VoQA: Visual-only Question Answering

Jianing An[1] Luyang Jiang[1] Jie Luo[1] Wenjun Wu[1,2] Lei Huang[1,2,✉]
[1]SKLCCSE, School of Artificial Intelligence, Beihang University, Beijing, China
[2]Hangzhou International Innovation Institute, Beihang University, Hangzhou, China

Abstract

Visual understanding requires interpreting both natural scenes and the textual information that appears within them, motivating tasks such as Visual Question Answering (VQA). However, current VQA benchmarks overlook scenarios with visually embedded questions, whereas advanced agents should be able to see the question without separate text input as humans. We introduce ***Visual-only Question Answering (VoQA)****, where both the scene and the question appear within a single image, requiring models to perceive and reason purely through vision. This setting supports more realistic visual understanding and interaction in scenarios where questions or instructions are embedded directly in the visual scene. Evaluations under pure visual-only zero-shot, prompt-guided and OCR-assisted settings show that current models exhibit a clear performance drop compared to traditional VQA. To address this, we investigate question-alignment fine-tuning strategies designed to guide models toward interpreting the visual question prior to reasoning. Leveraging the VoQA dataset together with these strategies yields robust vision-only reasoning while preserving cross-task generalization to traditional VQA, reflecting complementary visual and textual reasoning capabilities fostered through VoQA training.*

1. Introduction

Visual understanding is a core capability in both human perception and artificial intelligence systems, including comprehension of natural scenes and interpretation of visual scenarios with embedded textual cues. Research in multimodal reasoning has largely been driven by the Visual Question Answering (VQA) [2, 8] task, where a model receives an image together with an explicitly provided textual question and is required to produce an answer. This formulation has enabled systematic evaluation of visual reasoning and the development of a wide range of benchmarks [10, 18, 19, 24] and models.

Q. Where is the cat positioned in the image?

A. The cat is positioned on top of the back of the couch in the living room.

(a) Traditional VQA Task

A. The cat is positioned on top of the back of the couch in the living room.

(b) VoQA Task

Figure 1. Comparison between (a) the traditional VQA task and (b) the Visual-only Question Answering (VoQA) task. Traditional VQA provides an image and a textual question as separate inputs, whereas VoQA embeds the question directly within the image, requiring reasoning purely through visual perception.

Fueled by advances in Large Vision-Language Models (LVLMs) [1, 4, 13, 16, 22], recent models have achieved notable results on various VQA benchmarks. However, current VQA benchmarks ignore cases where questions are embedded within the visual scene, yet advanced agents should be able to directly *see* and interpret such questions without relying on separate textual input, similar to human perception. Such situations commonly occur in natural environments, digital interfaces, and many embodied AI settings. In these cases, models must detect, interpret, and reason over visually embedded text without relying on separate textual inputs. Developing the ability to handle these scenarios allows embodied AI systems, including robots, autonomous vehicles, and graphical user interface (GUI) agents, to interact more efficiently and naturally with real-world environments.

This paper introduces *Visual-only Question Answering* (VoQA), a new multimodal reasoning task that removes the explicit language channel. As shown in Figure 1, the model receives a single composite image that visually encodes both the question and the scene. It must comprehend the embedded question, align it with the corresponding visual content, and generate the correct answer, all through a unified visual input stream. This reframing moves from explicit language-conditioned reasoning in VQA to implicit

✉Correspondence to: Lei Huang <huangleiAI@buaa.edu.cn>. Our code and data are available at https://github.com/AJN-AI/VoQA and https://huggingface.co/datasets/AJN-AI/VoQA.

vision-only reasoning in VoQA, where the model must infer and answer questions purely from visual cues.

To support this task, we construct a large-scale *VoQA Dataset* and a comprehensive *VoQA Benchmark*, containing over 3.35M training and 134k evaluation samples. The training set is derived by converting LLaVA [16] instruction-tuning data into visual-only inputs via text–image rendering. For evaluation, the benchmark is built by converting existing VQA datasets into the same visual-only format, covering tasks adapted from VQAv2 [8], GQA [10], POPE [14], TextVQA [24] and ScienceQA [18].

To assess model capabilities without modifying their weights, we evaluate both open-source and closed-source LVLMs under pure visual-only zero-shot, prompt-guided, and OCR-assisted experiments, revealing a substantial performance drop compared to traditional VQA. Analysis using the *Question Alignment Accuracy* (QAA, see Section 3.3.3) confirms that correctly aligning visual questions is crucial, as higher QAA consistently correlates with better answer accuracy. This finding highlights that effective reasoning depends critically on question alignment, which involves the ability to locate and interpret visual text, motivating supervised fine-tuning (SFT) strategies that guide models to first identify and align with the embedded question before producing an answer.

We conducted a comprehensive investigation of SFT strategies. For VoQA Baseline-SFT (see Section 4.1), we found that models often either repeat the embedded question or generate answers unrelated to it, reflecting poor alignment between visual prompts and the intended response. To address this, we introduce *question-alignment fine-tuning strategies* that guide the model to first identify the visual question before reasoning, effectively enhancing its ability to comprehend visually embedded questions and perform reasoning purely within the visual modality.

