Ayeshaamjad at ImageCLEF 2025 Multimodal Reasoning: Visual Question Answering with Structured Data Extraction and Robust Reasoning

Ayesha Amjad[1,†], Fatima Seemab[1,†], Saima Kausar[1,†], Seemab Latif[2,*,†] and Mehwish Fatima[2,†]

[1]National University of Sciences and Technology, Islamabad, Pakistan
[2]National University of Sciences and Technology, Islamabad, Pakistan

Abstract

Visual Question Answering (VQA) is a powerful tool for evaluating the generalization and reasoning capabilities of artificial intelligence (AI) models in educational contexts. However, VQA has a variety of challenges, including a wide range of visual components, a large number of question types, and a multitude of visual elements. This work proposes a multimodal visual question answering framework (MVQF) for exam-style VQA that combines Gemini's structured data extraction with DeepSeek's robust reasoning capabilities, aiming to overcome the persistent challenges in multilingual, multimodal question answering that these studies have collectively identified. The framework focuses on the EXAMS-V 2025 challenge in supports in English, Arabic, and Chinese. Our model navigates the dataset's diverse visuals and multimodal demands like a seasoned scholar. We compare qualitative results with alternative models and provide an in-depth analysis of the performance of subject and visual elements.

Keywords: Multimodal Visual Question Answering (MVQA) , Exam-style VQA , Structured Data Extraction (Gemini), Robust Reasoning Capabilities (DeepSeek), Multilingual VQA (English, Arabic, Chinese), EXAMS-V 2025 Challenge, Visual & Textual Question Types, Educational AI Application

1. Introduction

Visual Question Answering (VQA) has emerged as a critical multimodal task that combines visual and textual information to generate precise answers to queries grounded in images. This capability mimics human-like cognitive reasoning and has transformative potential in domains including autonomous systems, medical diagnostics, and educational tools. This is demonstrated by the ImageCLEF Multimodal Visual Question Answering task, which tests systems' ability to analyze exam-style images with text, diagrams, or tables and provide precise answers [1]. Its impact lies in advancing AI's ability to handle real-world scenarios, such as analyzing medical images or educational content, enhancing decision-making and accessibility.

The intricacy of merging visual and linguistic data makes VQA assignments extremely difficult. Robust visual processing is necessary to extract pertinent features from a variety of image elements, such as schematics, labeled arrows, or scientific notations. Current systems often have trouble processing noisy

CLEF 2025 Working Notes, September 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
†These authors contributed equally.
aamjad.msai24seecs@seecs.edu.pk (A. Amjad); fseemab.msai24seecs@seecs.edu.pk (F. Seemab); skausar.msai24seecs@seecs.edu.pk (S. Kausar); seemab.latif@seecs.edu.pk (S. Latif); mehwish.fatima@seecs.edu.pk (M. Fatima)
0009-0001-1670-565X (A. Amjad); 0009-0007-6546-2467 (F. Seemab); 0000-0002-0877-7063 (S. Kausar); 0000-0002-5801-1568 (S. Latif); 0000-0002-9421-8566 (M. Fatima)

or insufficient visual input, generalizing across a variety of domains, and making appropriate decisions without overfitting to particular datasets. In high-stakes real-world applications, these difficulties restrict the dependability of VQA systems [2].

Recent VQA systems leverage deep learning architectures such as transformer-based models for language understanding and convolutional or vision transformer (ViT) backbones for image feature extraction. Architectures such as ViLBERT, LXMERT, and CLIP have demonstrated efficacy on general-purpose datasets by aligning visual and textual embeddings in a shared space. However, these models often struggle with domain-specific challenges like those presented in exam-style VQA. Limitations include poor generalization to structured visuals (e.g., tables, schematics), inadequate handling of sparse or noisy input, and overreliance on large-scale pretraining that fails to transfer well to specialized datasets such as MBZUAI/EXAMS-V.[3]

To address these deficiencies, we propose a modular VQA pipeline tailored for the ImageCLEF 2025 challenge. Our approach integrates Gemini, a state-of-the-art visual extraction model capable of structured parsing of complex image elements, with DeepSeek, a reasoning engine designed for logical inference over structured data. Gemini processes each image to produce a JSON representation encapsulating textual elements, labels, arrows, diagrams, and tabular data. This structured intermediate representation enables DeepSeek to perform targeted reasoning and generate well-informed responses with minimal hallucination or contextual drift.

Contributions

This work makes the following contributions to the field of multimodal question answering:

- **Development of a novel two-stage VQA pipeline** that combines *Gemini* for structured visual data extraction with *DeepSeek* for reasoning, significantly improving performance on images style exam.
- **Integration of a language-filtering preprocessing module** to isolate English-language samples in the MBZUAI/EXAMS-V dataset[3], enhancing the precision of multilingual benchmarking.
- **Robust handling of complex image structures** such as tables, labels, and multilingual annotations through structured JSON outputs, facilitating more accurate downstream reasoning.
- **Demonstration of domain generalizability and scalability** in educational VQA applications, addressing key limitations of prior VQA systems in structured visual environments.

2. Related Work

Multimodal question answering (QA) has progressed from basic image-question pairs toward more advanced tasks involving complex visual structures, diverse subject matter, and multilingual text. Early benchmarks such as the VQA [4] laid the foundation for this field by introducing image-question-answer triplets for natural scene understanding. While influential, VQA lacked structural diversity and was not suited for reasoning tasks involving diagrams, charts, or domain-specific educational content.

