

VoQA: Visual-only Question Answering

Supplementary Material

A. Dataset Construction

A.1. Text-to-Visual Question Synthesis

Given a scene image I_s and a textual question T_q , to generate the composite image I_v in the VoQA task, we first convert T_q into RGB-format images I_q of size 224×224 using the *DejaVuSans-Bold* font.

A.2. Concatenation Methods and Results

Concatenation Process. Following CLIPPO [5], we first try to concatenate the scene image I_s with the rendered text image I_q . Considering the difference of side lengths between I_s and I_q , we use two concatenation methods with different effects, and also take into account the four relative positions of the textual question on the top, bottom, left and right of the composite image.

For the concatenation method without resizing (Figure A1, left), we align the center of the two images and fill in the blank space with the padding of the white background. For the concatenation method with resizing (Figure A1, right), we fix the larger size in I_s and I_q , and enlarge the smaller size to the same side length.

Experiments. We first explored concatenation as a simple baseline, which performed better than watermark rendering thanks to its fixed layout and clean background. We averaged results across four concatenation directions for each method and evaluated model performance in three settings: the traditional VQA task, the *pure visual-only* VoQA task, and the VoQA task with Light Prompt (as described in Section B.1). Results are shown in Table A1 and Table A2. Even under this simplified setup, where the scene image and the embedded question are completely separated, models still struggle to perform well. The effect of Light Prompt also varies across models, sometimes improving but sometimes reducing performance. Overall, both concatenation methods outperform watermark rendering in the pure visual-only setting, consistent with their simpler visual composition.

A.3. Details of Watermark Rendering Method

We first identify the most suitable region for rendering the watermark to minimize the impact of poor text readability. The watermark’s side length is set to 1/4 of the short side of the scene image I_s . To balance efficiency and quality, candidate regions are generated using a stride of 1/4 the image side length. Each region is scored based on gradient, variance, and contrast (weighted 0.4, 0.4, and 0.2, respec-

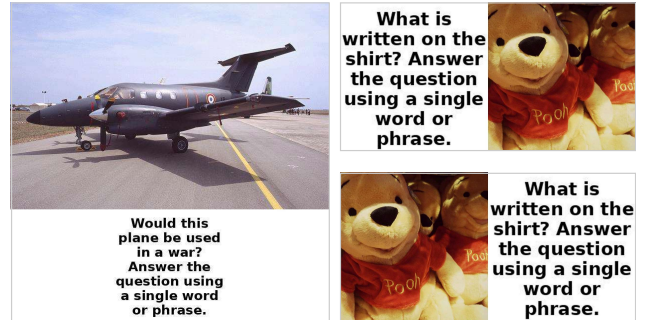


Figure A1. Examples of concatenation methods, showing bottom concatenation without resizing (left) and left / right concatenation with resizing (right).

tively), reflecting smoothness and color contrast to ensure readability.

Next, we determine the watermark color in HSV space based on the average hue, saturation, and value of the selected region. The hue is set to the complementary hue (offset by 180° on the color wheel). If the saturation exceeds 200, the watermark saturation is set to 0.8 times that of the region; otherwise, it is set to the maximum value of 255. The value (brightness) is inverted: if the region’s value exceeds 127, the watermark value is set to 0; otherwise, 255.

Finally, we render I_q onto I_s . Following the WCAG standard¹, if the contrast ratio between the watermark and background exceeds 4.5, we retain the computed color; otherwise, we choose either black or white, selecting the one with greater contrast.

A.4. Details of VoQA Benchmark

We remove reference OCR tokens from the TextVQA questions to prevent excessively long text from reducing font size and interfering with the question information already present in the image. To ensure visual clarity, we filter out overly long questions in SQA to avoid unreadable text regions.

Our benchmark includes three question types: open-ended questions (VQAv2, GQA, TextVQA), binary questions (POPE), and multiple-choice questions (SQA, with A/B/C/D options embedded directly in the images). The final number of evaluation samples retained for each dataset is 107,394 (VQAv2), 12,578 (GQA), 8,910 (POPE), 5,000 (TextVQA), and 1,087 (SQA).

¹<https://www.w3.org/TR/WCAG21/>

Table A1. Evaluation results of *concatenation method without resizing*. We evaluate four open-source models under three settings—the traditional VQA task, the *pure visual-only* VoQA task, and the VoQA task with Light Prompt. Each result is averaged over four concatenation directions. All results are reported in accuracy (%). Although the image content and embedding questions are presented separately, which helps maintain visual consistency, the model’s performance still drops significantly compared to the traditional VQA setup, highlighting the inherent difficulties of VoQA.

Model	Settings	VQAv2	GQA	POPE	TextVQA	SQA	Avg.
InternVL2.5-1B	Traditional VQA	76.3	56.9	89.9	72.0	96.7	78.4
	VoQA zero-shot	4.1	3.0	54.9	15.3	4.1	16.3
	Light Prompt	<u>35.8</u>	<u>20.1</u>	<u>65.8</u>	<u>43.7</u>	<u>17.5</u>	<u>36.6</u>
DeepSeek-VL2-Tiny	Traditional VQA	83.5	62.7	88.6	79.2	94.3	81.7
	VoQA zero-shot	<u>9.3</u>	<u>4.5</u>	<u>71.1</u>	<u>31.2</u>	<u>40.3</u>	<u>31.3</u>
	Light Prompt	8.3	4.0	54.0	4.3	6.7	15.5
Qwen2.5-VL-3B-Instruct	Traditional VQA	82.0	60.1	88.2	79.0	84.9	78.8
	VoQA zero-shot	<u>70.7</u>	<u>47.9</u>	<u>81.2</u>	<u>65.4</u>	<u>69.2</u>	<u>66.9</u>
	Light Prompt	25.1	15.9	80.9	38.6	32.8	38.6
BLIP-3 (4B)	Traditional VQA	81.7	61.6	87.0	71.0	89.8	78.2
	VoQA zero-shot	10.1	5.8	53.6	<u>16.4</u>	<u>26.2</u>	<u>22.4</u>
	Light Prompt	<u>11.7</u>	<u>8.6</u>	<u>61.5</u>	15.6	6.3	20.8

Table A2. Evaluation results of *concatenation method with resizing*. Experiment settings are identical to those in Table A1, except for the difference in the concatenation method used during data construction. All results are reported in accuracy (%). Despite the simpler setup compared to watermark rendering, all models still show a significant performance drop relative to traditional VQA.

Model	Settings	VQAv2	GQA	POPE	TextVQA	SQA	Avg.
InternVL2.5-1B	Traditional VQA	76.3	56.9	89.9	72.0	96.7	78.4
	VoQA zero-shot	4.8	2.7	54.6	14.4	5.8	16.5
	Light Prompt	<u>35.2</u>	<u>21.0</u>	<u>68.2</u>	<u>42.9</u>	<u>16.0</u>	<u>36.6</u>
DeepSeek-VL2-Tiny	Traditional VQA	83.5	62.7	88.6	79.2	94.3	81.7
	VoQA zero-shot	11.0	5.6	<u>69.9</u>	<u>36.4</u>	<u>40.7</u>	<u>32.7</u>
	Light Prompt	<u>11.9</u>	<u>6.6</u>	57.4	3.9	6.8	17.3
Qwen2.5-VL-3B-Instruct	Traditional VQA	82.0	60.1	88.2	79.0	84.9	78.8
	VoQA zero-shot	<u>69.2</u>	<u>45.4</u>	<u>80.6</u>	<u>60.9</u>	<u>72.7</u>	<u>65.8</u>
	Light Prompt	24.1	15.2	80.2	30.4	47.6	39.5
BLIP-3 (4B)	Traditional VQA	81.7	61.6	87.0	71.0	89.8	78.2
	VoQA zero-shot	10.6	7.6	53.6	<u>17.7</u>	<u>26.6</u>	<u>23.2</u>
	Light Prompt	<u>15.3</u>	<u>10.3</u>	<u>55.6</u>	16.4	7.9	21.1

B. LVLM Evaluation Settings and Results

B.1. Examples of Designed Prompts

Light Prompt. We explore a series of carefully designed light prompts to guide the model in performing the VoQA task under a zero-shot setting. By analyzing the impact of different key phrases in the prompts, we identify the most effective formulation. Table A3 summarizes the various prompt designs and highlights their key instructional components.