Given the inherent difficulty of the VoQA task, it is desirable for the model to not only perform well on VoQA but also acquire *cross-task generation capability* (see Section 4.2) to traditional VQA. Among these strategies, *Question–Role–Answer Supervised Fine-Tuning* (QRA-SFT, see Section 4.3) guides the model to predict the complete Question–Role–Answer sequence, which both enforces prior alignment with the visually embedded question, ensuring strong performance on the VoQA task, and preserves the input-output format of traditional VQA, thereby maintaining the model's ability to generalize to text-based VQA.

2. Related Works

2.1. Visual Question Answering

Early Visual Question Answering (VQA) research focused on object recognition and basic reasoning over static images, as in VQAv1 [2], VQAv2 [8], and CLEVR [11], while datasets such as Visual7W [36], GQA [10], and VizWiz-VQA [9] expanded to region grounding, compositional reasoning, and real-world robustness. Subsequent work explored bias analysis [12], external knowledge [19], and multilingual or large-scale benchmarks such as MMBench [17] and MEGA-Bench [6]. To evaluate text understanding, TextVQA [24], OCR-VQA [21], and ChartQA [20] evaluate models' ability to read and reason over scene text or structured data.

Although these methods vary and explore different aspects of multimodal capabilities, they all provide the question as a separate textual input. In contrast, our VoQA task *embeds the question as rendered text* within the image, requiring unifying perception, text recognition, and reasoning without any explicit textual prompt.

2.2. Large Vision-Language Models

Large Vision-Language Models (LVLMs) bridge vision and language via large-scale pretraining on image–text pairs. Early works such as CLIP [23] and SigLIP [32] established strong cross-modal alignment, while recent systems combine vision encoders with Large Language Models (LLMs) [1, 13, 16, 22, 34], enabling unified multimodal reasoning across captioning and VQA. Representative open-source models such as Qwen3-VL [26], and InternVL3.5 [27] achieve performance comparable to proprietary systems like GPT-5 and Gemini2.5-Pro.

Despite their progress, most LVLMs still depend on explicit textual input for perception and reasoning. Their reliance on language priors leaves open the question of whether multimodal understanding can emerge purely from the visual channel when textual information is embedded within the image itself.

2.3. OCR Systems and OCR-Enhanced LVLMs

Traditional OCR systems (e.g., Tesseract [25]) follow a modular detect–recognize pipeline, which limits robustness in complex or noisy scenes. Recent Transformer-based approaches such as Nougat [5], GOT-OCR2.0 [28], and DeepSeek-OCR [29] pursue unified end-to-end designs, extending OCR to structured documents, formulas, and diagrams. Recent LVLMs demonstrate emergent OCR-like capabilities through instruction-tuned multimodal pretraining [3, 26, 27, 35], yet they remain dependent on vision-language instruction tuning.

In contrast, our *VoQA* task removes textual input entirely, embedding questions visually within the image. This setting moves beyond text recognition, requiring models to both perceive and reason over in-image textual content purely through the visual modality.

3. VoQA Task, Dataset and Benchmark

We first introduce the VoQA task along with the VoQA dataset and benchmark designed for it. To investigate the model capability on this task, we evaluate both open-source and closed-source models across multiple settings.

3.1. Task Definition

The **Visual-only Question Answering (VoQA)** task is defined as a *visual reasoning problem* in which all information required for inference, including both visual content and a textual question, is conveyed solely through the visual modality. Formally, let $\mathcal{I}$ denote the space of natural images, and $\mathcal{T}$ denote the space of textual sequences. In VoQA, the input to the model is a *composite visual input* $I_v \in \mathcal{I}_v \subset \mathbb{R}^{H\times W\times 3}$, which visually encodes both $I_s \in \mathcal{I}$ (scene image) and $T_q \in \mathcal{T}$ (question text). Given this visual-only input, a reasoning function $\mathcal{M} : \mathcal{I}_v \to \mathcal{T}_a$ produces an answer $\hat{T}_a = \mathcal{M}(I_v)$. This formulation abstracts VoQA as a unified visual reasoning task where both the scene and the query are presented within a single visual stream, and all semantic inference is performed in the visual space without explicit textual input.

3.2. Dataset Construction

To build the VoQA dataset and benchmark, we introduce the text-image rendering methods used to generate visual-only inputs and the data sources used for training and evaluation.

3.2.1. Text-Image Rendering Methods

To generate the VoQA input image I_v, we first render the question T_q into an RGB image I_q. While a straightforward approach is to concatenate I_q and I_s (details in Supplementary Material, *SM*), this maintains a distinct spatial separation between the two modalities.

Our primary approach, Watermark Rendering (Fig. 2), addresses this by embedding I_q as an integrated watermark within I_s. Inspired by the sliding window approach, we select the embedding region based on multiple criteria to ensure minimal occlusion and maximal readability. Final watermark colors are selected based on WCAG contrast guidelines. Technical details, including candidate region scoring, contrast-aware color selection, are provided in *SM*.