To address this limitation, the [5] introduced annotated science diagrams that required models to reason over labeled structures. The initial models used rule-based parsing and handcrafted matching techniques, but these approaches were noisy and lacked scalability. This shift toward layout-sensitive visual tasks was expanded by Infographic VQA.

[6], which introduced visually complex content like charts, posters, and data tables. Pipelines for this dataset often combined OCR tools (such as Tesseract) with transformer-based models like BERT and T5. However, these hybrid models while slightly better still struggled with numeric reasoning and aligning textual elements accurately.

Lu et al. introduced ScienceQA [7], which combined images, questions, and answers with supporting explanations. They used chain-of-thought (CoT) prompting, leveraging models like UnifiedQA and GPT-3. This led to a noticeable performance boost in few-shot and zero-shot tasks. However, ScienceQA's reliance on clean visual input meant that performance dropped sharply in the presence of noisy diagrams or poor OCR, emphasizing the fragility of current approaches.

While ScienceQA emphasized reasoning through CoT in English science education, M3Exam expanded the scope toward multilinguality and script diversity. In an attempt to broaden the multilingual scope of multimodal QA, Liang et al. presented M3Exam [8], a benchmark that spanned nine languages. This paper tested popular models like GPT-4, ChatGPT-3.5, Claude, and Google Bard. Results revealed that these models significantly underperformed on scripts like Arabic, Thai, and Hindi due to unreliable OCR outputs from tools like Google Vision and PaddleOCR, leading to reasoning failures and hallucinations.

Zhang et al. proposed MMMU [9], targeting domain-diverse QA spanning 30 academic subjects. They evaluated 14 vision-language models, including BLIP-2, LLaVA, and GPT-4V. Despite GPT-4V leading with a 55.7% score, the models underperformed on structured visual inputs like diagrams or flowcharts underscoring limitations in spatial reasoning.

Yue et al. addressed the issue of shortcut learning in MMMU-Pro [10], where all content was embedded into images and answer locations were randomized. Models were forced to rely on visual layout understanding rather than text matching. Despite this, top models like GPT-4V and Claude achieved less than 30% accuracy, revealing a clear gap in layout parsing and robust visual reasoning. Li et al. tackled math-specific challenges in MathVista [11], a dataset with over 6,000 visual math questions. Despite integrating symbolic solvers like Mathpix and SymPy with GPT-4 and CoT prompting, models only achieved approximately 34%. Errors often arose from misreading axes, legends, or interpreting geometric relationships again pointing to weak visual grounding.

Das et al. culminated these efforts with EXAMS-V [3], a multilingual, multimodal benchmark featuring 24,000 questions across 13 languages and domains. The models evaluated including GPT-4V, Claude, Gemini, and Bard performed poorly on layout-heavy and noisy visual inputs. Despite leveraging large-scale models and prompt engineering, they struggled with low-resource languages, diagram parsing, and multi-hop reasoning.

These findings highlight that despite architectural scaling, vision-language models fall short in layout reasoning, multilingual robustness, and structural interpretation. Table 1 provides a comparative overview of major multimodal QA benchmarks, the nature of QA tasks, modeling approaches, performance trends, and persistent limitations in layout reasoning and multilingual robustness.

In conclusion, while recent models and prompting methods have improved, current vision-language systems still face major challenges in understanding structured visuals and supporting multiple languages. These weaknesses affect their ability to answer questions accurately in real-world educational settings. To overcome these issues, we propose a modular architecture that separates visual parsing from multilingual reasoning. This approach improves the understanding of the layout, makes error diagnosis easier, and leads to more reliable and transparent question answering in diverse visual and language formats.

Table 1
Summary of Key Multimodal QA Benchmarks and Model Limitations

Benchmark	Task Focus	Top Models	Score	Key Limitations
VQA [4]	Natural scenes QA	CNN + LSTM	60–70%	No reasoning, lacks structured input
AI2D [5]	Science diagrams	Rule-based	–	Fragile parsing, low scalability
InfographicVQA [6]	Charts, posters	BERT + OCR	50–60%	OCR misalignment, dense layout
ScienceQA [7]	CoT-based science QA	UnifiedQA, GPT-3	+3–18%	Sensitive to OCR noise
M3Exam [8]	Multilingual exams	GPT-4, Bard	<40%	OCR fails on low-resource scripts
MMMU [9]	Subject-rich QA	GPT-4V, LLaVA	55.7%	Poor at structured diagrams
MMMU-Pro [10]	Shortcut-resistant QA	GPT-4V, Claude	<30%	Layout parsing weak
MathVista [11]	Visual math QA	GPT-4 + SymPy	34%	Fails on units, axes, geometry
EXAMS-V [12]	Multilingual + layout	GPT-4V, Claude	<40%	Logic, script, layout failures

3. Dataset

The EXAMS-V dataset([3]), hosted on Hugging Face, is a comprehensive multimodal and multilingual benchmark designed for the ImageCLEF Visual Question Answering ([1]) task. It is based on actual high school exam questions from several nations and represents a range of curriculum and educational systems. The EXAMS-V dataset comprises 20,932 samples across 20 subjects, covering grades 4–12, and includes 11 languages from 7 language families, making it a diverse resource for multimodal and multilingual assessment of large language models (LLMs) and vision-language models (VLMs). Key highlights include:

- **Language Diversity**: The dataset features high-resource languages (e.g., English: 724 questions, Chinese: 2,635 questions) and low-resource languages (e.g., Bulgarian: 2,132 questions, Croatian: 3,969 questions, Serbian: 1,434 questions). It spans Germanic, Slavic, Romance, Sino-Tibetan, Semitic, and Finno-Ugric language families, with Arabic introducing right-to-left script. This diversity supports evaluating closely related languages and multilingual capabilities.