Short Workflow Prompt. This prompt guides the model through image analysis, ensuring it extracts the question, understands its context, and provides a precise answer based on visual cues. The placeholders `<bbox>`, `<top-left-location>`, `<bottom-right-location>`, `<picture-width>` and `<picture-height>` are replaced by the actual question location and image dimensions, derived from the dataset generation process. The model is required to output the result in a strict JSON format, ensuring consistency and enabling automated processing. Prompt content is as follows:

Task Definition:

You will receive an image with a watermark question. Your task is to:

1. Detect and extract the full question text.
2. Locate the question bounding box `<bbox>` (top-left `<top-left-location>`, bottom-right `<bottom-right-location>`) in the `<picture-width>x<picture-height>` image.
3. Understand the question type.
4. Answer accurately based only on the visual content.

Output Format (strict JSON):

```
{
  "Detected Question": "<recognized question text>",
  "Answer": "<concise answer based on the image>",
  "Reasoning": "<brief explanation of how the answer was derived>"
}
```

Important:

- Ensure full and coherent question extraction.
- Base the answer strictly on visual evidence.

- If uncertain or unclear, output "Unknown".
- Do not add commentary outside the JSON.

Now, analyze the image, detect the question and its location, and output the result in the required JSON format.

Long Workflow Prompt. The Long Workflow Prompt offers a more detailed step-by-step process compared to the Short Workflow Prompt. It includes tasks like watermark question detection, visual information extraction, and answer generation. However, both workflow prompts share the same setup regarding placeholders, such as `<bbox>`, `<top-left-location>`, `<bottom-right-location>`, `<picture-width>` and `<picture-height>`, which are replaced by the actual question location and image dimensions based on the dataset generation process.

While the Short Workflow Prompt is more concise, the Long Workflow Prompt provides more structured guidance, ensuring that the model follows a clear workflow for accurate answers based solely on the visual content. Prompt content is as follows:

(1) Task Definition:

You will receive an image containing a watermark-embedded question. Your task is to:

1. Detect the full question text embedded as a watermark in the image.
2. Understand the meaning of the question.
3. Answer the question based solely on the visual content of the image.
4. Provide an accurate and relevant answer.

(2) Workflow Steps:

Step 1: Watermark Question Detection

- Scan the image for textual content, including semi-transparent overlays and repeated patterns.
- Extract the complete question sentence.
- Tip: The question is located at bounding box `<bbox>`, from top-left `<top-left-location>` to bottom-right `<bottom-right-location>` in a `<picture-width>x<picture-height>` image.

Table A3. Light prompt design variants for zero-shot VoQA. Each prompt aims to instruct the model to first locate the embedded question in the image and then generate an answer based solely on the visual content. Key phrases used in the prompts are listed to highlight their instructional focus. The first prompt is the one ultimately adopted for our light prompt evaluations.

Designed prompts	Key phrases
There is a question in this image, you need to find the question and answer the question based on the visual information of the entire image. (Ours)	find the question, answer the question, the visual information
There is a question in this image. You need to find the question and answer the question based on the visual information of the entire image. Please do not repeat the question and answer it directly.	find the question, answer the question, the visual information, do not repeat the question, answer it directly
Please answer the question in the image directly, and do not repeat the question. You need to find the answer based on the visual information of the entire image.	answer the question directly, do not repeat the question, the visual information
Please answer the question in the image directly, and do not repeat the question.	answer the question, do not repeat the question
Please find the question in the image, and answer it directly based on the whole image.	find the question, answer it directly
Please find the question in the image, and answer it directly based on the whole image. Do not repeat the question.	find the question, answer it directly, do not repeat the question
Please find the question in the image and answer it based on the image. Do not repeat the question you find in your answers.	find the question, answer it, do not repeat the question

Step 2: Visual Information Extraction

- Analyze the image to find information relevant to the question.
- Focus on object recognition, counting, attribute description, spatial relations, scene understanding, text recognition, or reasoning as needed.

Step 3: Answer Generation

- Provide an accurate, concise answer based solely on visual evidence.
- Give a brief explanation of how the answer was derived.
- If the answer cannot be determined, state it honestly.

(3) Output Format (JSON):

Strictly return a valid JSON object as shown below:

```
{
  "Detected Question": "<recognized question text>",
  "Answer": "<concise answer based on the image>",
  "Reasoning": "<brief explanation of how the answer was derived>"
}
```

Example:

```
{
  "Detected Question": "What is the brand of this camera?",
  "Answer": "Canon",
  "Reasoning": "The text 'Canon' is clearly visible on the camera body."
}
```

(4) Important Notes:

1. Ensure question detection and answer are grounded in the image.
2. Provide the full, coherent question text.
3. Base the answer strictly on visual evidence.
4. Be concise and clear.
5. State uncertainty clearly if the image lacks information.

Now, process the input image and execute the VoQA task following the above workflow.

B.2. Response Filtering Method

To fairly evaluate model performance on the VoQA Benchmark, we apply standardized response filtering, since models may produce answers in varying formats, while the benchmark datasets require concise responses, typically a single letter, word, or phrase.

For zero-shot and few-shot (see section B.5) mod-

Table A4. Performance comparison between traditional VQA and VoQA under several zero-shot settings for VoQA. The evaluation includes six open-source models and one closed-source model. All results are reported in accuracy (%).