3.2.2. Data Preparation

Training Dataset. We construct the VoQA training dataset based on the vision–language instruction-tuning data released by LLaVA [16]. Each multi-turn dialogue is split into individual question–answer pairs, and each pair is rendered as a separate composite image by overlaying the question onto the corresponding scene image. The resulting dataset comprises over 3.35 million VoQA samples.

Evaluation Benchmark. We build the VoQA Benchmark by transforming five widely-used VQA datasets, VQAv2 [8], GQA [10], POPE [14], TextVQA [24], and ScienceQA-IMG [18] (denoted as SQA), into the visual-only format. Each image–question pair is rendered into a single composite image using our text–image rendering process. The benchmark spans a broad range of question types and reasoning skills, and includes over 134k evaluation samples. Details of each sub-task are provided in *SM*.

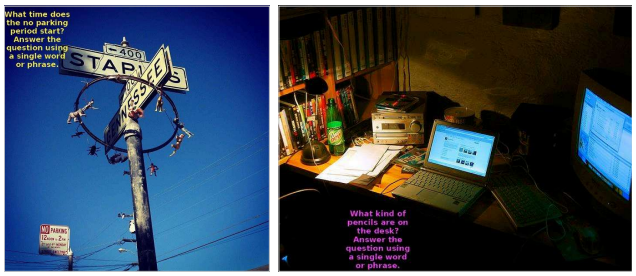


Figure 2. Two examples of watermark rendering with different text colors.

3.3. Evaluation on the VoQA Benchmark

We evaluate both open-source and closed-source models on traditional VQA benchmarks as baseline performance and on the VoQA benchmark under three main settings: (1) *pure visual-only zero-shot*, (2) *prompt-guided evaluation*, where models are explicitly instructed to reconstruct and answer the visually embedded question, and (3) *OCR-assisted evaluation*, where a strong OCR system provides auxiliary textual input.

3.3.1. Evaluation Setup

Overview of Evaluation Settings. We evaluate seven LVLMs, including six open-source models: InternVL3-1B [35], DeepSeek-VL2-Tiny (1B) [30], Qwen2.5-VL-3B-Instruct [4], TinyLLaVA-3.1B [33], BLIP-3 (4B) [31], and LLaVA-v1.5-7B [16], as well as the closed-source model Doubao-1.5-thinking-vision-pro. All models are evaluated under four settings: (1) traditional VQA benchmarks; (2) the VoQA benchmark under a *pure visual-only zero-shot* setting, where models receive only images with visually embedded questions; (3) the VoQA benchmark with *prompt-guided evaluation*, where prompts explicitly instruct the model to locate and interpret the embedded question before answering; and (4) the VoQA benchmark with *OCR-assisted evaluation*, where the OCR output from a strong system (*GOT-OCR 2.0* [28]) is provided as auxiliary textual input together with the composite image.

Prompt Engineering Setup. We design three prompting configurations (see *SM* for templates): (1) **Light Prompt**, which briefly instructs the model to locate and answer the embedded question; (2) **Short Workflow Prompt**, which adds minimal reasoning guidance and requires *structured*

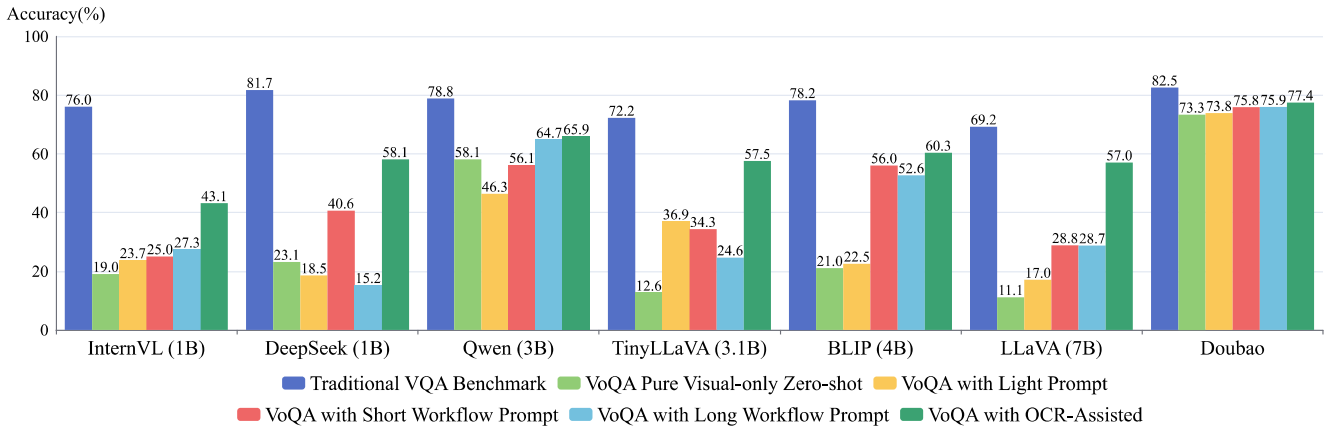


Figure 3. Average Accuracy (%) of all models on the VoQA benchmark under various zero-shot settings (*pure visual-only*, *prompt-guided*, and *OCR-assisted*) across all datasets, compared with traditional VQA benchmarks. The models correspond to those introduced in Section 3.3.1. Across all models and zero-shot settings, performance on VoQA is noticeably lower than on traditional VQA, highlighting the challenge of reasoning over visually embedded questions. *SM* provides comprehensive per-dataset performance along with evaluations for more closed-source models.

