Visual Question Answering (VQA) - Model layer

Model zoo (Hugging Face) — practical picks

BLIP (VQA head)

- **Checkpoints:** `Salesforce/blip-vqa-base`, `Salesforce/blip-vqa-capfilt-large`
- **Why:** Strong open-ended answers; clean processor API; works well without heavy tricks.
- **Core idea:** Dual-encoder pretraining (ITC/ITM/LM) + VQA head; end-to-end on pixels (no region features).

BLIP-2 (ViT + Q-Former + LLM)

- **Checkpoints:** `Salesforce/blip2-flan-t5-xl`, `Salesforce/blip2-opt-2.7b`, `Salesforce/blip2-flan-t5-xxl`
- **Why:** Bridges vision features to an LLM via **Q-Former** → strong reasoning, good few-shot.
- **Core idea:** Freeze ViT + LLM, learn a small **querying transformer** (Q-Former) for efficient alignment.

InstructBLIP (instruction-tuned BLIP-2)

- **Checkpoints:** `Salesforce/instructblip-vicuna-7b`, `Salesforce/instructblip-flan-t5-xl`
- **Why:** Better follows prompts (“Answer concisely”, “use units …”); robust on diverse VQA styles.
- **Core idea:** Instruction tuning on mixed VQA/vision-lang corpora to improve controllability.

OFA (Unified Sequence-to-Sequence)

- **Checkpoints:** `OFA-Sys/ofa-base`, `OFA-Sys/ofa-large`
- **Why:** One seq2seq framework for many vision-language tasks (captioning, VQA, grounding).
- **Core idea:** Everything is text generation conditioned on visual tokens; multitask pretraining.

ViLT (Vision-and-Language Transformer)

- **Checkpoints:** `dandelin/vilt-b32-finetuned-vqa`

- **Why: No region detector** — patches + text tokens in a single transformer; lightweight & fast.
- **Core idea:** Early fusion of image patches + subword tokens; end-to-end pretraining objectives (MLM/ITM).

LLaVA / MiniGPT-4 (community MLLMs)

- **Checkpoints:** `liuhaotian/llava-v1.5-7b`, `liuhaotian/llava-v1.6-vicuna-7b`, `OpenGVLab/minigpt-4-v1_7b`
- **Why:** Chat-style VQA (multi-turn, chain-of-thoughty answers), strong zero-/few-shot on open images.
- **Core idea: CLIP/ViT visual encoder → projection → LLM** with visual-instruction tuning.

OCR-aware VQA (TextVQA / DocVQA)

- **Checkpoints:** `naver-clova-ix/donut-base-finetuned-docvqa`, `microsoft/layoutlmv3-base` (+ heads)
- **Why:** When reading text in images is essential (menus, receipts, signs).
- **Core idea:** End-to-end OCR-free (Donut) or layout-aware encoders (LayoutLMv3) + QA decoding.

(Legacy but notable) **LXMERT / UNITER / ViLBERT**

- **Checkpoints:** `unc-nlp/lxmert-base-uncased` (others often require conversion)
- **Why:** Classic **region-feature** (Faster R-CNN) + text fusion baselines; still useful for ablations.
- **Core idea:** Late-fusion with pre-extracted object regions ("bottom-up attention").

Architectural innovations to know

- **Region features → End-to-end pixels:** Older VQA used Faster R-CNN region proposals; newer (ViLT/BLIP) learn directly from patches.
- **Q-Former bridging (BLIP-2):** A small trainable transformer queries frozen vision features and speaks to a frozen LLM → efficiency + strong reasoning.
- **Instruction tuning for vision-language:** InstructBLIP/LLaVA align outputs to natural prompts and constraints.
- **Multitask seq2seq (OFA):** Unifies many tasks as text generation with visual conditioning.
- **OCR-aware pathways:** Either OCR-free (Donut) or OCR+layout (LayoutLMv3) for TextVQA/DocVQA.
- **Contrastive & matching pretraining:** ITC/ITM/MLM to align modalities (BLIP family, ViLT).