- **Parallel Questions**: The dataset includes parallel question sets for Croatian exams in Serbian (1,207 questions) and Italian (1,147 questions), and for Arabic exams in English (262 questions across Science, Physics, Chemistry, and Biology), enabling cross-lingual analysis.

- **Subject Diversity**: Initially, 83 subjects were collected, but after aggregation to address naming inconsistencies, 20 subjects were grouped into three categories: Natural Sciences (53.02%), Social Sciences (27.15%), and Others (Applied Studies, Arts, Religion, etc., 19.82%).

- **Question Complexity**: Questions, primarily from high school exams, vary by subject. Natural Sciences (e.g., Physics, Chemistry, Biology, Mathematics) require foundational knowledge and complex reasoning. Geography and History demand region-specific knowledge. The Polish section includes 55 professional exam questions across fields like accounting and motor vehicle services, requiring precise professional understanding. - **Question Types**: The dataset includes both multimodal (visual) questions (e.g. 700 in Croatian, 1,991 in Chinese) and text-only questions (e.g., 3,269 in Croatian, 644 in

Chinese), with varying distributions across languages.

Table 2
Statistics by language including grade levels, number of subjects, and question types [13].

Language	Family	Grade	# Subjects	# Questions	# visual Q.	# text Q.
English	Germanic	11, 12	4	724	181	543
Chinese	Sino-Tibetan	8–12	6	2,635	1,991	644
French	Romance	12	3	439	50	389
German	Germanic	12	5	819	144	675
Italian	Romance	12	11	1,645	292	1,353
Arabic	Semitic	4–12	6	823	117	706
Polish	Slavic	12	1	2,511	422	2,089
Hungarian	Finno-Ugric	12	6	3,801	495	3,306
Bulgarian	Slavic	4, 12	4	2,132	435	1,697
Croatian	Slavic	12	13	3,969	700	3,269
Serbian	Slavic	12	11	1,434	259	1,175

The data set was divided into training sets (66%), validation sets (20%), and test sets (14%) to facilitate model learning, hyperparameter tuning, and objective evaluation. The data set contains no duplicates and no missing details as it was pre-processed by the organizers. The data set was filtered to separate samples in English, Arabic, and Chinese language. The multilingual, multimodal, and diverse subject coverage of this data set makes it ideal for testing the reasoning and generalizability of AI models in educational contexts.

4. Proposed Methodology

We propose a Multimodal Visual Question Answering Framework (MVQF) for multilingual exam-style VQA. The framework targets the ImageCLEF 2025 challenge across supports in English, Arabic, and Chinese. At a high level, our framework operates in two main stages.

1. An Image Description Module using Gemini-1.5 Flash for structured content extraction from exam-style question images.
2. An Answer Generation Module powered by DeepSeek-R1-Distill-LLaMA for reasoning and answer selection.

4.1. Prompt-Guided Visual Decomposition

This module uses a zero-shot high-precision instructional prompt which directs Gemini-1.5 Flash to decompose images into structured output fields (e.g., question_text, diagram_caption) while suppressing reasoning. This separation ensures modularity and avoids any bias from early reasoning. We design the prompt to handle multilingual inputs. Gemini correctly identifies the language of the question and processes it accordingly. The output is returned in a JSON format that encodes all relevant visual information in structured form.

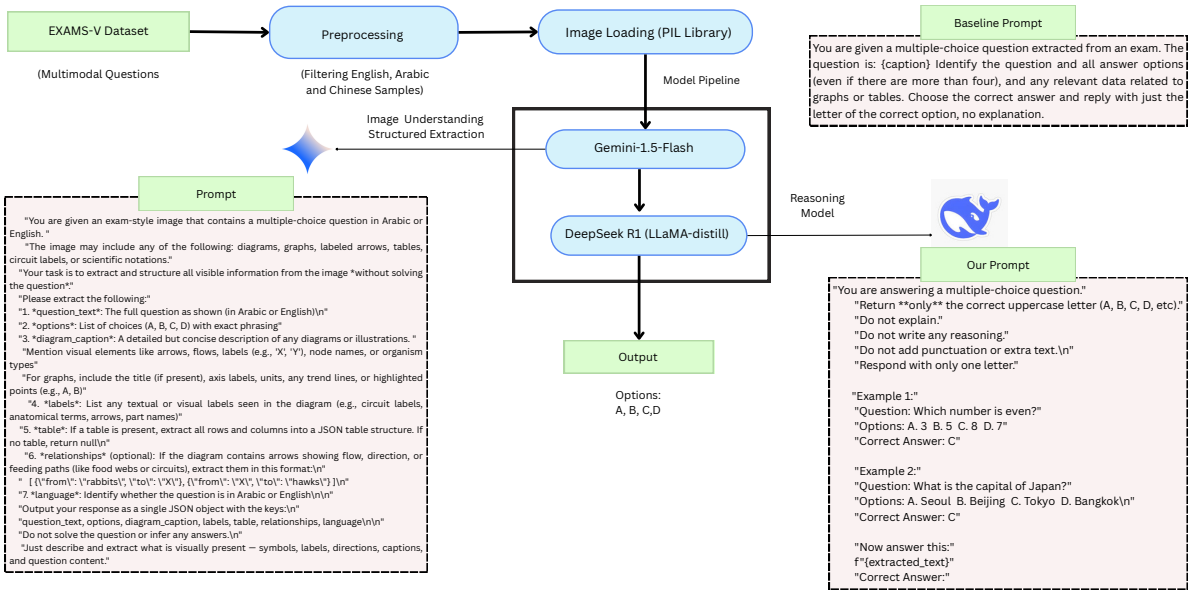

Figure 1: Architecture Diagram of Proposed System

4.2. Answer Generation Module

The Answer Generation Module takes the extracted content and selects the correct option. We use the DS-R1-Distill-LLaMA 70B model for this task. We construct a Strict Single-Letter Extraction (SSLE) Prompt that embeds the question and its options. The model is instructed to return only the correct option letter. We prevent it from generating explanations or additional text.