Model	Benchmark	Setting	VQAv2	GQA	POPE	TextVQA	SQA	Avg.
InternVL3-1B	Traditional VQA	/	71.8	52.1	88.8	71.8	95.4	76.0
	VoQA	pure zero-shot	5.5	3.2	53.9	12.8	19.5	19.0
		light prompt	14.6	9.2	57.8	19.0	18.0	23.7
		short workflow prompt	13.6	7.6	53.6	26.5	23.6	25.0
		long workflow prompt	16.8	10.0	59.9	30.8	18.9	27.3
		OCR-assisted	45.3	29.4	63.3	31.5	45.7	43.0
DeepSeek-VL2-Tiny (1B)	Traditional VQA	/	83.5	62.7	88.6	79.2	94.3	81.7
	VoQA	pure zero-shot	2.0	0.2	59.3	21.7	32.0	23.1
		light prompt	7.5	3.7	57.0	6.4	18.0	18.5
		short workflow prompt	35.1	21.5	58.5	48.7	39.3	40.6
		long workflow prompt	8.3	5.2	53.1	4.9	4.5	15.2
		OCR-assisted	69.1	44.7	82.3	41.3	52.9	58.1
Qwen2.5-VL-3B-Instruct	Traditional VQA	/	82.0	60.1	88.2	79.0	84.9	78.8
	VoQA	pure zero-shot	57.4	35.0	76.3	60.7	61.2	58.1
		light prompt	40.1	26.9	71.0	37.1	56.2	46.3
		short workflow prompt	57.0	36.9	55.7	64.9	66.2	56.1
		long workflow prompt	69.9	45.7	80.0	66.6	61.2	64.7
		OCR-assisted	68.0	47.8	84.2	58.7	70.9	65.9
TinyLLaVA-3.1B	Traditional VQA	/	80.1	62.1	87.2	55.9	75.5	72.2
	VoQA	pure zero-shot	0.8	0.2	50.5	4.5	6.7	12.6
		light prompt	46.0	22.2	71.6	31.8	12.7	36.9
		short workflow prompt	41.4	25.0	65.2	28.3	11.9	34.3
		long workflow prompt	28.8	16.2	59.9	12.3	6.0	24.6
		OCR-assisted	70.0	51.8	71.4	46.3	47.9	57.5
BLIP-3 (4B)	Traditional VQA	/	81.7	61.6	87.0	71.0	89.8	78.2
	VoQA	pure zero-shot	7.3	2.7	49.7	16.2	29.1	21.0
		light prompt	11.5	13.3	58.0	11.3	18.8	22.5
		short workflow prompt	65.7	42.7	80.1	54.3	37.4	56.0
		long workflow prompt	63.9	40.3	79.6	50.7	28.6	52.6
		OCR-assisted	68.6	46.4	79.3	57.1	50.1	60.3
LLaVA-v1.5-7B	Traditional VQA	/	78.5	62.0	85.9	46.1	73.7	69.2
	VoQA	pure zero-shot	0.2	0	50.5	3.5	1.5	11.1
		light prompt	13.7	9.3	55.0	2.9	4.1	17.0
		short workflow prompt	36.4	18.4	60.3	22.6	6.2	28.8
		long workflow prompt	38.1	18.4	62.1	21.5	3.6	28.7
		OCR-assisted	71.9	55.1	74.7	44.2	39.0	57.0
Doubao-1.5-thinking-vision-pro	Traditional VQA	/	82.7	61.7	89.1	82.1	96.9	82.5
	VoQA	pure zero-shot	75.7	49.1	86.2	78.1	77.6	73.3
		light prompt	77.0	49.7	86.2	78.2	77.6	73.8
		short workflow prompt	78.3	52.5	86.6	79.0	82.3	75.8
		long workflow prompt	78.9	53.3	86.5	79.5	81.2	75.9
		OCR-assisted	79.5	55.1	87.7	78.8	86.0	77.4

els, we adopt multiple parsing strategies to extract the actual answer from the response. These include formats such as The answer is [Actual Answer] or Answer: [Actual Answer]. Dataset-specific rules are also considered. For instance, POPE only evaluates whether the response contains no or not, so when the exact answer location is uncertain, we retain the full sentence to preserve contextual clues.

For experiments using JSON-formatted outputs (including long/short workflow prompt and few-shot settings), we first check whether the model’s response is in valid JSON

format. If so, we parse the JSON and extract the value of the ANSWER field as the candidate answer; otherwise, we retain the original output. Finally, we apply the multi-strategy filtering approach described above to obtain the final answer.

For fine-tuned models, filtering strategies vary depending on the fine-tuning approach. For all models fine-tuned using Baseline-SFT, we retain the original output. For QRA-SFT, we extract the segment following the last occurrence of a role token as the answer. Specifically for QA-SFT, where responses often follow a *Question + Answer* format, we directly use the last sentence as the final pre-

Table A5. QAA and ACC Results under short workflow settings.

Model	Metric	GQA	POPE	TextVQA	SQA	Avg.
InternVL (1B)	QAA(Correct)	92.4	87.4	92.6	66.6	84.7
	QAA(Incorrect)	83.8	87.2	76.1	55.4	75.6
	ACC	7.6	53.6	26.5	23.6	27.8
DeepSeek (1B)	QAA(Correct)	99.2	99.3	98.6	80.5	94.4
	QAA(Incorrect)	98.9	99.3	95.2	69.4	90.7
	ACC	21.5	58.5	48.7	39.3	42.0
Qwen (3B)	QAA(Correct)	59.4	61.1	79.0	53.8	63.3
	QAA(Incorrect)	59.1	57.5	78.3	47.1	60.5
	ACC	36.9	55.7	64.9	66.2	55.9
TinyLLaVA (3.1B)	QAA(Correct)	34.9	52.8	21.2	31.2	35.0
	QAA(Incorrect)	24.8	45.5	13.2	16.2	24.9
	ACC	25.0	65.2	28.3	11.9	32.6
BLIP (4B)	QAA(Correct)	86.4	93.2	86.3	73.1	84.7
	QAA(Incorrect)	79.1	91.6	64.4	55.3	72.6
	ACC	42.7	80.1	54.3	37.4	53.6
LLaVA (7B)	QAA(Correct)	38.7	36.0	37.5	40.0	38.0
	QAA(Incorrect)	35.9	36.7	34.2	26.3	33.3
	ACC	18.4	60.3	22.6	6.2	26.9
Doubao	QAA(Correct)	94.3	98.1	99.5	57.6	87.4
	QAA(Incorrect)	93.6	97.3	98.9	50.9	85.2
	ACC	52.5	86.6	79.0	82.8	75.2

dicted answer.

B.3. Zero-shot Evaluation Results

Detailed ACC Results for Zero-shot Evaluation. The complete results of the traditional VQA benchmarks and the VoQA zero-shot evaluation across different settings are summarized in Table A4. All sub-datasets show a clear and consistent performance degradation on VoQA, highlighting the increased complexity and distinct challenges posed by the VoQA task.

QAA and ACC under Workflow-based Prompt-Guided Evaluation. As shown in Table A5 and Table A6, across most sub-tasks, models exhibit higher QAA on correctly answered samples than on incorrect ones, indicating that recognizing the embedded question is a prerequisite for reliable reasoning. Moreover, in the majority of sub-tasks, higher QAA correlates with higher ACC, further reinforcing the importance of accurate question understanding.

Additional Results on Closed-source Models. We have included GPT-4o and Claude Sonnet 4.5 results (Table A7), in addition to Doubao-1.5-Thinking-Vision-Pro in the main paper. Across all models, we consistently observe a non-negligible VQA–VoQA gap.

Effects of Model Scale and Architecture. We performed additional controlled comparisons to assess the impact of model scale and architecture on VoQA performance. Experimental results (see Table A8) show that the VQA–VoQA gap consistently persists, indicating that the degradation

Table A6. QAA and ACC Results under long workflow settings.

Model	Metric	GQA	POPE	TextVQA	SQA	Avg.
InternVL (1B)	QAA(Correct)	89.0	82.9	89.8	65.4	81.8
	QAA(Incorrect)	79.2	80.7	71.7	45.9	69.4
	ACC	10.0	59.9	30.8	18.9	29.9
DeepSeek (1B)	QAA(Correct)	57.0	4.5	2.2	0.9	16.1
	QAA(Incorrect)	3.9	4.0	0.4	1.8	2.5
	ACC	5.2	53.1	4.9	4.5	16.9
Qwen (3B)	QAA(Correct)	66.2	63.8	63.9	55.0	62.2
	QAA(Incorrect)	64.6	64.0	63.7	44.7	59.3
	ACC	45.7	80.0	66.6	61.2	63.4
TinyLLaVA (3.1B)	QAA(Correct)	4.2	3.9	6.0	0	3.5
	QAA(Incorrect)	2.8	2.6	4.0	1.1	2.6
	ACC	16.2	59.9	12.3	6.0	23.6
BLIP (4B)	QAA(Correct)	81.0	89.6	64.5	68.9	76.0
	QAA(Incorrect)	65.4	86.9	45.8	44.1	60.5
	ACC	40.3	79.6	50.7	28.6	49.8
LLaVA (7B)	QAA(Correct)	37.4	34.6	35.0	34.5	35.4
	QAA(Incorrect)	33.2	32.6	31.5	22.6	30.0
	ACC	18.4	62.1	21.5	3.6	26.4
Doubao	QAA(Correct)	94.5	97.7	98.9	63.8	88.7
	QAA(Incorrect)	93.4	97.1	98.4	50.1	84.8
	ACC	53.3	86.5	79.5	81.2	75.1

Table A7. Accuracy (%) comparison between VQA and VoQA zero-shot evaluation on closed-source models.

