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

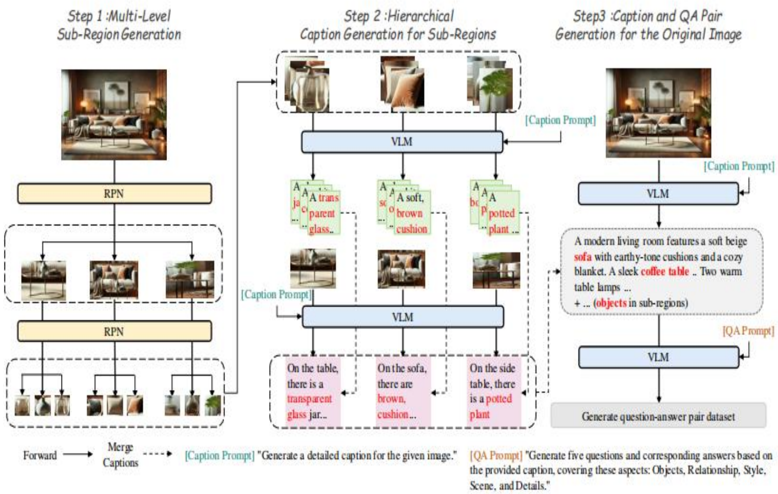

Figure 2. The process of multi-level scene understanding and VQA dataset generation in the OURO framework.

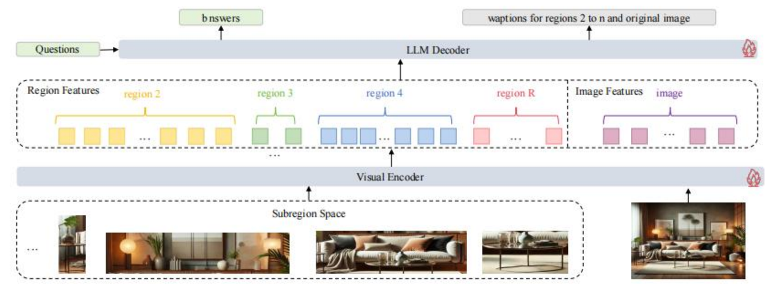

Figure 3. The overall architecture of the OURO model training process. The original image and its sub-regions are input into the Visual Encoder, which extracts both region and global image features. These features are then passed to the LLM Decoder to generate answers to the given questions, while simultaneously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

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.

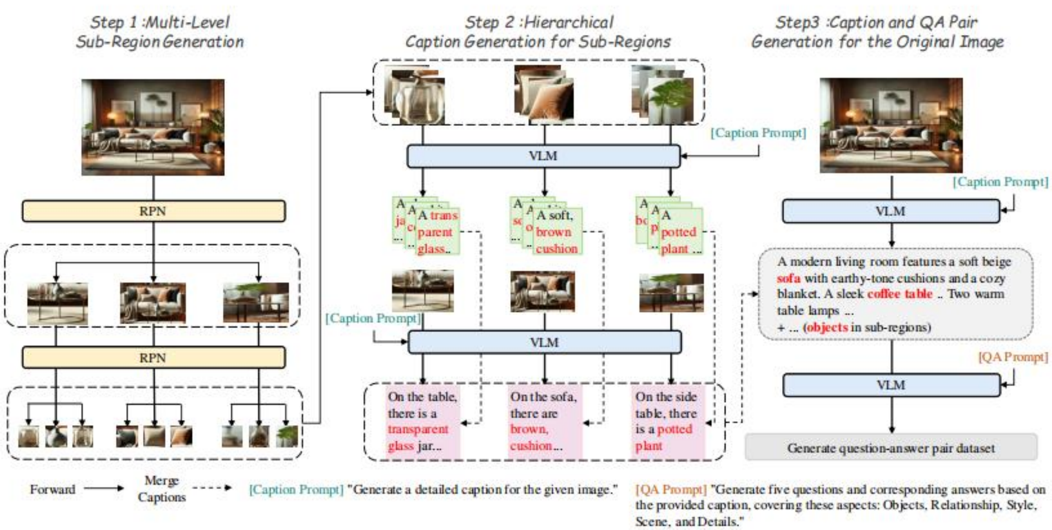

Figure 2. The process of multi-level scene understanding and VQA dataset generation in the OURO framework.