This minimalistic format ensures consistency across languages. The model reads the question context and reasons over the textual description, producing a single-letter answer.

A key strength of MVQF is its language independence. No separate models or prompt translations were needed for:

- English
- Arabic
- Chinese

The same pipeline and prompts were used across all languages, and Gemini and DeepSeek handled multilingual input natively. This makes our solution scalable and easily extensible to additional languages without retraining or fine-tuning.

5. Experimental Design

This section presents our evaluation strategy, model configurations, and experimental setup for the EXAMS-V benchmark [3]. We broke the reasoning pipeline into intermediate steps and performed experiments separately for each stage. This helped us pinpoint performance bottlenecks, whether they

originated from visual parsing (OCR or captioning) or from the reasoning model itself. Due to hardware limitations, all evaluations were conducted on curated subsets of the dataset.

5.1. Models

We explored two system types: multimodal models and modular pipelines. Multimodal models attempt end-to-end question answering directly from images. Modular pipelines, on the other hand, separate the process into two stages visual content extraction (via OCR or captioning), followed by text-based reasoning. This setup allowed us to diagnose weaknesses at each stage more effectively. Our experiments span three EXAMS-V languages: English, Arabic, and Chinese.

5.1.1. English

The English subset of EXAMS-V included a variety of question formats such as tables, graphs, labeled diagrams, and multi-step reasoning questions. The questions were chosen to test both visual perception and logical reasoning. We tested several combinations:

- **Multimodal Models:** Mistral LLaVA was tested on 50 English MCQs to assess reasoning directly from images. **DeepSeek-VL** was also considered, but could not be executed due to its extremely high memory requirements (>80GB VRAM).
- **Visual Parsing:** We evaluated three tools. **Tesseract OCR** was used for basic text extraction on clean layouts but performed poorly on rotated, dense, or complex visuals. **BLIP**, used for generating image captions, struggled with detailed or scientific content and often missed key elements. In contrast, **Gemini 1.5 Flash** provided layout-aware OCR and structured parsing, and was ultimately selected as the preferred parser for its ability to accurately retain spatial structure and handle diverse visual formats across languages.
- **Modular Pipelines:**
  - **Gemini + DeepSeek-R1 Distill LLaMA (Proposed):** Gemini extracted structured content from images; DeepSeek-R1 handled reasoning.
  - **Gemini + Mistral-7B:** Used for a simpler pipeline pairing Gemini with Mistral-7B.
  - **BLIP + Mistral-7B:** Used BLIP captions as input to the reasoning model.

Gemini + DeepSeek-R1 was selected as the final approach for English due to its consistent performance and better structural handling.

5.1.2. Arabic

The Arabic subset of EXAMS-V posed unique challenges such as right-to-left formatting, variable fonts, and missing answer labels. These issues made parsing and reasoning more difficult.

Experiment Performed are:

- **Multimodal Models: Qwen-VL** was tested on a small number of Arabic MCQs. It supports Arabic input and multilingual reasoning but can only be evaluated in small batches due to high memory demands.

- **Visual Parsing:** We evaluated three tools. **Tesseract OCR** was applied for Arabic text extraction but showed poor support for Right to Left formatting and frequently reversed question structure or answer order. **BLIP**, used in Arabic captioning mode, often dropped key scientific terms and failed to generate coherent sentence structure. **Gemini 1.5 Flash** demonstrated stronger Right To Left alignment, more accurate sentence preservation, and domain-aware parsing, making it more reliable for downstream reasoning.
- **Modular Pipeline:**
  - **Gemini + DeepSeek-R1 Distill LLaMA (Proposed):** Gemini was used to extract Right to left aware structured text, while DeepSeek-R1 handled reasoning with prompt adjustments to maintain sentence clarity.

Gemini + DeepSeek-R1 was selected as the final approach for improved layout parsing, and consistent reasoning performance.

5.1.3. Chinese

The Chinese subset of EXAMS-V presented the greatest difficulty across all languages. These MCQs are based on Gaokao exams and feature complex diagrams, scientific plots, mathematical tables, and domain-specific terminology. The visual complexity and abstract reasoning required make them particularly challenging for vision-language models.We tested several configurations:

- **Multimodal Models: Qwen-VL** was evaluated on a small subset due to high computational requirements. While it supports Chinese input and demonstrated strong multilingual capabilities, we could not scale its evaluation across the full test set.
- **Visual Parsing: Gemini 1.5 Flash** preserved structural layout more effectively, extracted numerical information reliably, and was better suited to the visual density of Chinese MCQs.
- **Modular Pipeline:**
  - **Gemini + DeepSeek-R1 Distill LLaMA (Proposed):** Gemini was used for structured OCR and layout parsing, while DeepSeek-R1 served as the reasoning engine for handling scientific and numeric logic.