JSON outputs containing both the detected question and final answer; (3) **Long Workflow Prompt**, which further enforces multi-step reasoning and stricter output constraints to ensure stronger question alignment.

To ensure fair comparison, we apply consistent response filtering across all VoQA settings. Details of the implementation are provided in *SM*.

3.3.2. Evaluation Results

As shown in Figure 3, performance of all models drops significantly on *pure visual-only zero-shot* VoQA compared with traditional VQA, highlighting the challenge of reasoning over visually embedded questions, even for the closed-source Doubao model.

Under prompt-guided evaluation, models achieve moderate gains by explicitly parsing the embedded question. Incorporating OCR results, as an external auxiliary system that directly provides the model with the detected embedded question, achieves the highest accuracy among all VoQA settings. Nevertheless, performance remains noticeably below traditional VQA, indicating that current LVLMs still struggle when textual information is presented solely in the visual modality.

Notably, the closed-source Doubao model performs relatively well, likely due to its large size and extensive pretraining, which help it interpret visual questions even without specialized guidance.

3.3.3. Result Analysis

Model behavior Analysis. To better understand model behaviors under the *pure visual-only* zero-shot setting, we categorize their responses to the VoQA task into five types: (1) **Question-unaware image captioning**: ignores the embedded question and generates a generic caption; (2) **Question-aware image captioning**: partially recognizes the question and produces a loosely related description; (3) **Repeating the question**: restates the question without answering; (4) **Answering incorrectly**: identifies the question but gives a wrong or irrelevant answer; (5) **Answering correctly**: accurately understands and answers the embedded question. We show a representative example in Table 1, which refers to a road construction scene that requires understanding both the overall context and multiple text-rich elements such as warning signs and road symbols. In this case, most models fail to produce meaningful answers. Instead, they tend to either repeat the question or generate image descriptions without giving responses.

Question Recognition Analysis. To analyze the performance gap, we measure models' performance in recognizing the visually embedded question. We define *Question Alignment Accuracy (QAA)* as:

$$\text{QAA} = 1 - \frac{\min(\text{EditDistance}(\hat{q}, q))}{\text{len}(q)}, \tag{1}$$

where q is the ground-truth question, $\hat{q}$ is the model-predicted question, and *EditDistance* counts the character-level insertions, deletions, or substitutions to match the strings. This metric measures how accurately a model identifies the question before reasoning.

We compute the Question Alignment Accuracy (QAA) on the VoQA benchmark, reporting results by reasoning correctness for GQA, POPE, and SQA, and by confidence threshold (not less than 0.5) for TextVQA, excluding VQAv2 due to unavailable correctness labels. QAA is calculated under both workflow prompts.

As shown in Figure 4, higher QAA generally correlates with stronger reasoning performance, as correctly an-

Table 1. A representative example from the *pure visual-only* zero-shot VoQA benchmark, showing a visually embedded question and comparing responses from different models. The question *"Where will these things eventually be seen? Answer the question using a single word or phrase."* is asked about a road construction site. The models correspond to those introduced in Section 3.3.1. Note: ellipses in some model outputs are manually added for brevity and do not affect the classification of behavior type. More examples are provided in *SM*.

Input	Model	Response	Behavior Type
	LLaVA (7B)	*The image features a construction site with several road signs and barricades. There are ...*	Question-unaware image captioning
	InternVL (1B)	*The text in the image suggests that the question "where will these things eventually be? Answer ... phrase." is used to refer to road construction. The related sign ...*	Question-aware image captioning
	TinyLLaVA (3.1B)	*The image captures a scene of road construction ... The image also includes a text that reads "WHERE WILL THESE THINGS EVENLY BE ..."*	Question-aware image captioning
	DeepSeek (1B)	*Where will these things eventually be seen? Answer the question using a single word or phrase.*	Repeating the question
	BLIP (4B)	*The text 'Where will these things eventually be seen? Answer ... phrase.' is overlaid on the image.*	Repeating the question
	Qwen (3B)	*Disappeared*	Answering incorrectly
	Doubao	*road*	Answering correctly

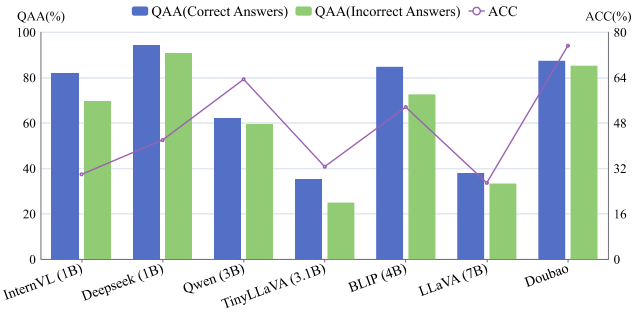


Figure 4. Average Question Alignment Accuracy (QAA) and Answer Accuracy (ACC) for all models across four VoQA sub-tasks under the two workflow prompt settings, except VQAv2. *Correct Answers* and *Incorrect Answers* indicate averages computed over correctly and incorrectly answered samples, respectively. The models correspond to those introduced in Section 3.3.1. For each model, the results shown correspond to the workflow setting (short or long) that yields the higher average ACC. Complete QAA and ACC results are provided in *SM*.

swered samples consistently exhibit higher QAA than incorrect ones. Incorporating an external OCR system provides direct access to the embedded question and achieves the highest VoQA accuracy, but performance still falls short of traditional VQA due to challenges like background text unrelated to the question.