Algorithm 1 Recursive Scene Annotation with VLM

Require: Image I, Prompt P, Confidence threshold τ
Ensure: Hierarchical descriptions $d^{(0)}$ and QA pairs QA

```
function RECURSIVEANNOTATE(I)
    d^(0) ← RecursiveDescribe(I, 0)
    QA ← VLM(I, P, d^(0))          ▷ Generate QA pairs using descriptions
    return d^(0), QA
end function
function RECURSIVEDESCRIBE(r^(t), t)
    d^(t) ← VLM(r^(t))             ▷ Generate description
    R^(t+1) ← RPN(r^(t))           ▷ Generate sub-regions
    if R^(t+1) ≠ ∅ then
        D^(t+1) ← ∅
        for each r_i^(t+1) ∈ R^(t+1) do
            d_i^(t+1) ← RecursiveDescribe(r_i^(t+1), t + 1)
            D^(t+1) ← D^(t+1) ∪ {d_i^(t)}
        end for
        d^(t) ← Merge(d^(t), D^(t+1))
    end if
    return d^(t)
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.

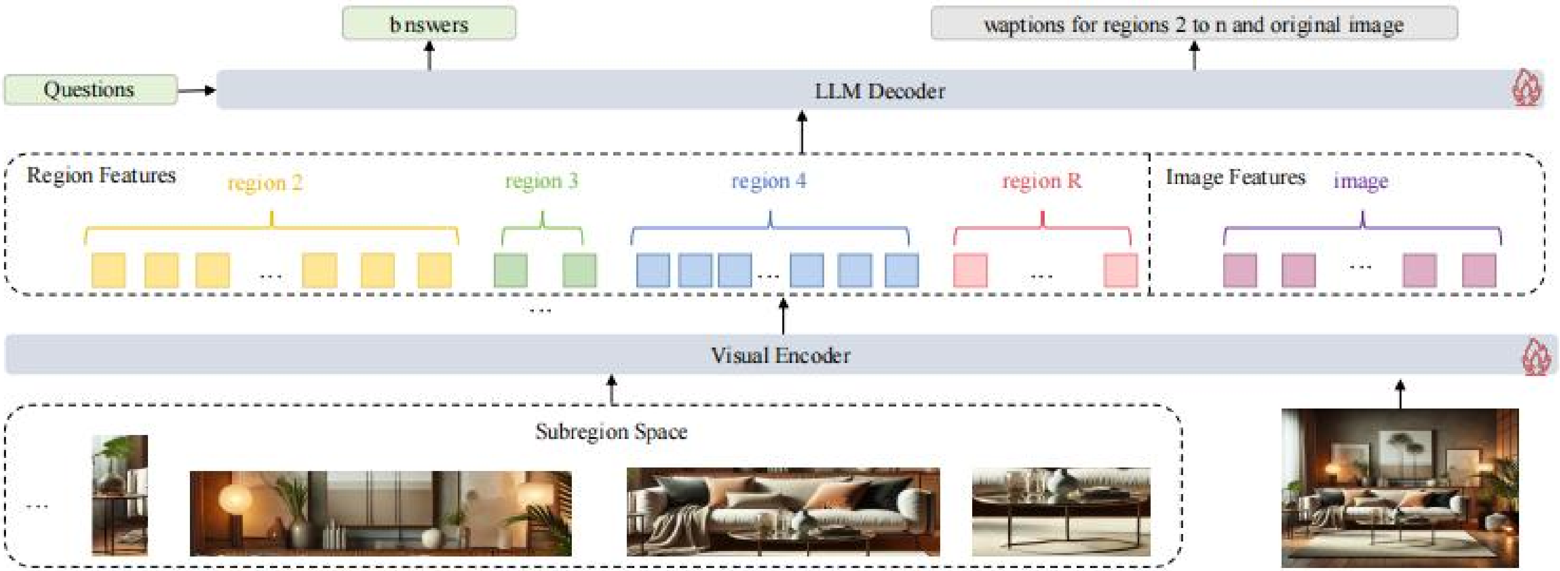

Figure 3. The overall architecture of the OURO model training process. The original image and its sub-regions are input into the Visual Encoder, which extracts both region and global image features. These features are then passed to the LLM Decoder to generate answers to the given questions, while simultaneously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

Data & Training Setup

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

