VQA

Here is the English translation of your content:

VQA:

VQA typically refers to **Visual Question Answering**, a research task at the intersection of computer vision and natural language processing. Its core objective is to enable models to understand an image along with a related question and generate an appropriate answer.

Datasets

Name	Summary	Data Volume	Link
VQAv2	Large-scale dataset covering multiple question types: binary (Yes/No), numerical, and open-ended	214k	https://huggingface.co/datasets/lmms-lab/VQAv2
textVQA	Focused on textual understanding within images, requiring models to read and comprehend text	34.6k	https://huggingface.co/datasets/lmms-lab/textvqa
documentVQA	OCR-based document understanding, including questions about tables and complex layouts	39.5k	https://huggingface.co/datasets/HuggingFaceM4/DocumentVQA
VQA-RAD + Path-VQA	Medical-focused dataset for radiology and pathology,	20k	https://huggingface.co/datasets/

Path-VQA	includes Yes/No and open-ended questions		flaviagiammarino/vqa-rad
			https://huggingface.co/datasets/flaviagiammarino/path-vqa
OK-VQA	Categorized by image type (e.g., plants, transportation), focuses mainly on open-ended questions	5k	https://huggingface.co/datasets/lmms-lab/OK-VQA
hiyouga/journeybench	Multi-image VQA dataset	—	https://huggingface.co/datasets/hiyouga/journeybench-multi-image-vqa
JourneyBench	Another variant of multi-image VQA	—	https://huggingface.co/datasets/JourneyBench/JourneyBench_Multi_Image_VQA

Model Leaderboards

1. HuggingFace OpenVLM Leaderboard

(Not integrated with the latest models; relatively outdated)
Leaderboard: https://huggingface.co/spaces/opencompass/open_vlm_leaderboard

Three sections:

- **TextVQA Evaluation Results**
  TextVQA evaluates visual reasoning based on textual elements in images. It requires models to read and understand embedded text.
- **OCRVQA Evaluation Results**
  Based on OCR-VQA-200K, this dataset comprises over 200k book cover images, combining OCR and VQA techniques.

- **MTVQA Leaderboard**
  (Content not currently displayable outside Feishu Docs)

---

2. Recent Top-performing Multimodal Models (VLMs)

Leaderboard: https://huggingface.co/spaces/opencompass/open_vlm_leaderboard

Model Name	Params (B)	Summary	Avg. VQA Score	Low-Score Count (<5)	Link
Qwen2.5-Omni	7	End-to-end multimodal model that handles text, images, audio, and video, with streaming text/speech output	7.36	108	https://huggingface.co/Qwen/Qwen2.5-Omni-7B
Qwen2.5-VL-7B	7		8.67	52	Qwen2.5-VL-7B-Instruct_low_score_rows.csv
InternVL2.5-26B	26		9.71	0	—
InternVL3-8B	8		8.32	68	InternVL3-8B_low_score_rows.csv
InternVL3-9B	9		8.63	28	InternVL3-9B_low_score_rows.csv
InternVL3-14B	14		8.69	43	InternVL3-14B_low_score_rows.csv
Google Gemma-3	27	Built using the same research and technology as Gemini, supports multimodal text and image processing	—	—	—
Ovis2 (16B / 8B / 4B)	Various		8.98–9.08	32–35	Ovis2_XB_low_score_rows.csv
Ola	7		—	—	—

Analysis for Ovis2-8B VQA Results

- **Average Score**: 9.08

- **Low-Scoring Items**: 32
- **Summary**: Strong image understanding, capable of recognizing details. Some low-score cases were too difficult even for humans. Human review suggests Ovis2-8B may be more accurate than GPT-based evaluators in certain cases.

Examples:

✅ **Good Cases**

Question	Image	Answer	Score
What is the child doing?	[img]	The child is sitting at a table with a group of people, likely enjoying a meal together.	10
What color is his shirt?	[img]	The man in the image is wearing a gray shirt.	10

❌ **Bad Cases (Error Analysis)**

Question	Image	Answer	Score	Analysis
What street is this?	[img]	This is Avenue Montaigne in Paris.	0	Query is too difficult; even humans may find it hard to determine.
What building is outside the window?	[img]	Historic Bank of Macon, Macon, Georgia.	0	Hard to confirm; answer cannot be reliably verified.
What sexual kink is represented?	[img]	Describes foot fetishism.	2	Misunderstanding of the intent behind the question.
Girls soccer team or coed team?	[img]	Suggests it's coed based on jersey colors.	2	Hard to confirm; inference might be incorrect.
What size bed is this?	[img]	Suggests queen size based on pillow size.	4	Difficult to determine without a reference; evaluation judged the answer inaccurate.
What bug is holding people aloft?	[img]	Describes a giant butterfly, but it was actually a parasol with butterfly features.	6	Misinterpretation of the image.
What kind of tree is this?	[img]	Suggests oak or similar broadleaf species.	6	Answer is too vague and possibly misleading.

---

3. **Most Downloaded VQA Models on HuggingFace**

Model	Summary	Link
BLIP	Developed by Salesforce. A unified vision-language pretraining framework for understanding and generation tasks.	https://huggingface.co/Salesforce/blip-vqa-base
		https://huggingface.co/Salesforce/blip-vqa-capfilt-large
ViLT	Lightweight vision-language model based on “Vision-and-Language Transformer Without Convolution or Region Supervision”	https://huggingface.co/dandelin/vilt-b32-finetuned-vqa