Overall, these observations highlight that robust VoQA reasoning requires effective question localization and understanding, highlighting the critical role of supervised fine-tuning in teaching the model to first recognize the question and then generate the answer, as follows in the next section.

4. Question-Alignment Fine-Tuning for VoQA

By Supervised Fine-Tuning (SFT) on instruction–response pairs from the VoQA training data, models are expected to better recognize embedded questions and reason over visual content, building on the proven effectiveness of prior SFT frameworks [7, 16, 34]. We first introduce its standard application in traditional VQA.

VQA Baseline-SFT (denoted as VQA SFT). In the traditional VQA setting, the model is fine-tuned with both the image and the *textual question* as inputs, using the corresponding answer as supervision (Figure 5, Line 1). This straightforward strategy enables the model to learn visual-text reasoning conditioned on an explicitly provided question.

However, due to the difference in input formats between traditional VQA and VoQA, the VQA fine-tuning approach must be adapted when applied to the VoQA task.

4.1. Naive SFT Adaptation to VoQA

VoQA Baseline-SFT. In VoQA Baseline-SFT, the explicit textual question is removed, and the model generates the answer directly from a composite image I_v that visually embeds the question (Figure 5, Line 2), preserving the basic fine-tuning structure. Subsequent experiments use three base models: *TinyLLaVA-1B-Pretrained*, *InternVL3-1B-Pretrained*, and *Qwen2-VL-2B*. For details of experimental settings and base model zero-shot results, please refer to *SM*.

Result Analysis. As shown in Figure 6, VoQA Baseline-SFT moderately improves model performance over zero-shot evaluation on the VoQA task. However, we further observe that the model frequently struggles to interpret the embedded question correctly: it may repetitively restate the question without answering, or generate responses that are semantically irrelevant (See *SM* for examples). These behaviors reveal a fundamental limitation: without explicit

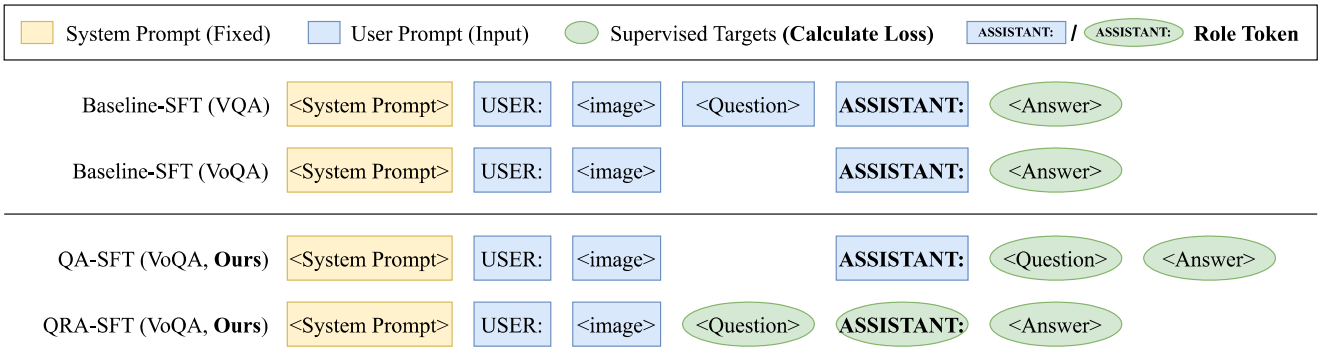


Figure 5. Comparison of four supervised fine-tuning strategies. The first two represent baseline fine-tuning under the VQA and VoQA settings, respectively. The bottom two are our proposed VoQA-specific methods that first align the visually embedded question before generating the answer.

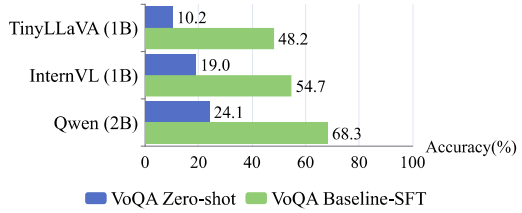


Figure 6. Comparison of average accuracy on the VoQA benchmark between zero-shot evaluation and VoQA Baseline-SFT fine-tuned models. Detailed results for each dataset are provided in *SM*.