Gemini + DeepSeek-R1 was selected as the final approach for Chinese due to its stability, structural accuracy, and robustness in handling visually complex, domain-specific content.

5.2. Experimental Setup

We run our experiments using a combination of APIs, inference servers, and cloud-based notebooks. Our final approach integrates two independent modules for parsing and reasoning.

- *Visual Parsing:* We use **Gemini 1.5 Flash**[1] via Google Cloud Vertex AI for OCR and layout-aware captioning. Gemini receives image inputs and extracts structured textual content from visual questions. The model was used with default generation parameters: `temperature = 1.0`, `top_k = 40`, `top_p = 0.95`, `max_output_tokens = 1024`, and no specified `seed` (non-deterministic behavior).

[1] https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/1-5-flash

- *Reasoning:* The extracted text is passed to **DeepSeek-R1 Distill (LLaMA-70B)**[2], which serves as our main reasoning engine. We use **Groq Inference Servers**[3] to execute this model with low-latency inference. The prompt strategy follows a few-shot, answer-only format for consistency and control. We explicitly set `temperature = 0` to ensure deterministic answers. Other hyperparameters were left at default.
- *Environment:* All processing was conducted in **Google Colab**[4]. Data handling and batch operations were implemented using `pandas`, `tqdm`, and Python's built-in `json` module.
  For image loading and preprocessing, we used the `PIL.Image` module. Prompt formatting and Gemini interaction were handled using the official `google.generativeai` SDK[5]. DeepSeek-R1 was accessed via **Groq Inference Servers**
  Additional runtime utilities such as `inspect.signature` were used to validate prompt structure and automate input formatting.

5.3. Evaluation Criteria for Model's Performance Evaluation

We use **accuracy** as the primary metric, following the official ImageCLEF 2025 evaluation protocol [1]. Our main pipeline Gemini (visual parsing) and DeepSeek-R1 Distill (reasoning) is submitted to the full multilingual test set. Accuracy is computed using the official leaderboard, which includes 14 sub-leaderboards: one per language and one for overall multilingual performance.

All other models and pipeline variants are evaluated on small, curated subsets of 20–50 MCQs per language. These tests are **exploratory**, aimed at understanding model behavior rather than producing benchmark scores. We focus on specific challenges such as layout parsing, symbolic reasoning, and cross-language variability.

While only one system is evaluated at scale, these targeted experiments provide valuable insights into where models succeed or fail—and why. Together, they offer a broader perspective on current limitations in multimodal reasoning across diverse visual and linguistic formats.

All models are tested in **zero-shot or few-shot settings**, without any fine-tuning.

6. Results and Analysis

We evaluate the Gemini + DeepSeek-R1-Distill-Llama-70B model on the `MBZUAI/EXAMS-V` dataset, tackling English, Chinese, and Arabic MCQs. With overall test accuracies of 0.8125 (English), 0.6560 (Chinese), and 0.4775 (Arabic), our model excels in structured, science-based tasks but faces hurdles with complex visuals and cultural nuances. We begin with qualitative insights into our model's performance, followed by language-specific results, a detailed comparison with alternative models, and an in-depth analysis of subject and visual element performance. Tables 3, 7, and 8 summarize findings, while Figures 2 and 3 visualize trends.

[2] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B
[3] https://groq.com/
[4] https://colab.research.google.com/
[5] https://pypi.org/project/google-generativeai/

6.1. Qualitative Insights into Our Model

Our model, blending Gemini's multimodal extraction with DeepSeek's robust reasoning, navigates the EXAMS-V dataset's diverse visuals and multilingual demands like a seasoned scholar. Table 3 captures its strengths, weaknesses, error patterns, and vivid scenarios, framing its performance across languages.

Table 3
Qualitative Insights into Gemini + DeepSeek Performance Across Languages

Language	Key Strengths	Key Weaknesses
English	Master Physics, text, graphs;	Struggles with dense figures;
Chinese	Strong in Math, Biology;	Weak in Chemistry, chemical structures
Arabic	Excels in chemical structures, tables;	Poor in graphs, Biology;

English thrives on science-heavy questions, leveraging robust training. Chinese excels in logical tasks like math but stumbles in the notation of Chemistry. Arabicrowess in tables contrasts with graph struggles, likely from script challenges. These insights guide our comparison with other models and our detailed analysis.

6.2. Results by Language

We analyze performance by language, integrating overall accuracies and baseline comparisons.

6.2.1. English

Performance Overview: Our model achieves 0.8125, tripling the baseline 0.2701. Physics scores 0.8125, with visual elements at 0.8125 (text), 0.625 (graphs), and 0.4468 (figures). (chemical structures — not in English test dataset).

Qualitative Insights: Gemini's extraction excels in text and graphs, likely answering a Physics question on projectile motion. Figures (0.4468) pose challenges, possibly misreading dense biology annotations. The weak visual processing of the baseline fails on the graphs, unlike the precision of our model. Errors include overcomplicating simple MCQs, but structured tasks shine.

Table 4
English Test Accuracy

Model	Accuracy
Gemini + DeepSeek-R1-Distill-Llama-70B (Ours)	0.8125
Baseline	0.2701

6.2.2. Chinese

Performance Overview: Chinese scores 0.6560, doubling the baseline's 0.2678. Subject-wise results show 0.7714 (Math), 0.73 (Biology), 0.65 (Physics), and 0.4706 (Chemistry). Visual elements are 0.6560 (text), 0.5397 (figures), 0.5122 (graphs), 0.5 (tables), and 0.4348 (chemical structures).