Model	Setting	VQAv2	GQA	POPE	TextVQA	SQA	Avg.
Doubao	VQA	82.7	61.7	89.1	82.1	96.9	82.5
	VoQA	75.7	49.1	86.2	78.1	77.6	73.3
GPT	VQA	72.5	51.3	86.5	75.2	90.7	75.2
	VoQA	70.5	48.3	85.7	69.4	78.3	70.4
Claude	VQA	70.6	48.3	84.7	71.5	68.7	68.8
	VoQA	36.6	27.1	79.1	49.1	64.7	51.3

cannot be explained by model capacity or architecture alone.

B.4. Additional Representative Zero-shot Examples

To further illustrate the challenges of the VoQA task and the diverse behaviors of various models, we provide three additional representative examples (Table A9 to A11). These examples supplement Table 1 in the main paper and cover different scene types and question complexities within the *pure visual-only* zero-shot setting. As observed in these cases, models often struggle with basic visual reasoning when the question is embedded within the visual modality.

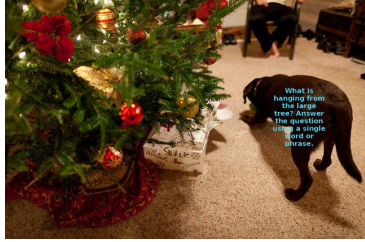
B.5. Few-shot Evaluation

To help models learn the task pattern from examples, we incorporate few-shot settings into our prompt-engineering experiments. We evaluate three open-source LVLMS (InternVL3-1B, Qwen2.5-VL-3B-Instruct, and BLIP-3 (4B)) in a few-shot setting, providing $k \in \{1, 2, 4, 8\}$ randomly sampled examples from the VoQA training dataset. Each example follows a *JSON format* containing the question and answer.

Table A8. Performance comparison across various model scales and architectures. All results are reported in accuracy (%).

Model	Benchmark	Setting	VQAv2	GQA	POPE	TextVQA	SQA	Avg.
InternVL3-1B	Traditional VQA	/	71.8	52.1	88.8	71.8	95.4	76.0
	VoQA	pure zero-shot	5.5	3.2	53.9	12.8	19.5	19.0
		light prompt	14.6	9.2	57.8	19.0	18.0	23.7
		short workflow prompt	13.6	7.6	53.6	26.5	23.6	25.0
		long workflow prompt	16.8	10.0	59.9	30.8	18.9	27.3
		OCR-assisted	45.3	29.4	63.3	31.5	45.7	43.0
InternVL3-2B	Traditional VQA	/	74.5	54.0	88.8	75.3	96.7	77.9
	VoQA	pure zero-shot	11.4	6.5	57.7	16.1	34.8	25.3
		light prompt	19.0	11.6	62.1	22.1	29.1	28.8
		short workflow prompt	43.0	25.2	62.6	42.4	32.3	41.1
		long workflow prompt	49.8	30.6	68.5	47.8	32.2	55.3
		OCR-assisted	59.8	37.5	76.6	42.6	59.7	55.3
TinyLLaVA-1B	Traditional VQA	/	72.3	55.8	86.6	42.1	62.3	63.8
	VoQA	pure zero-shot	0.3	0	50.7	2.4	1.9	11.1
		light prompt	17.0	5.3	55.2	6.3	1.0	17.0
		short workflow prompt	23.4	14.8	51.7	12.7	1.1	20.8
		long workflow prompt	1.5	0.5	50.5	0.2	0.6	10.7
		OCR-assisted	54.9	42.8	54.6	29.1	34.8	43.2
TinyLLaVA-3.1B	Traditional VQA	/	80.1	62.1	87.2	55.9	75.5	72.2
	VoQA	pure zero-shot	0.8	0.2	50.5	4.5	6.7	12.6
		light prompt	46.0	22.2	71.6	31.8	12.7	36.9
		short workflow prompt	41.4	25.0	65.2	28.3	11.9	34.3
		long workflow prompt	28.8	16.2	59.9	12.3	6.0	24.6
		OCR-assisted	70.0	51.8	71.4	46.3	47.9	57.5

Table A9. A supplementary case study based on a sample from the GQA dataset, illustrating the *pure visual-only* zero-shot VoQA setting. The question “What is hanging from the large tree? Answer the question using a single word or phrase.” is asked about a decorated Christmas tree. The models correspond to those introduced in Section 3.3.1. Note: ellipses in some model outputs are added for brevity and do not affect the classification of behavior type.

Input	Model	Response	Behavior Type
	TinyLLaVA (3.1B)	In the image, a black dog ... In the background, there is a Christmas tree adorned with red and gold ornaments ...	Question-unaware image captioning
	LLaVA (7B)	The image features a dog standing in front of a Christmas tree ... The tree is adorned with various ornaments ...	Question-unaware image captioning
	InternVL (1B)	The question hanging from the tree is "What is hanging from the large tree?" The answer should use a single word or phrase.	Repeating the question
	Qwen (3B)	What is hanging from the large tree? Answer using a single word or phrase.	Repeating the question
	BLIP (4B)	Answer is hanging from the large tree.	Answering incorrectly
	DeepSeek (1B)	The large tree has ornaments hanging from it.	Answering correctly
	Doubao	Ornaments	Answering correctly

Few-shot settings. We randomly sampled 10,000 instances from the VoQA training set to construct a demonstration pool. To ensure balanced coverage, the sampling process was stratified according to the proportion of samples from different sub-tasks in the training data. For each of the 134k evaluation samples, we randomly selected 1, 2, 4, or 8 examples from this pool as few-shot demonstrations. Once selected, the examples were fixed to ensure consistent comparison across models and reproducibility of results.

Each selected example’s image, embedded question text,

and corresponding answer were inserted into the following prompt template (shown below using the 2-shot setting as an example). In this template, the placeholders <Ground Truth Question in Example x> and <Ground Truth Answer in Example x> are replaced with the actual question and answer from the sampled examples.

Each image contains a question written inside it. Your task is to extract the question from the image and answer it

Table A10. Another illustrative example from the VoQA benchmark, featuring a sample from the ScienceQA (SQA) dataset. The image contains a geographical map with the visually embedded question: “What is the name of the colony shown? A. Maryland B. New Hampshire C. Rhode Island D. Vermont Answer with the option’s letter from the given choices directly.” The models correspond to those introduced in Section 3.3.1. Note: ellipses in some model outputs are added for brevity and do not affect the classification of behavior type.

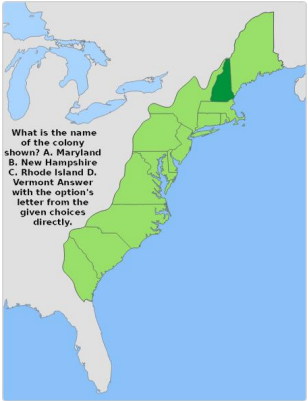

Input	Model	Response	Behavior Type
	LLaVA (7B)	The image is a political map of the United States, showing the states and their borders ... The states are colored green, and the map is accompanied by a text description of the states and their locations.	Question-unaware image captioning
	TinyLLaVA (3.1B)	The image presents a detailed map of the United States ... The question reads, “What is the name of the colony shown A, Maryland, C, Rhode Island, D, Vermont Answer in the option at the bottom of the map given choices directly.” This suggests that the image might be part of a quiz or educational material, testing the viewer’s knowledge about the historical context of these states ...	Question-aware image captioning
	InternVL (1B)	The name of the colony shown is **D. Vermont**.	Answering incorrectly
	DeepSeek (1B)	The name of the colony shown is D. Vermont.	Answering incorrectly
	BLIP (4B)	Answer: A	Answering incorrectly
	Qwen (3B)	B. New Hampshire	Answering correctly
	Doubao	B	Answering correctly

Table A11. An additional example of the VoQA benchmark, featuring a street scene from the TextVQA dataset. The image depicts a convenience store exterior with a bicycle, where the visually embedded question is: “How much for a can of skoal? Answer the question using a single word or phrase.” The task requires identifying the specific price tag for “Skoal” displayed on the store door. The models correspond to those introduced in Section 3.3.1. Note: ellipses in some model outputs are added for brevity and do not affect the classification of behavior type.