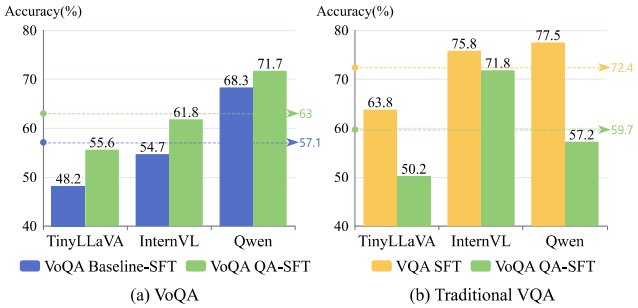


Figure 7. Evaluation of models fine-tuned with different strategies. (a) Average VoQA performance of VoQA Baseline-SFT and VoQA QA-SFT models. (b) Average traditional VQA performance of VQA Baseline-SFT and VoQA QA-SFT models. All models are fine-tuned on either the LLaVA instruction-following dataset (for VQA) or the VoQA dataset (for VoQA), with differences only in input presentation. The base models correspond to the three models shown in Section 4.1. Horizontal lines denote the mean performance across three models under each setting. Detailed results for each dataset are provided in *SM*.

textual guidance, answer-only supervision is insufficient for the model to align its instruction-following behavior with visually embedded questions.

4.2. Enhancing Supervised Fine-Tuning with Question Alignment

To address the limitations observed with Baseline-SFT, we introduce an explicit question alignment strategy to improve reasoning over visually embedded questions.

QA-SFT. We propose *Question–Answer Supervised Fine-Tuning* (*QA-SFT*, Figure 5, Line 3), which incorporates a question reconstruction objective, supervising both the reconstructed question and its answer. This guides the model to first extract the embedded question before reasoning, thereby enforcing question–answer alignment and improving the connection between visual understanding and reasoning. As shown in Figure 7 (a), QA-SFT consistently improves average performance over VoQA Baseline-SFT across all models, indicating that aligning the model with the visually embedded question helps enhance reasoning.

Cross-task Generation to Traditional VQA. Considering that VoQA is an inherently challenging task, we further investigate whether QA-SFT enables models to acquire *cross-task generation capability*—namely, the capacity to generalize acquired reasoning skills beyond the specific training setting. To assess this ability, we evaluate QA-SFT models on the traditional VQA task, where questions are provided in text rather than visually embedded. As VQA and VoQA share the same question–answering objective but differ in input modality, this evaluation serves as a natural test of cross-task generalization and provides a simpler setting to isolate the model's visual understanding capability.

As shown in Figure 7, QA-SFT models perform significantly worse on traditional VQA than models fine-tuned under the standard VQA SFT setting, and in some cases even underperform their VoQA results. These findings indicate that while QA-SFT strengthens visual question alignment in VoQA, it exhibits limited cross-task generation capability toward traditional VQA.

Structure Misalignment. Although QA-SFT enhances alignment with visually embedded questions in VoQA, its

Table 2. Performance comparison of fine-tuning strategies on the VoQA benchmark. Results for VoQA QA-SFT and QRA-SFT are reported, with both models based on the same pre-trained backbone. All results are reported in accuracy (%).

Base Model	Settings	VQAv2	GQA	POPE	TextVQA	SQA	Avg.
TinyLLaVA (1B)	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**
InternVL (1B)	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**
Qwen (2B)	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

input–output structure differs from traditional VQA fine-tuning. In VQA SFT, the textual question is explicitly provided before the role token (*ASSISTANT:*), giving the model a clear reasoning context. In contrast, QA-SFT treats the question as part of the generated output after the role token. This structural discrepancy may cause the model to misalign its reasoning or produce inaccurate answers when facing the traditional VQA format, leading to limited generalization to traditional VQA tasks.

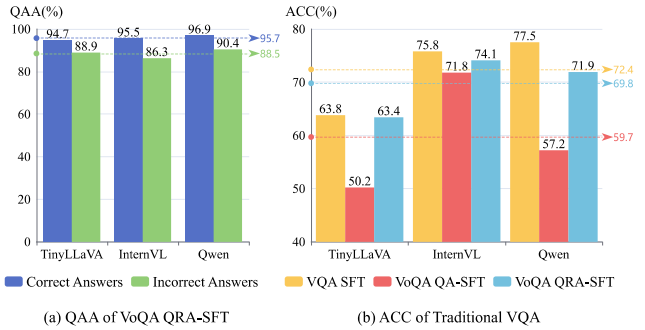
4.3. Enhancing Cross-Task Generalization in Question-Alignment Fine-Tuning

Building on question-alignment fine-tuning, we propose a strategy that explicitly separates question parsing from answer generation while preserving the traditional VQA format, ensuring strong alignment with visually embedded questions in VoQA and enhanced cross-task generalization to VQA compared with QA-SFT.

QRA-SFT. We propose *Question–Role–Answer Supervised Fine-Tuning* (QRA-SFT, Figure 5, Line 4). In QRA-SFT, the model sequentially generates the predicted question, a role token, and the answer. This output design mirrors the traditional VQA input format, keeping compatibility while explicitly enforcing question parsing before answering. In addition, the output sequence follows a structured *Question–Role–Answer* format, helping the model separate question interpretation from answer generation by predicting the role token during inference.