Qualitative Insights: DeepSeek's reasoning drives Math and Biology success, solving table-based algebra or figure-based questions. Chemistry (0.4706) and chemical structures (0.4348) lag, likely from

Table 5
Chinese Test Accuracy

Model	Accuracy
Gemini + DeepSeek-R1-Distill-Llama-70B (Ours)	0.6560
Baseline	0.2678

Gemini's notation misreads. The baseline's visual limitations contrast with our model's versatility. Errors stem from character recognition issues, but logical tasks excel.

6.2.3. Arabic

Performance Overview: Arabic achieves 0.4775, nearly doubling the baseline's 0.2703. Subjects include 0.65 (Chemistry), 0.5 (Math), 0.4868 (Physics), and 0.4028 (Biology). Visual elements show 0.8889 (tables), 0.875 (chemical structures), 0.4775 (text), 0.4567 (figures), and 0.2703 (graphs).

Qualitative Insights: Gemini's extraction excels in tables and chemical structures, mastering Chemistry questions. Graphs (0.2703) and Biology (0.4028) struggle, likely from script misalignment or cultural gaps. The baseline's poor visual handling falls short. Errors arise from graph misreads, but structured formats thrive.

‘

Table 6
Arabic Test Accuracy

Model	Accuracy
Gemini + DeepSeek-R1-Distill-Llama-70B (Ours)	0.4775
Baseline	0.2678

6.3. Qualitative Comparison with Alternative Models

To underscore why Gemini + DeepSeek outperforms alternatives, we compare qualitative performance on EXAMS-V's multilingual, visually complex MCQs.

Table 7 details the strengths, limitations and scenarios of the few-shot and small-batch testing, revealing the edge of our model.

Our model's synergy tackles EXAMS-V's challenges—multilingual text, graphs, tables, and chemical structures—more effectively than alternatives. Mistral's spatial reasoning falters, misinterpreting graph layouts (e.g., a Physics slope), while our model excels (English graphs: 0.625). VLaVA struggles with dense figures, like biology diagrams, unlike our model's moderate success (Chinese figures: 0.5397). Qwen-VL's resource demands slow it on structured visuals, such as Arabic tables, where our model shines (0.8889). DeepSeek-VL's hardware requirements make it impractical, unlike our model's accessibility. Gemini + Mistral-7B's shallow reasoning fails complex MCQs, whereas DeepSeek's logic drives success (Chinese Math: 0.7714). These contrasts highlight our model's adaptability to EXAMS-V's visual and linguistic complexity, though Arabic graphs and Chinese notations need refinement.

Table 7
Qualitative Comparison of Models on EXAMS-V Dataset

Model	Type	Strengths	Limitations
Gemini + DeepSeek (Best)	OCR + Reasoning Pipeline	Precise extraction of text, tables, graphs; robust science reasoning; moderate hardware needs	Weak on Arabic graphs due to script issues; struggles with Chinese Chemistry notations
Mistral	Vision-Language	Basic text processing in English; lightweight	Poor spatial reasoning; fails on graphs, figures; English-only
VLaVA	Vision-Language	Handles simple visuals like text boxes	Weak on complex visuals (e.g., graphs, dense figures); no multilingual support
Qwen-VL	Multilingual VLM	Supports English, Chinese, Arabic; decent text extraction	High computational cost; weak visual reasoning; slow on structured visuals
DeepSeek-VL	Multimodal VLM	Potential for end-to-end multimodal reasoning	Unusable on moderate hardware (>80 GB VRAM); unstable performance
Gemini + Mistral-7B	OCR + Reasoning Pipeline	Effective OCR via Gemini; handles simple visuals	Shallow reasoning; struggles with complex science MCQs

6.4. Performance by Subjects and Visual Elements

Table 8 details performance across subjects and visual elements, with – indicating absent elements, enriched with qualitative insights to unpack trends.

Table 8
Performance by Subjects and Visual Elements Across Languages

Language	Subjects				Visual Elements				
	Phys.	Bio.	Math	Chem.	Text	Chem. Struct.	Table	Figure	Graph
English	0.8125	–	–	–	0.8125	–	–	0.4468	0.625
Chinese	0.65	0.73	0.7714	0.4706	0.6560	0.4348	0.5	0.5397	0.5122
Arabic	0.4868	0.4028	0.5	0.65	0.4775	0.875	0.8889	0.4567	0.2703

Qualitative Insights: English's Physics dominance (0.8125) stems from Gemini's graph extraction, likely acing questions on projectile motion or energy plots, where clear axes and labels align with training data. The low figure score (0.4468) reflects struggles with dense annotations, as in biology diagrams with overlapping labels, possibly due to Gemini's OCR limitations. Chinese's Math (0.7714) and Biology (0.73) strengths showcase DeepSeek's logical reasoning, excelling at table-based algebra or figure-based genetics questions. However, Chemistry (0.4706) and chemical structures (0.4348) suffer