Input	Model	Response	Behavior Type
	LLaVA (7B)	The image features a bicycle parked outside a store, possibly a liquor store ...	Question-unaware image captioning
	TinyLLaVA (3.1B)	The image captures a scene outside a convenience store ... The bicycle is also in front of a trash can, which has a sign that reads “How much for a can of Skoal? Answer the question using a single phrase or phrase.” ...	Question-aware image captioning
	InternVL (1B)	The question in the image is: “How much for a can of skool? Answer the question using a single word or phrase.” The bicycle ...	Question-aware image captioning
	DeepSeek (1B)	How much for a can of skoal? Answer the question using a single word or phrase.	Repeating the question
	Qwen (3B)	4.52	Answering incorrectly
	BLIP (4B)	Answer is 2.	Answering incorrectly
	Doubao	3.82	Answering correctly

accurately.

Example 1:

Input: <image>

Output: {"The question in the image": "<Ground Truth Question in Example 1>", "Answer": "<Ground Truth Answer in Example 1>"}

Example 2:

Input: <image>

Output: {"The question in the image": "<Ground Truth Question in Example 2>",

"Answer": "<Ground Truth Answer in Example 2>"}

Now answer the next one:

Input: <image>

Output:

Result analysis. As shown in Table A12, the few-shot evaluation results on the VoQA benchmark show that few-shot examples help models better grasp the task pattern, yielding improvements over the zero-shot setting but still falling far short of performance on traditional VQA. The detailed results of each model on each sub-task can be found

Table A12. Comparison of model performance on traditional VQA, VoQA *pure visual-only* zero-shot, and few-shot settings across the VoQA benchmark. All results are reported in accuracy (%). For few-shot results, each model shows the setting with the highest average accuracy.

Model	Evaluation Setting	VQAv2	GQA	POPE	TextVQA	SQA	Avg.
InternVL3-1B	Traditional VQA	71.8	52.1	88.8	71.8	95.4	76.0
	VoQA zero-shot	5.6	3.2	53.9	12.8	19.5	19.0
	VoQA few-shot	<u>20.2</u>	<u>12.8</u>	<u>59.5</u>	<u>26.6</u>	<u>26.0</u>	<u>29.0</u>
Qwen2.5-VL-3B-Instruct	Traditional VQA	82.0	60.1	88.2	79.0	84.9	78.8
	VoQA zero-shot	57.4	35.0	76.3	60.7	61.2	58.1
	VoQA few-shot	<u>64.1</u>	<u>41.4</u>	<u>76.6</u>	<u>64.4</u>	<u>55.7</u>	<u>60.4</u>
BLIP-3 (4B)	Traditional VQA	81.7	61.6	87.0	71.0	89.7	78.2
	VoQA zero-shot	7.3	2.7	49.7	16.2	29.1	21.0
	VoQA few-shot	<u>54.0</u>	<u>35.8</u>	<u>75.3</u>	<u>51.3</u>	<u>28.2</u>	<u>48.9</u>

Table A13. *InternVL3-1B* few-shot Answer Accuracy (ACC) and Question Alignment Accuracy (QAA) results. *Correct* and *Incorrect* indicate correctly and incorrectly answered samples, respectively. All results are reported in accuracy (%).

Setting	Metric	VQAv2	GQA	POPE	TextVQA	SQA	Avg.
1-shot	QAA(Correct)	/	56.2	45.1	50.5	25.0	44.2
	QAA(Incorrect)	/	43.5	40.1	31.9	18.2	33.4
	ACC	18.6	11.0	58.3	23.7	23.5	27.0
2-shot	QAA(Correct)	/	53.5	42.4	51.2	31.0	44.5
	QAA(Incorrect)	/	45.3	39.7	34.9	18.1	34.5
	ACC	20.2	12.8	59.5	26.6	26.0	29.0
4-shot	QAA(Correct)	/	45.3	33.9	40.8	27.9	37.0
	QAA(Incorrect)	/	34.3	30.4	27.9	17.5	27.5
	ACC	20.1	11.9	57.7	24.0	19.6	26.7
8-shot	QAA(Correct)	/	38.3	29.8	31.9	15.7	28.9
	QAA(Incorrect)	/	28.9	27.0	22.0	13.8	22.9
	ACC	19.2	11.8	57.1	21.0	14.5	24.7

Table A14. *Qwen2.5-VL-3B-Instruct* few-shot Answer Accuracy (ACC) and Question Alignment Accuracy (QAA) results.

Setting	Metric	VQAv2	GQA	POPE	TextVQA	SQA	Avg.
1-shot	QAA(Correct)	/	67.3	63.7	71.3	42.3	61.1
	QAA(Incorrect)	/	63.9	61.8	69.6	30.5	56.4
	ACC	64.1	41.4	76.6	64.4	55.7	60.4
2-shot	QAA(Correct)	/	69.9	68.4	73.6	42.7	63.6
	QAA(Incorrect)	/	64.4	65.6	71.0	30.9	58.0
	ACC	62.2	40.2	73.8	63.1	50.4	57.9
4-shot	QAA(Correct)	/	66.3	67.7	69.7	44.0	61.9
	QAA(Incorrect)	/	58.6	63.6	66.2	29.1	54.4
	ACC	59.5	38.5	71.1	61.7	46.7	55.5
8-shot	QAA(Correct)	/	57.1	49.1	61.1	40.5	52.0
	QAA(Incorrect)	/	42.4	40.6	49.1	24.1	39.0
	ACC	48.0	30.6	62.6	54.6	30.9	45.3

in Tables A13, A14, and A15.

Consistent with our workflow analysis, all models exhibit higher Question Alignment Accuracy (QAA) for correctly answered samples across all few-shot configurations, underscoring that accurate question recognition remains the main factor for reliable reasoning. However, models still struggle to consistently identify and interpret embedded questions, leading to only modest overall gains.

Table A15. *BLIP-3 (4B)* few-shot Answer Accuracy (ACC) and Question Alignment Accuracy (QAA) results.

Setting	Metric	VQAv2	GQA	POPE	TextVQA	SQA	Avg.
1-shot	QAA(Correct)	/	66.8	63.1	57.6	25.6	53.3
	QAA(Incorrect)	/	53.2	58.5	48.5	13.7	43.5
	ACC	53.6	35.4	77.3	51.1	26.6	48.8
2-shot	QAA(Correct)	/	67.5	63.5	55.3	19.8	51.5
	QAA(Incorrect)	/	52.2	58.4	47.0	13.2	42.7
	ACC	54.0	35.8	75.3	51.3	28.2	48.9
4-shot	QAA(Correct)	/	56.5	40.7	38.8	4.5	35.1
	QAA(Incorrect)	/	27.3	26.2	22.9	4.7	20.3
	ACC	37.7	26.2	64.3	42.4	26.3	39.4
8-shot	QAA(Correct)	/	37.6	18.0	20.2	1.0	19.2
	QAA(Incorrect)	/	8.1	8.5	6.9	0.4	6.0
	ACC	23.3	14.3	57.0	28.7	29.2	30.5

Table A16. Performance comparison between traditional VQA and VoQA under several zero-shot settings for VoQA. The evaluation covers the official instruction-tuned versions of the three models introduced in Section 4.1. All results are reported in accuracy (%).