4.3.1. Evaluation on VoQA and Cross-Task VQA Generalization

Effectiveness on VoQA. As shown in Table 2, QRA-SFT matches or slightly outperforms QA-SFT, demonstrating that explicit supervision of question interpretation preserves alignment with visually embedded questions. Figure 8 (a) further shows that correctly answered samples consistently exhibit higher QAA than incorrect ones, exceeding 95% on average across the three models, suggesting that strong question–answer alignment underlies the improved VoQA performance.



(a) QAA of VoQA QRA-SFT (b) ACC of Traditional VQA

Figure 8. Evaluation of models fine-tuned with different strategies. (a) Average QAA results of VoQA QRA-SFT models on correctly and incorrectly answered samples across the four VoQA sub-tasks, excluding VQAv2. (b) Average traditional VQA ACC of VQA Baseline-SFT, VoQA QA-SFT, and VoQA QRA-SFT models. The fine-tuning strategies and base models are consistent with those in Figure 7. Horizontal lines denote the mean performance across three models under each setting. Detailed results for each dataset are provided in *SM*.

Generalization to VQA. Figure 8 (b) shows that models fine-tuned with QRA-SFT generalize substantially better to traditional VQA than those trained with QA-SFT. Their performance approaches that of models fine-tuned directly on VQA SFT, indicating that QRA-SFT improves question alignment in VoQA while also supporting cross-task generalization to traditional VQA.

4.3.2. Token Activation Map Visualization

To further illustrate how QRA-SFT improves the model's reasoning process, we employ the *Token Activation Map (TAM)* [15] technique to visualize token-level attentions on VoQA samples. Figure 9 shows the comparison between the *InternVL3* model and the QRA-SFT fine-tuned model. In Figure 9 (a), we present the original input image, where the question text is embedded in the visual scene. Figure 9 (b) shows the visualization results of the *InternVL3* model. We observe that the model does not explicitly attend to the question area in the image, leading to a misunderstanding of the question and an incorrect answer. In contrast, Figures 9 (c) and (d) show the *QRA-SFT* model's visualization results, where the model first focuses on the textual region containing the question, then shifts its attention to the relevant visual region (the man's hat) when generating the final answer. These results indicate that QRA-SFT effectively guides the model to perform step-wise reasoning: first understanding the question text within the image, and then locating the visual evidence to produce the correct answer.

4.3.3. Ablation Study

Impact of Role Token Design. We further analyze the effect of role token form and semantics in QRA-SFT. As

Figure 9. **Token Activation Map (TAM) visualization results.** (a) Original VoQA input image. (b) Visualization of the *InternVL3* model. (c–d) Visualization of the *QRA-SFT* model. The QRA-SFT model first attends to the textual question region and subsequently to the relevant visual area (the hat), demonstrating a more structured reasoning process. **Note:** In TAM visualizations, the text below each image represents the sequence of generated tokens. The token currently being visualized is highlighted with a black box. The brightness of the overlay indicates attention strength, brighter areas correspond to higher attention on the corresponding token or image region.

Table 3. Influence of role token format and content on VoQA performance. All results are reported in accuracy (%).

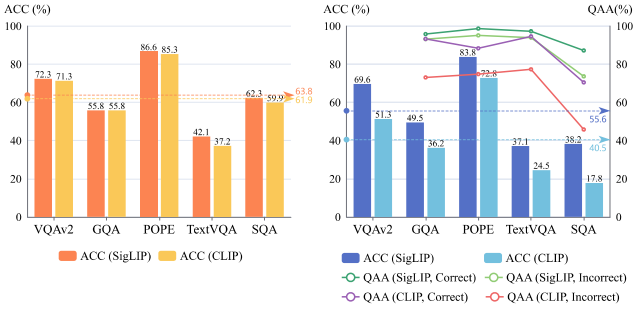
Model	Role Token	VQAv2	GQA	POPE	TextVQA	SQA	Avg.
InternVL (1B)	*ASSISTANT:*	**73.0**	52.3	**86.2**	55.4	**56.4**	**64.6**
	\nassistant\n	72.6	**52.7**	85.8	**56.3**	49.1	63.3
TinyLLaVA (1B)	*ASSISTANT:*	69.6	49.5	**83.8**	**37.1**	**38.2**	**55.6**
	HELPER:	**69.7**	**49.6**	83.7	36.8	36.9	55.3
	CAT:	69.4	49.3	83.6	36.9	36.6	55.2

shown in Table 3, different token formats (e.g., *ASSISTANT:* vs. others) and contents (e.g., *ASSISTANT* vs. *HELPER*) yield nearly identical results, suggesting that the role token mainly acts as a weak structural separator to stabilize formatting rather than providing semantic guidance.

Impact of Vision Encoders on VoQA. We compare two vision encoders under identical fine-tuning pipelines to isolate their effect on performance and visual–text alignment. For each vision encoder, the corresponding model is fine-tuned separately on the LLaVA instruction-tuning dataset (for VQA SFT) and the VoQA dataset. As shown in Figure 10, both encoders perform similarly on traditional VQA, indicating comparable general visual features. On VoQA, however, SigLIP outperforms CLIP by over 15%, emphasizing the importance of stronger visual grounding and text-sensitive representations for interpreting embedded questions. Consistently, models with higher QAA on VoQA also achieve better accuracy, confirming that visual–text alignment remains a main bottleneck in this task.

