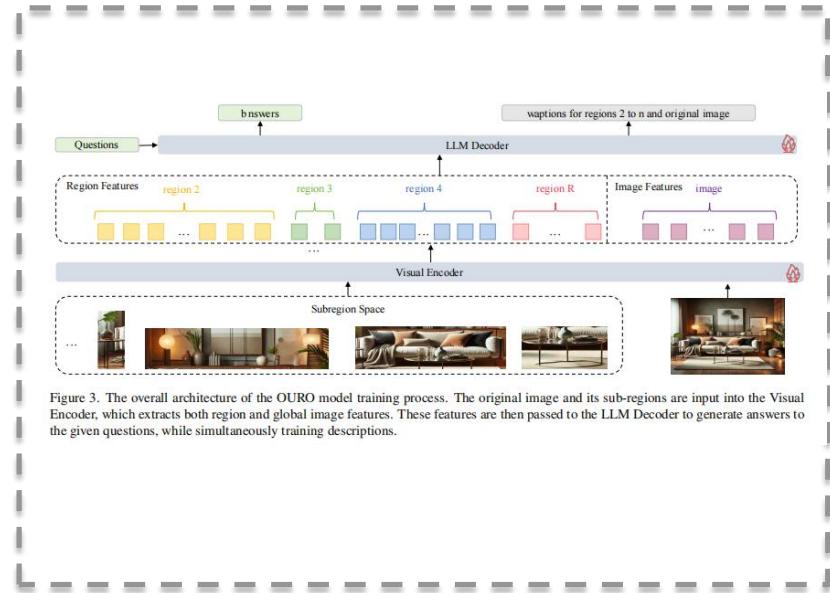


Figure 3. The overall architecture of the OIRO model training process. The original image and its sub-regions are input into the Visual Encoder, which extracts both region and global image features. These features are then passed to the LLM Decoder to generate answers to the given questions, while simultaneously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

Two-stage framework: generate hierarchical data, then train jointly on global+local inputs.

Promotes interpretability via hierarchy and robustness via joint objectives.

# Stage I — Multi-Level Scene Annotation (Intuition)

RPN proposes sub-regions; VLM describes each region.  
 Merge local descriptions back to parents for hierarchical captions.  
 Generate QA pairs from the hierarchical descriptions.

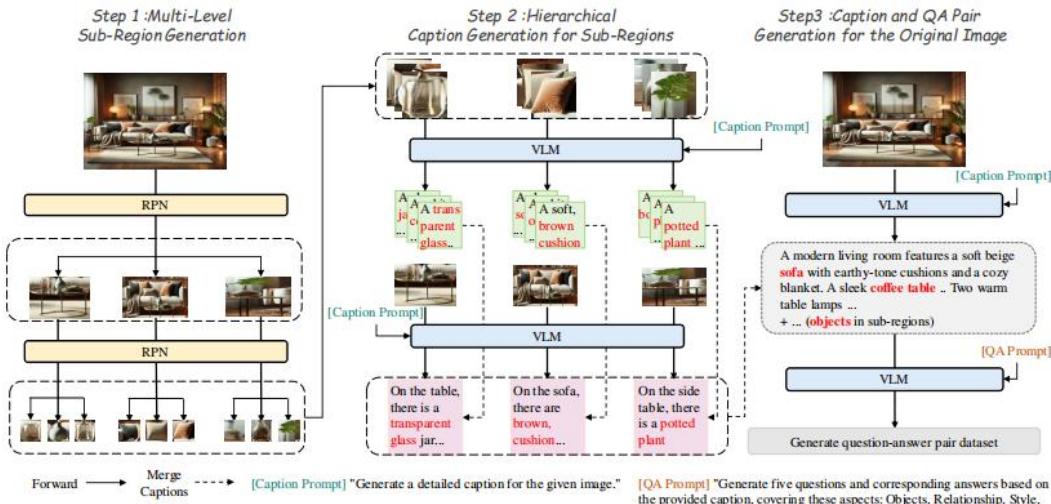


Figure 2. The process of multi-level scene understanding and VQA dataset generation in the OURO framework.