Model	Benchmark	Setting	VQAv2	GQA	POPE	TextVQA	SQA	Avg.
TinyLLaVA-1B	Traditional VQA	/	72.3	55.8	86.6	42.1	62.3	63.8
	VoQA	pure zero-shot	0.3	0	50.7	2.4	1.9	11.1
		light prompt	17.0	5.3	55.2	6.3	1.0	17.0
		short workflow prompt	23.4	14.8	51.7	12.7	1.1	20.8
		long workflow prompt	1.5	0.5	50.5	0.2	0.6	10.7
		OCR-assisted	54.9	42.8	54.6	29.1	34.8	43.2
InternVL3-1B	Traditional VQA	/	71.8	52.1	88.8	71.8	95.4	76.0
	VoQA	pure zero-shot	5.5	3.2	53.9	12.8	19.5	19.0
		light prompt	14.6	9.2	57.8	19.0	18.0	23.7
		short workflow prompt	13.6	7.6	53.6	26.5	23.6	25.0
		long workflow prompt	16.8	10.0	59.9	30.8	18.9	27.3
		OCR-assisted	45.3	29.4	63.3	31.5	45.7	43.0
Qwen2-VL-2B-Instruct	Traditional VQA	/	81.7	60.4	88.1	79.1	64.2	74.7
	VoQA	pure zero-shot	62.5	38.8	79.3	4.5	62.8	33.2
		light prompt	11.4	7.7	56.9	27.6	14.6	23.6
		short workflow prompt	32.5	19.5	49.4	48.9	21.2	34.3
		long workflow prompt	61.4	36.9	72.2	60	10.6	48.2
		OCR-assisted	70.3	46.8	82.0	59.7	54.7	62.7

Table A17. ACC comparison of three fine-tuned models on the VoQA benchmark. All models share the same pre-trained backbone, and results are reported as accuracy (%).

Base Model	Settings	VQAv2	GQA	POPE	TextVQA	SQA	Avg.
TinyLLaVA (1B)	VoQA zero-shot	0.2	0.1	50.4	0	0	10.2
	Baseline-SFT	63.4	45.3	75.5	32.5	24.4	48.2
InternVL (1B)	VoQA zero-shot	5.5	3.2	53.9	12.8	19.5	19.0
	Baseline-SFT	67.4	50.1	82.1	35.2	38.6	54.7
Qwen (2B)	VoQA zero-shot	11.6	6.9	55.5	25.3	21.3	24.1
	Baseline-SFT	75.9	56.9	85.9	68.5	54.3	68.3

Table A18. Evaluation of models fine-tuned on either VQA or VoQA using different strategies, all tested on traditional VQA benchmarks. Results are reported in accuracy (%).

Base Model	Training Setting	VQAv2	GQA	POPE	TextVQA	SQA	Avg.
TinyLLaVA (1B)	VQA SFT	<u>72.3</u>	55.8	86.6	<u>42.1</u>	62.3	63.8
	VoQA QA-SFT	65.6	48.9	85.8	34.3	16.3	50.2
	VoQA QRA-SFT	72.6	<u>54.6</u>	87.1	42.2	<u>60.3</u>	<u>63.4</u>
InternVL (1B)	VQA SFT	76.8	56.6	88.9	69.0	88.0	75.8
	VoQA QA-SFT	74.6	50.9	85.7	62.5	85.3	71.8
	VoQA QRA-SFT	<u>76.3</u>	<u>56.1</u>	88.0	<u>64.5</u>	<u>85.4</u>	<u>74.1</u>
Qwen (2B)	VQA SFT	80.4	60.7	87.8	76.0	82.6	77.5
	VoQA QA-SFT	62.1	50.7	88.3	46.1	38.9	57.2
	VoQA QRA-SFT	<u>70.6</u>	<u>58.6</u>	85.9	<u>66.7</u>	<u>77.7</u>	<u>71.9</u>

C. Question-Alignment Fine-Tuning

C.1. Evaluation of Pre-fine-tuned Models

This section presents the full zero-shot results for the base models prior to our fine-tuning stage. All evaluations follow the same experimental configurations as detailed in Section 3.3.1. The models evaluated in Table A16 are the official instruction-tuned versions of the models that we subsequently fine-tune in Section 4.1.

C.2. Fine-Tuning Settings

All models are fine-tuned for one epoch with consistent hyperparameters, and all experiments are conducted on $8 \times$ NVIDIA A100 (40GB) GPUs. TinyLLaVA (1B) is trained on both the connector and language model using AdamW[3], a cosine scheduler[2] with 0.03 warmup, a learning rate of 2×10^{-5} , and a batch size of 128. For InternVL (1B) and Qwen (2B), only the language model is updated using LoRA[1] (rank 8), AdamW, and a cosine scheduler with 0.1 warmup; the learning rates are 1×10^{-5} and 1.4×10^{-5} , respectively, with a batch size of 64.

C.3. Case Study of Fine-Tuning Strategies

Figure A2 presents representative examples illustrating the inference behaviors of Fine-tuning Strategies on the VoQA benchmark. Each subfigure includes the composite image input, the model’s prediction, a brief behavior analysis, and the corresponding model name. The examples cover diverse outcomes such as irrelevant responses, question repetition, and correct answers.

C.4. Fine-tuning Results on VoQA and traditional VQA Benchmarks

Complete Results of VoQA ACC. Table A17 shows that Baseline-SFT yields consistent performance gains over the zero-shot setting across all VoQA sub-tasks.

Complete Results of VQA ACC. From the results in Table A18, QRA-SFT demonstrates clear advantages over QA-SFT on most VQA sub-tasks and even surpasses VQA-SFT in several cases. These findings highlight the strength of QRA-SFT in maintaining the traditional VQA input format while effectively aligning visual questions during VoQA-oriented fine-tuning.

Complete Results of VoQA QAA. As summarized in Table A19, the average QAA of correct samples consistently exceeds that of incorrect ones across all sub-tasks, indicating that accurately detecting the embedded question is a crucial prerequisite for reliable answer generation.

Table A19. Question Alignment Accuracy (QAA, %) of models after QRA-SFT fine-tuning. *Correct* and *Incorrect* indicate averages computed over correctly and incorrectly answered samples, respectively. Higher QAA indicates better recognition of visually embedded questions.

Model	Results	GQA	POPE	TextVQA	SQA	Avg.
TinyLLaVA	Correct	95.8	98.6	97.2	87.2	94.7
	Incorrect	93.1	95.1	93.9	73.6	88.9
InternVL	Correct	96.6	98.3	97.6	89.7	95.5
	Incorrect	96.1	97.6	93.6	57.9	86.3
Qwen	Correct	97.7	98.7	98.6	92.8	96.9
	Incorrect	97.5	98.8	91.7	73.8	90.4

C.5. Influence of Role Tokens in VQA Inference Templates

Purpose. Since QA-SFT alters the input sequence format originally used in traditional VQA, we examine how the inclusion of the role token (i.e., the *ASSISTANT:* token shown in Figure 5) influences model behavior during VQA inference. Specifically, we analyze whether retaining this token helps preserve VQA performance and how its presence or absence affects QRA-SFT.