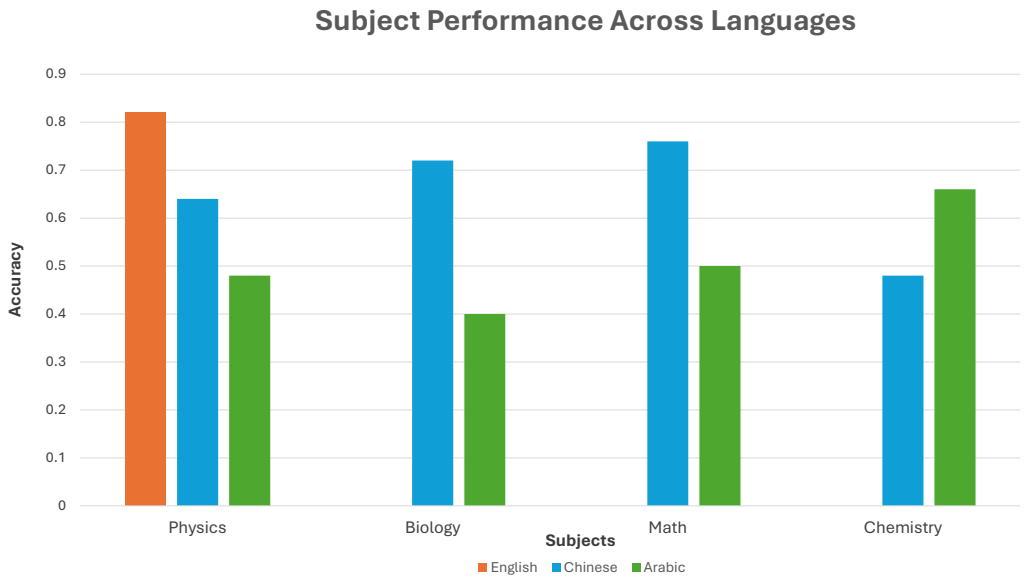

Figure 2: Subject performance across English, Chinese, and Arabic, highlighting strengths in Physics (English) and Math (Chinese), and weaknesses in Biology (Arabic).

from Gemini's misreads of complex notations, like mistaking a benzene ring's bonds, reflecting limited training on chemical symbols. Arabic's table (0.8889) and chemical structure (0.875) success highlights Gemini's structured data handling, mastering reaction tables in Chemistry. Graphs (0.2703) and Biology (0.4028) lag, likely from right-to-left script misalignment (e.g., misreading a mechanics graph's axes) or cultural gaps in Biology (e.g., unfamiliar terminology). These patterns contrast with the baseline's uniform visual struggles and alternatives' limitations (e.g., Qwen-VL's slow table processing). Figures 2 and 3 vividly illustrate these trends, guiding future improvements like script handling for Arabic graphs.

6.5. Why Our Model Outperforms the Baseline and Alternatives

Our model's superiority arises from Gemini's multimodal extraction and DeepSeek's reasoning, tailored to EXAMS-V's challenges:

- **Baseline**: Its low accuracies (0.2701–0.2703) reflect basic text processing, failing on visuals like graphs. Our model triples English accuracy (0.8125) and doubles Chinese (0.6560) and Arabic (0.4775), excelling in Physics (English: 0.8125) and Arabic tables (0.8889), like decoding a motion graph.
- **Alternatives**: Table 7 shows Mistral/VLaVA's spatial weaknesses, Qwen-VL's inefficiency, DeepSeek R1's hallucination, DeepSeek-VL's impracticality, and Gemini + Mistral-7B's shallow reasoning. Our model's efficient pipeline and robust logic handle Chinese Math (0.7714) and Arabic chemical structures (0.875) effectively.

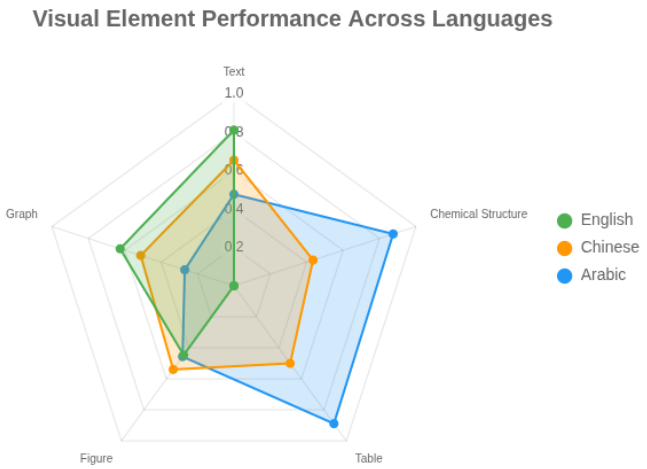

Figure 3: Visual element performance across languages, showcasing Arabic's strength in tables and chemical structures, and weakness in graphs.

This synergy minimizes errors, unlike the baseline's broad failures or alternatives' specific limitations.

6.6. Overarching Insights

Our model, like a multilingual science scholar, excels in structured tasks but needs coaching for complex visuals and cultural nuances. Key takeaways from Tables 3, 7, and 8, and Figures 2 and 3:

- **Strengths**: Dominates Physics (English: 0.8125), Math (Chinese: 0.7714), and Arabic tables (0.8889). Gemini's extraction and DeepSeek's reasoning outperform alternatives.
- **Weaknesses**: Struggles with Arabic graphs (0.2703), Chinese Chemistry (0.4706), and figures (e.g., English: 0.4468). Script and cultural gaps (e.g., Arabic Biology: 0.4028) pose challenges.
- **Error Patterns**: OCR misreads (e.g., Arabic script, Chinese notations) and reasoning gaps (e.g., cultural knowledge) drive errors, like misreading a molecular structure.
- **Improvements**: Fine-tune Gemini for script and notation handling; expand DeepSeek's chemistry and humanities training.

These insights highlight our model's adaptability and growth areas.

6.7. Example Output

Below is an example Figure 4 of the model's output for a sample exam-style image, demonstrating the extracted text from Gemini and the predicted answer from DeepSeek.