---

#### Algorithm 1 Recursive Scene Annotation with VLM

---

**Require:** Image  $I$ , Prompt  $P$ , Confidence threshold  $\tau$   
**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:       $D^{(t+1)} \leftarrow D^{(t+1)} \cup \{d_i^{(t)}\}$ 
14:    end for
15:     $d^{(t)} \leftarrow \text{Merge}(d^{(t)}, D^{(t+1)})$ 
16:  end if
17:  return  $d^{(t)}$ 
18: end function

```

---

# Stage II — Joint Bootstrapping Training

Input: full image +  $k$  sampled sub-regions per instance.

Objectives: caption loss + QA loss; shared encoder–decoder.

Balances global context with local details.

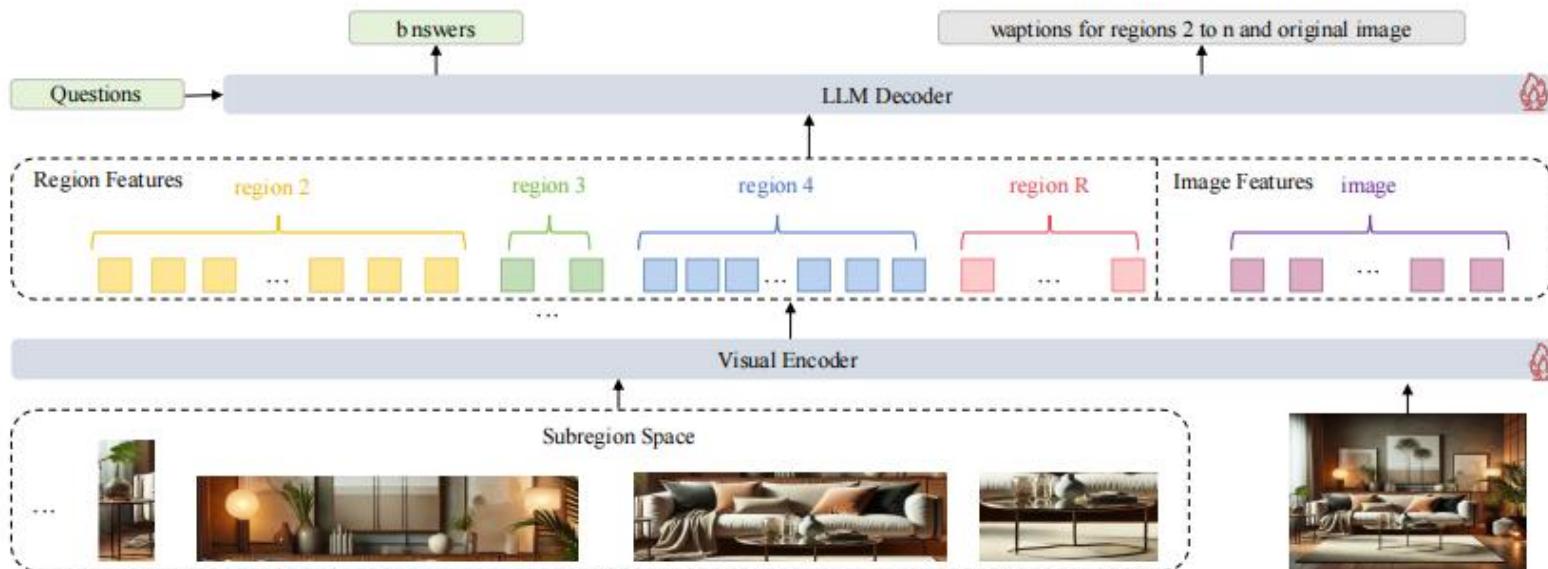


Figure 3. The overall architecture of the OURO model training process. The original image and its sub-regions are input into the Visual Encoder, which extracts both region and global image features. These features are then passed to the LLM Decoder to generate answers to the given questions, while simultaneously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

# Data & Training Setup

Tasks: Captioning, General VQA, Scene-Text VQA, Document VQA.  
LoRA rank, epochs, LR schedule, precision, hardware (fill from paper).

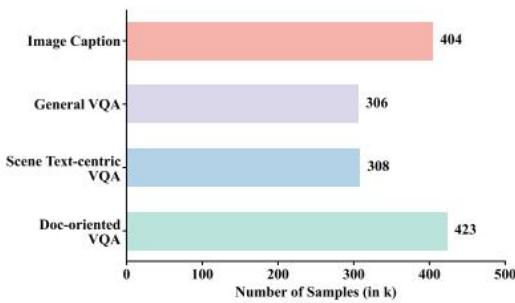


Figure 4. Training dataset distribution across different tasks

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], In\_x0002\_foVQA [48], and others, with 423k samples, are employed.

# Results — General VQA

OURO outperforms base and peer models on multiple general VQA datasets. Highlight key numbers (e.g., OKVQA, VQAv2, VizWiz, GQA).

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	<u>61.5</u>	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	<b>52.7</b>
Qwen2-VL-7B [64]	57.9	75.5	<u>64.7</u>	<u>77.3</u>	-	<b>95.4</b>	-
OURO-7B	<b>66.2</b>	<b>80.8</b>	<b>70.4</b>	<b>77.7</b>	<b>77.0</b>	87.0	<u>51.6</u>

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

# Results — Document-Oriented VQA and Scene Text-Centric VQA

Strong results on DocVQA/ChartQA/InfoVQA/WTQ;

Model	DocVQA	ChartQA	InfoVQA	DeepForm	KLC	WTQ
<b>Closed-source Models</b>						
GPT-4o [49]	92.8	<b>85.7</b>	66.4	38.4	29.9	46.6
GeminiPro-1.5 [13]	91.2	34.7	73.9	32.2	<b>24.1</b>	50.3
Claude-3.5 [4]	88.5	51.8	59.1	31.4	24.8	<b>47.1</b>
<b>Open-source Models</b>						
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	<b>19.0</b>	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	<b>63.3</b>	<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	<b>93.5</b>	<u>84.1</u>	<b>79.1</b>	<u>52.5</u>	<b>56.2</b>	<b>72.0</b>

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	<b>67.7</b>	82.6
Qwen2-VL-7B [64]	<u>82.2</u>	<u>77.6</u>	<u>61.3</u>	<u>83.7</u>
OURO-7B	<b>85.3</b>	<b>80.2</b>	<b>77.0</b>	<b>90.2</b>

Table 3. Results on Scene Text-centric VQA.

# Visualization

Is this bathroom high or low end?

✓ This bathroom is high end. ✗

✓ This bathroom leans towards high-end but not ultra-luxury. With some modern updates... ✗

✗ It is ambiguous to determine if the bathroom is high or low end. ✗

✗ The bathroom appears to be of the lower-end based on its design elements such as well as overall aesthetics. ✗