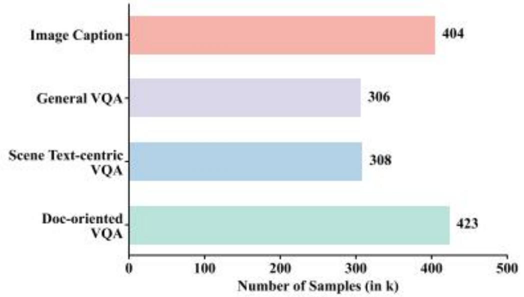

Figure 4. Training dataset distribution across different tasks

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 sam_x0002_ples. 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	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]	57.9	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.

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

Strong results on DocVQA/ChartQA/InfoVQA/WTQ;

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**	35.5	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	76.8	42.6	30.6	57.9
OURO-7B	**93.5**	84.1	**79.1**	52.5	**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	67.7	82.6
Qwen2-VL-7B [64]	82.2	77.6	61.3	83.7
OURO-7B	**85.3**	**80.2**	**77.0**	**90.2**

Table 3. Results on Scene Text-centric VQA.

Visualization

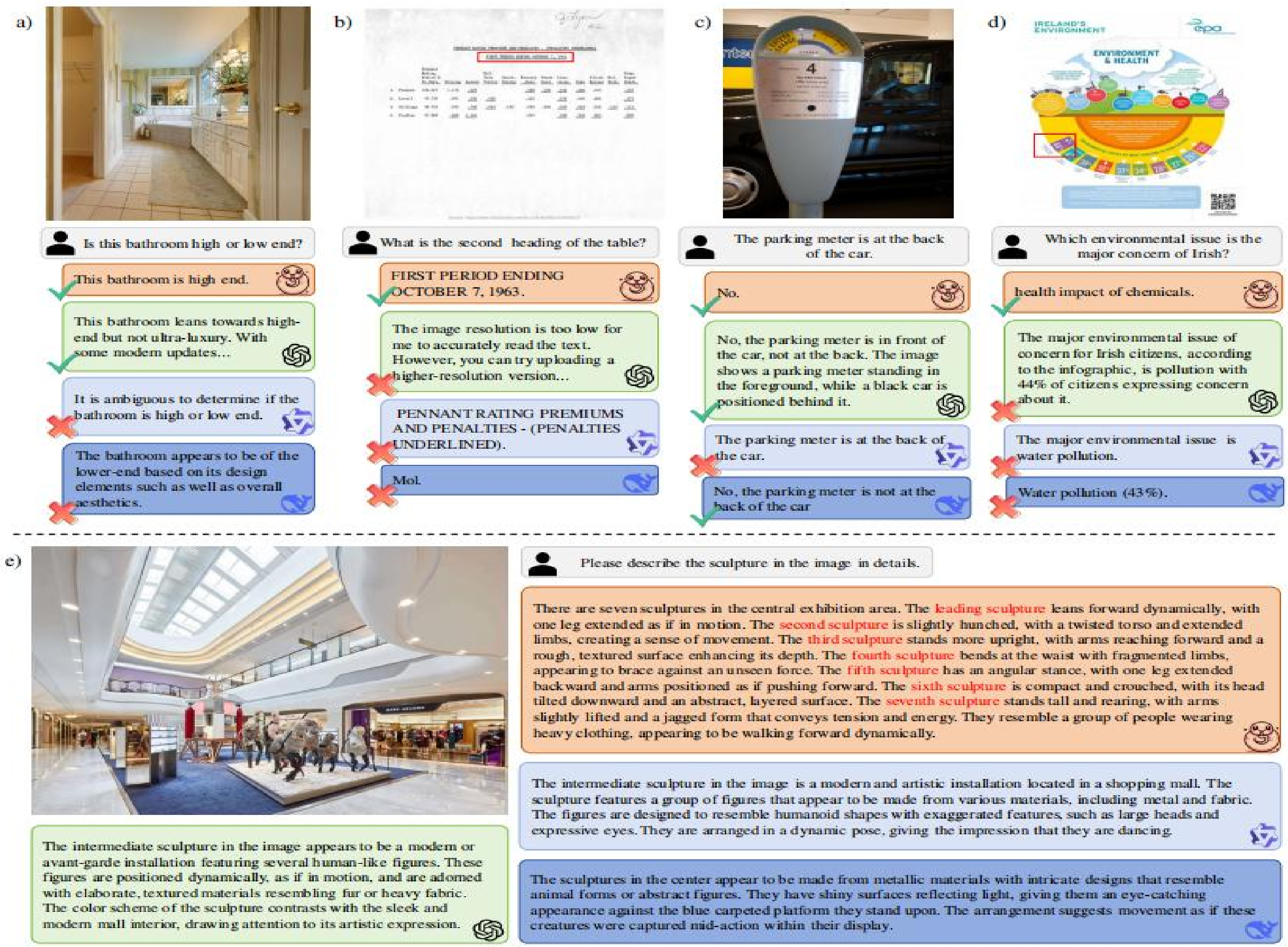

Figure 5. Qualitative comparison of scene descriptions and VQA responses across different datasets, illustrating outputs from our model, ChatGPT-4o, Qwen2-VL and DeepSeek-VL .

Limitations & Future Work

Conciseness & alignment for long hierarchical captions.
From random sub-region sampling to policy-guided, interpretable selection.
Further optimize training/inference efficiency.