7. Conclusion

In this work, we introduced a modular Visual Question Answering (VQA) system designed for exam-style questions with complex visuals and multiple languages. Our approach uses Gemini to extract

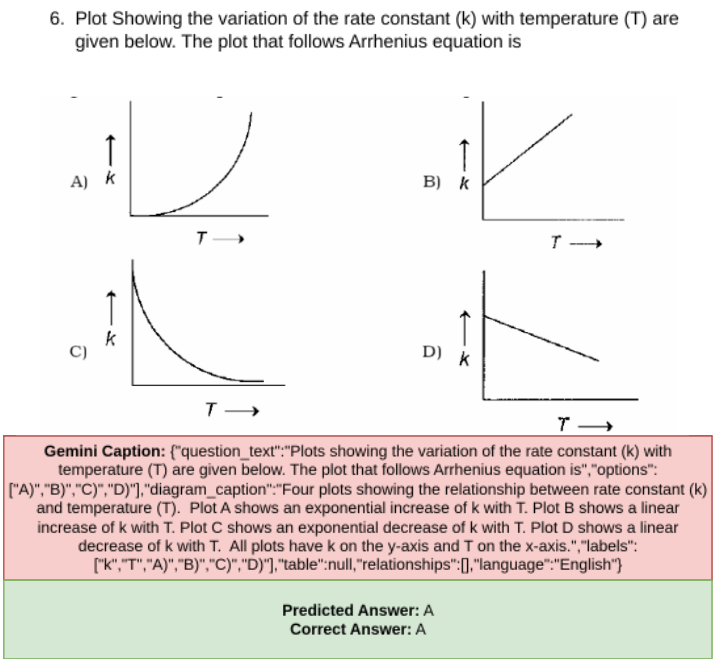

Figure 4: Sample exam-style image containing a multiple-choice question with graphs.

structured information from images and DeepSeek-R1-Distill-LLaMA to reason over that data and select the correct answer. This setup worked well on the EXAMS-V 2025 dataset, especially in subjects like Physics and Math, and handled English, Arabic, and Chinese without needing language-specific changes.

Our results show that the system performs better than existing models in understanding tables, diagrams, and multilingual questions. However, it still faces challenges with dense visuals (like Biology diagrams), right-to-left scripts in Arabic graphs, and special symbols in Chemistry.

Overall, this work shows that using a modular setup for visual question answering can be more flexible, accurate, and easier to improve. Future work can focus on improving image extraction for tricky visuals, adding more support for different languages, and making the system faster and lighter for real-world use.

Declaration on Generative AI

During the preparation of this work, the authors utilized tools such as ChatGPT and Grammarly to assist with grammar and spelling checks, as well as paraphrasing and rewording. All content generated or modified using these tools was subsequently reviewed and edited by the authors, who accept full responsibility for the final content of the publication.

References

[1] B. Ionescu, H. Müller, A.-M. Drăgulinescu, W.-W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, J. Rückert, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. A. Hicks, M. A. Riegler, V. Thambawita, A. M. Storås, P. Halvorsen, N. Papachrysos, J. Schöler, D. Jha, A.-G. Andrei, I. Coman, V. Kovalev, A. Radzhabov, Y. Prokopchuk, L.-D. Ştefan, M.-G. Constantin, M. Dogariu, J. Deshayes, A. Popescu, Overview of the imageclef 2023: Multimedia retrieval in medical, social media and internet applications, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2023, pp. 370–396.

[2] A. Agrawal, D. Batra, D. Parikh, A. Kembhavi, Don't just assume; look and answer: Overcoming priors for visual question answering, CoRR abs/1712.00377 (2017). URL: http://arxiv.org/abs/1712.00377. arXiv:1712.00377.

[3] R. Das, S. Hristov, H. Li, D. Dimitrov, I. Koychev, P. Nakov, EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 7768–7791. URL: https://aclanthology.org/2024.acl-long.420. doi:10.18653/v1/2024.acl-long.420.

[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, Vqa: Visual question answering, ICCV (2015).

[5] A. Kembhavi, M. Salvato, M. Seo, D. Schwenk, H. Hajishirzi, A. Farhadi, A diagram is worth a dozen images, 2016.

[6] M. Mathew, V. Bagal, R. Tito, D. Karatzas, E. Valveny, C. V. Jawahar, Infographicvqa, in: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2582–2591. doi:10.1109/WACV51458.2022.00264.

[7] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, A. Kalyan, Learn to explain: Multimodal reasoning via thought chains for science question answering, arXiv preprint arXiv:2209.09513 (2022).

[8] W. Zhang, S. M. Aljunied, C. Gao, Y. K. Chia, L. Bing, M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models, arXiv preprint arXiv:2306.05179 (2023).

[9] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, W. Chen, Mmmu: a massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, arXiv preprint arXiv:2311.16502 (2023).

[10] X. Yue, T. Zheng, Y. Ni, Y. Wang, K. Zhang, S. Tong, Y. Sun, B. Yu, G. Zhang, H. Sun, Y. Su, W. Chen, G. Neubig, Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark, arXiv preprint aarXiv:2409.02813 (2024).

[11] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, J. Gao, Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts, https://arxiv.org/abs/2310.02255 (2024).

[12] R. Das, S. Hristov, H. Li, D. Dimitrov, I. Koychev, P. Nakov, EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models, ACL (2024).

[13] D. Dimitrov, M. S. Hee, Z. Xie, R. Jyoti Das, M. Ahsan, S. Ahmad, N. Paev, I. Koychev, P. Nakov, Overview of imageclef 2025 – multimodal reasoning, 2025.