5. Conclusion

In this work, we present *VoQA*, a vision-only reasoning task designed to advance robust understanding and interaction in real-world visual scenes where textual questions are embedded directly in the image. To support this task, we build



(a) Traditional VQA (b) VoQA

Figure 10. Impact of visual encoders on VQA and VoQA tasks. *TinyLLaVA-1B-Pretrained* originally uses **SigLIP** (siglip-so400m-patch14-384) as its vision encoder. We also pretrain it with **CLIP** (clip-vit-large-patch14-336) under identical fine-tuning settings. We report ACC on both VQA and VoQA, and QAA on VoQA. Horizontal lines denote the mean performance across all sub-tasks under each setting.

the *VoQA Dataset* and *VoQA Benchmark*, revealing a substantial gap between human perception and current multimodal models under vision-only conditions. Furthermore, our study of question-alignment fine-tuning demonstrates that QRA-SFT not only markedly enhances VoQA performance but also preserves strong *cross-task generalization* to traditional VQA tasks. Together, these contributions establish VoQA as a practical and insightful testbed for future research on unified visual reasoning.

Limitations and Future Work. Our experiments are limited to small-scale LVLMs due to computational constraints, though the proposed strategies are model-agnostic and readily scalable. Future directions include applying QRA-SFT to larger backbones, exploring diverse visual text styles and multilingual scenarios, and extending VoQA to embodied or interactive environments.

Acknowledgments

This work was partially supported by the National Science and Technology Major Project (2022ZD0116310), National Natural Science Foundation of China (Grant No. 62476016 and 62441617), the Fundamental Research Funds for the Central Universities.

References

[1] 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. In *NeurIPS*, 2022. 1, 2

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *ICCV*, pages 2425–2433. IEEE Computer Society, 2015. 1, 2

[3] 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. *CoRR*, abs/2308.12966, 2023. 2

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025. 1, 3

[5] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. In *ICLR*. OpenReview.net, 2024. 2

[6] Jiacheng Chen, Tianhao Liang, Sherman Siu, Zhengqing Wang, Kai Wang, Yubo Wang, Yuansheng Ni, Ziyan Jiang, Wang Zhu, Bohan Lyu, et al. Mega-bench: Scaling multimodal evaluation to over 500 real-world tasks. In *ICLR*. OpenReview.net, 2025. 2

[7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023. 5

[8] 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 *CVPR*, pages 6325–6334. IEEE Computer Society, 2017. 1, 2, 3

[9] 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 *CVPR*, pages 3608–3617. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[10] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709. Computer Vision Foundation / IEEE, 2019. 1, 2, 3

[11] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pages 1988–1997. IEEE Computer Society, 2017. 2

[12] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *ICCV*, pages 1983–1991. IEEE Computer Society, 2017. 2

[13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742. PMLR, 2023. 1, 2

[14] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, pages 292–305. Association for Computational Linguistics, 2023. 2, 3

[15] Yi Li, Hualiang Wang, Xinpeng Ding, Haonan Wang, and Xiaomeng Li. Token activation map to visually explain multimodal llms. *CoRR*, abs/2506.23270, 2025. 7

[16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 2, 3, 5

[17] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV (6)*, pages 216–233. Springer, 2024. 2

[18] 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. In *NeurIPS*, 2022. 1, 2, 3

[19] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *CVPR*, pages 3195–3204. Computer Vision Foundation / IEEE, 2019. 1, 2

[20] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *ACL (Findings)*, pages 2263–2279. Association for Computational Linguistics, 2022. 2

[21] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. OCR-VQA: visual question answering by reading text in images. In *ICDAR*, pages 947–952. IEEE, 2019. 2

[22] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. 1, 2

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 2

[24] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *CVPR*, pages 8317–8326. Computer Vision Foundation / IEEE, 2019. 1, 2, 3

[25] R. Smith. An overview of the tesseract OCR engine. In *ICDAR*, pages 629–633. IEEE Computer Society, 2007. 2

[26] Qwen Team. Qwen3-vl technical report. *CoRR*, abs/2511.21631, 2025. 2

[27] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *CoRR*, abs/2508.18265, 2025. 2

[28] Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, Chunrui Han, and Xiangyu Zhang. General OCR theory: Towards OCR-2.0 via a unified end-to-end model. *CoRR*, abs/2409.01704, 2024. 2, 3

[29] Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-ocr: Contexts optical compression. *CoRR*, abs/2510.18234, 2025. 2

[30] 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. *CoRR*, abs/2412.10302, 2024. 3

[31] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S. Ryoo, et al. xgen-mm (BLIP-3): A family of open large multimodal models. *CoRR*, abs/2408.08872, 2024. 3

[32] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11941–11952. IEEE, 2023. 2

[33] Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. Tinyllava: A framework of small-scale large multimodal models. *CoRR*, abs/2402.14289, 2024. 3

[34] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*. OpenReview.net, 2024. 2, 5

[35] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *CoRR*, abs/2504.10479, 2025. 2, 3

[36] Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, pages 4995–5004. IEEE Computer Society, 2016. 2