What is the second heading of the table?

✓ FIRST PERIOD ENDING OCTOBER 7, 1963. ✗

✓ The image resolution is too low for me to accurately read the text. However, you can try uploading a higher-resolution version... ✗

✗ PENNANT RATING PREMIUMS AND PENALTIES - (PENALTIES UNDERLINED). ✗

✗ Mol. ✗

The parking meter is at the back of the car.

✓ No. ✗

✓ No, the parking meter is in front of the car, not at the back. The image shows a parking meter standing in the foreground, while a black car is positioned behind it. ✗

✗ The parking meter is at the back of the car. ✗

✓ No, the parking meter is not at the back of the car ✗

Which environmental issue is the major concern of Irish?

✓ health impact of chemicals. ✗

✗ The major environmental issue of concern for Irish citizens, according to the infographic, is pollution with 44% of citizens expressing concern about it. ✗

✗ The major environmental issue is water pollution. ✗

✓ Water pollution (43%). ✗

Please describe the sculpture in the image in details.

There are seven sculptures in the central exhibition area. The **leading sculpture** leans forward dynamically, with one leg extended as if in motion. The **second sculpture** is slightly hunched, with a twisted torso and extended limbs, creating a sense of movement. The **third sculpture** stands more upright, with arms reaching forward and a rough, textured surface enhancing its depth. The **fourth sculpture** bends at the waist with fragmented limbs, appearing to brace against an unseen force. The **fifth sculpture** has an angular stance, with one leg extended backward and arms positioned as if pushing forward. The **sixth sculpture** is compact and crouched, with its head tilted downward and an abstract, layered surface. The **seventh sculpture** stands tall and rearing, with arms slightly lifted and a jagged form that conveys tension and energy. They resemble a group of people wearing heavy clothing, appearing to be walking forward dynamically. ✗

The intermediate sculpture in the image is a modern and artistic installation located in a shopping mall. The sculpture features a group of figures that appear to be made from various materials, including metal and fabric. The figures are designed to resemble humanoid shapes with exaggerated features, such as large heads and expressive eyes. They are arranged in a dynamic pose, giving the impression that they are dancing. ✗

The sculptures in the center appear to be made from metallic materials with intricate designs that resemble animal forms or abstract figures. They have shiny surfaces reflecting light, giving them an eye-catching appearance against the blue carpeted platform they stand upon. The arrangement suggests movement as if these creatures were captured mid-action within their display. ✗

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 .

# Limitations & Future Work

Conciseness & alignment for long hierarchical captions.

From random sub-region sampling to policy-guided, interpretable selection.

Further optimize training/inference efficiency.