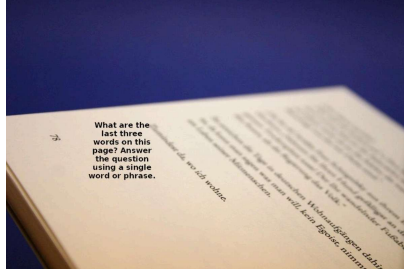
Findings. Table A20 shows that including the role token generally improves the performance of both SFT variants, suggesting that the token serves as a useful cue that prompts the model to begin generating an answer in traditional VQA settings. Nevertheless, QRA-SFT without this token remains highly competitive, indicating that the model implicitly learns to anticipate this token during VoQA fine-tuning and thus can still maintain robust question–answer alignment even without explicitly providing the structural cue.

C.6. Exploring Variants of Question-Alignment Fine-Tuning

C.6.1. Analyzing the Effect of Role Tokens in Question-Alignment Fine-Tuning

Since QA-SFT and QRA-SFT differ in both the placement of role tokens and whether these tokens appear in the supervised target, we further investigate how role-token positioning and supervision design influence model behavior.

R-QRA-SFT. As illustrated in Figure A3 (Line 3), *Role-guided Question–Role–Answer Supervised Fine-Tuning* (R-QRA-SFT) unifies the strengths of QA-SFT and QRA-SFT under a consistent structure. The input includes a role token that identifies the model as an *ASSISTANT*, serving as a cue to initiate answer generation. Meanwhile, the supervised target expands to the complete *Question–Role–Answer* sequence, making the boundary between question interpretation and answer production explicit. This design enforces



Answer: Last three words on page

Type: Repeat part of the question in the image
(Wrong answer)

(a) An example of TinyLLaVA-1B-Baseline



Answer: A bottle of whiskey is on a shelf.

Type: Answer irrelevantly in the form of a complete sentence (Wrong answer)

(b) An example of InternVL3-1B-Baseline



Answer: Toothbrush

Type: Give a correct answer directly
(Correct answer, Standard answer format for Baseline-SFT)

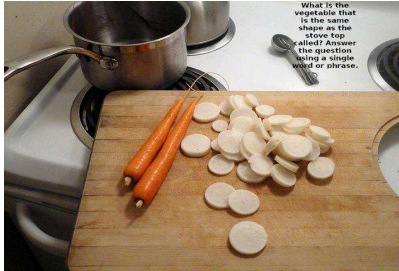
(c) An example of Qwen2-VL-2B-Baseline



Answer: Is there a car in the image? Answer the question using a single word or phrase. No

Type: Repeat the question, and then answer it
(Wrong answer)

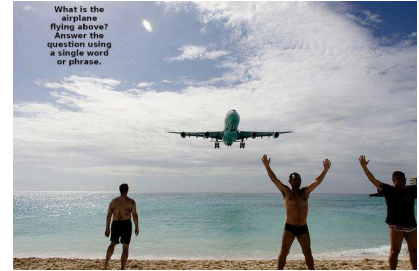
(d) An example of Qwen2-VL-2B-QA



Answer: What is the vegetable that is the same shape as the stove top stove top stove top stove top stove top stove top stove

Type: Repeat the question (Wrong answer)

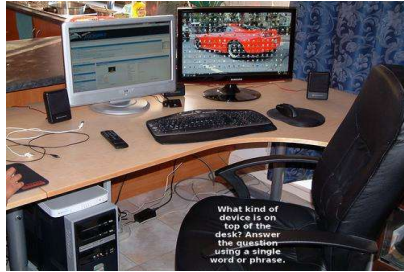
(e) An example of InternVL3-1B-QA



Answer: What is the airplane flying above? Answer the question using a single word or phrase. Ocean

Type: Repeat the question, and then answer it
(Correct answer, Standard answer format for QA-SFT)

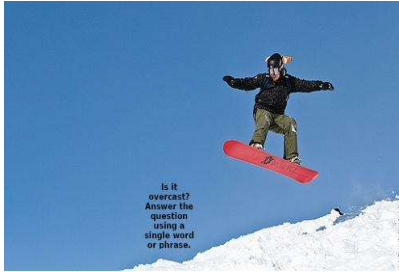
(f) An example of InternVL3-1B-QA



Answer: What kind of device is on top of the desk? Answer the question using a single word or phrase. Computer

Type: Repeat the question, and then answer it
(Wrong answer)

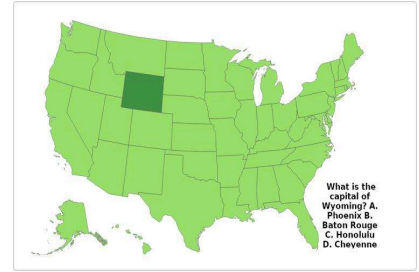
(g) An example of InternVL3-1B-QRA



Answer: Is it overcast? Answer the question using a single word or phrase. No

Type: Repeat the question, and output the role token 'nassistant/n', then answer the question
(Correct answer, Standard answer format for QRA-SFT)

(h) An example of InternVL3-1B-QRA



Answer: What is the capital of Wyoming? A. Phoenix B. Baton Rouge C. Honolulu D. Cheyenne

Type: Repeat the question, and output the role token 'nassistant/n', then answer the question
(Correct answer, Standard answer format for QRA-SFT)

(i) An example of Qwen2-VL-2B-QRA

Figure A2. Nine representative case study examples illustrating the inference behavior of Baseline-SFT, QA-SFT, and QRA-SFT on the VoQA benchmark. The fine-tuning method used by each model is indicated as a suffix (-Baseline, -QA, -QRA) next to the model name. In each subfigure, the model's predicted answer is shown in blue, while red and green indicate incorrect and correct answers, respectively, based on the *Type* field. The correctness of each sample is explicitly marked.

Table A20. Impact of including or excluding the role token during inference on traditional VQA benchmarks for both VoQA SFT variants. *w/ Role* and *w/o Role* indicate whether the role token is explicitly provided in the fixed input template at training or inference time, i.e., whether the model receives this token immediately before generating the answer. Results are reported in accuracy (%).

Model	SFT	Train	Inference	VQAv2	GQA	POPE	TextVQA	SQA	Avg.
TinyLLaVA (1B)	QA-SFT	w/ Role	w/ Role	65.6	48.9	85.8	34.3	16.3	50.2
			w/o Role	20.3	7.7	50.5	16.4	3.1	19.6
	QRA-SFT	w/o Role	w/ Role	72.6	54.6	87.1	42.2	60.3	63.4
			w/o Role	69.6	52.2	87.1	40.7	60.4	62.0
InternVL (1B)	QA-SFT	w/ Role	w/ Role	74.6	50.9	85.7	62.5	85.3	71.8
			w/o Role	12.2	4.4	50.6	3.4	48.8	23.9
	QRA-SFT	w/o Role	w/ Role	76.3	56.1	88.0	64.5	85.4	74.1
			w/o Role	74.6	54.5	85.9	63.7	81.6	72.0
Qwen (2B)	QA-SFT	w/ Role	w/ Role	62.1	50.7	88.3	46.1	38.9	57.2
			w/o Role	0	0	51.1	0	0.6	10.3
	QRA-SFT	w/o Role	w/ Role	70.6	58.6	85.9	66.7	77.7	71.9
			w/o Role	63.4	47.8	79.1	37.4	49.8	55.5

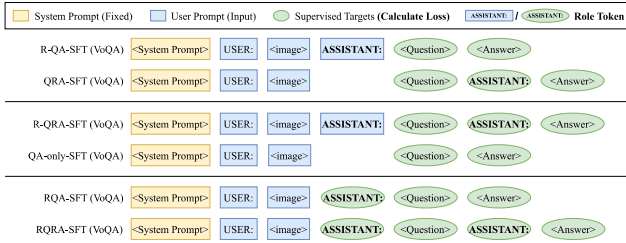


Figure A3. Comparison of six supervised fine-tuning strategies on the VoQA dataset. All methods are our proposed VoQA-specific approaches that first align the visually grounded question before generating the answer. The first two correspond to the main methods discussed in the paper: QA-SFT (denoted here as R-QA-SFT for consistency across variants) and QRA-SFT. The middle two are their variants, obtained by adding or removing the role token relative to the original methods. The last two correspond to variants of R-QA-SFT and R-QRA-SFT in which the model is required to explicitly predict the role token at the beginning of the output sequence.

structured reasoning while preserving tight visual–textual alignment.

QA-only-SFT. To isolate the effect of explicit role cues, we introduce *Question–Answer only Supervised Fine-Tuning* (QA-only-SFT, Figure A3, Line 4), where both the input and output sequences omit the *ASSISTANT* token, while the model still predicts the full question–answer pair. This variant tests whether removing role tokens weakens the model’s ability to align visually embedded questions with correct answers.

Result Analysis. Table A21 shows that QA-only-SFT performs slightly worse than R-QRA-SFT, though their results remain highly comparable. This indicates that role tokens provide only a weak segmentation cue in this setting.

When compared with QA-SFT and QRA-SFT introduced in Section 4, all four question-aligned fine-tuning strategies exhibit similar performance and consistently surpass Baseline-SFT by a substantial margin. Overall, the results confirm that precise question alignment is the primary factor driving the model’s success on the VoQA task, while the presence of role tokens plays a relatively minor role.

Table A21. Comparison between five fine-tuning strategies on the VoQA benchmark. All results are reported in accuracy (%).

Model	VoQA SFT	VQAv2	GQA	POPE	TextVQA	SQA	Avg.
TinyLLaVA (1B)	Baseline-SFT	63.4	45.3	75.5	32.5	24.4	48.2
	QA-SFT	70.0	49.8	84.1	37.4	36.9	55.6
	QRA-SFT	69.6	49.5	83.8	37.1	38.2	55.6
	R-QRA-SFT	69.7	49.6	84.0	37.4	37.3	55.6
	QA-only-SFT	69.2	48.7	83.8	37.5	36.1	55.1
InternVL (1B)	Baseline-SFT	67.4	50.1	82.1	35.2	38.6	54.7
	QA-SFT	73.2	53.0	86.6	53.7	42.5	61.8
	QRA-SFT	72.6	52.7	85.8	56.3	49.1	63.3
	R-QRA-SFT	72.5	53.2	87.3	56.2	50.3	63.9
	QA-only-SFT	69.5	51.2	86.3	53.5	50.5	62.2
Qwen (2B)	Baseline-SFT	75.9	56.9	85.9	68.5	54.3	68.3
	QA-SFT	79.2	60.1	87.6	70.8	60.6	71.7
	QRA-SFT	78.1	60.5	88.0	70.8	60.5	71.6
	R-QRA-SFT	78.0	60.5	88.1	71.9	63.9	72.5
	QA-only-SFT	79.3	60.6	87.9	72.3	58.1	71.6

C.6.2. Effect of Predicting the Initial Role Token

Two variants. To further understand how role tokens influence question-aligned fine-tuning, we analyze whether models can autonomously infer their role instead of relying on an explicit role token placed at the beginning of the output sequence, which is the default design in both R-QA-SFT and R-QRA-SFT. Building on these strategies, we extend the supervised targets to include the initial role token, yielding *Role–Question–Answer Supervised Fine-Tuning* (RQA-SFT, Figure A3, Line 5, from R-QA-SFT) and *Role–Question–Role–Answer Supervised Fine-Tuning* (RQRA-SFT, Figure A3, Line 6, from R-QRA-SFT).

Table A22. Comparison of two fine-tuning strategies on the VoQA benchmark using the *InternVL3-1B* model, evaluating their performance with and without providing the role token during inference. All results are reported in accuracy (%).

SFT	Train	Inference	VQA v2	GQA	POPE	TextVQA	SQA	Avg.
RQA	w/o Role	w/ Role	71.9	52.0	85.5	51.6	44.1	61.0
		w/o Role	37.0	22.5	68.1	29.8	25.6	36.6
RQRA	w/o Role	w/ Role	72.8	53.1	86.9	56.8	51.3	64.2
		w/o Role	5.2	4.2	52.0	6.6	30.8	19.8

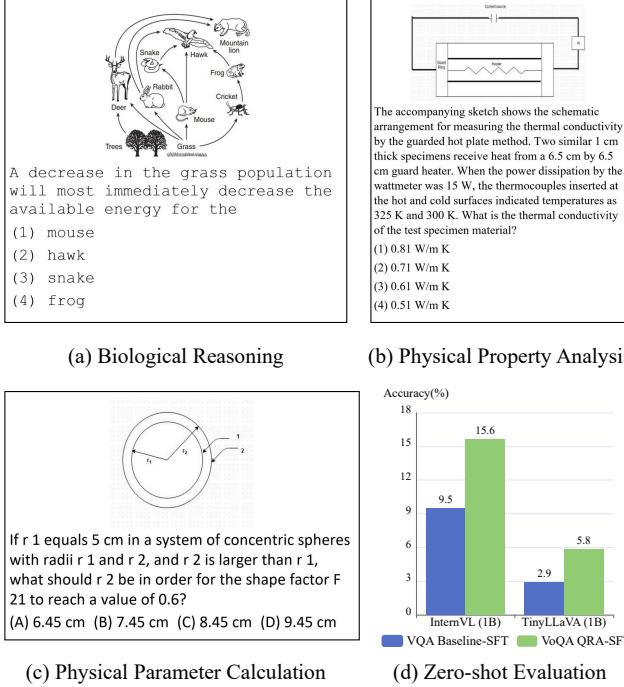


Figure A4. Zero-shot generalization to realistic scientific reasoning tasks. (a)-(c) Representative examples of our QRA-SFT models performing reasoning over complex questions with scientific diagrams. (d) Accuracy (%) comparison between VQA Baseline-SFT and our VoQA QRA-SFT across these out-of-distribution scenarios. Quantitative evaluations are conducted on a randomly sampled 10% subset of the original dataset.

Findings. Experiments on *InternVL* models (Table A22) reveal a consistent pattern. When the role token is not provided during inference, both variants show degraded performance because the model struggles to predict the token reliably. However, once the role token is supplied as an input cue, performance improves substantially. The findings demonstrate that VoQA performance is determined primarily by whether the model initiates generation with accurate alignment to the visually embedded question, while the presence or prediction of a role token has no substantive impact. In other words, successful VoQA performance depends far more on accurate question alignment at the first step than on predicting an explicit *ASSISTANT* role token.

C.7. Generalization to Real-World Scenarios

We evaluate the zero-shot generalization of our QRA-SFT models on a real-world dataset subset [4]. As illustrated by the representative results in Figure A4, our 1B-scale QRA-SFT models successfully maintain performance gains over their respective pre-fine-tuned versions. These results suggest that QRA-SFT can improve the model’s capability to handle visually-embedded questions in real-world scenarios, narrowing the gap between synthetic training and practical application.

References

- [1] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net, 2022. 11
- [2] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *ICLR (Poster)*. OpenReview.net, 2017. 11
- [3] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR (Poster)*. OpenReview.net, 2019. 11
- [4] Belal Shoer and Yova Kementchedjhieva. A simple data augmentation strategy for text-in-image scientific VQA. *CoRR*, abs/2509.20119, 2025. 14
- [5] Michael Tschannen, Basil Mustafa, and Neil Houlsby. CLIPPO: image-and-language understanding from pixels only. In *CVPR*, pages 11006–11017. IEEE, 2023. 1