Model (Provider/Platform)	**Input → Output**	**Access**	**Key/Signup**	**Free Limit**	**Quality/Notes**
BLIP-2 (Salesforce, HF)	Image (+ text prompt) → Text (caption, VQA)	Open-source (HF Transformers) / HF Inference	No for self-host; HF API key required	Self-host: unlimited; HF free tier limited	Strong zero-shot captioning & VQA. e.g. “impressively accurate” captions.
LLaVA (UCSD, HF)	Image + Text → Text (chat/VQA)	Open-source (HF Transformers)	No (self-host); HF API key	Self-host: unlimited; HF free tier limited	Multi-modal chat model; achieves SOTA on many vision-language benchmarks.
MiniGPT-4 (KAUST, HF)	Image + Text → Text (description, Q&A)	Open-source (HF Transformers)	No (self-host)	Self-host: unlimited	“GPT-4–like” vision-language model. Good at detailed descriptions, reasoning, problem-solving from images.
OpenFlamingo (MLFoundations, HF)	Image + Text → Text (caption, VQA)	Open-source (pip/HF)	No (self-host)	Self-host: unlimited	Open-source version of DeepMind’s Flamingo. 3B–9B models achieve strong benchmarks (e.g. 9B model gets CIDEr≈89.0 on COCO captions, VQA≈54.8%).
GIT (Microsoft)	Image + Text → Text	Open-source (HF Transformers)	No (self-host); HF API key	Self-host: unlimited; HF free tier limited	Decoder-only VLM. State-of-art on image captioning and VQA tasks; even surpasses human performance on TextCaps.
Qwen2.5-VL (Alibaba)	Image + Text → Text	Open-source (GitHub/HF)	No (self-host)	Self-host: unlimited	Latest flagship vision-language model. 3B–72B variants available. The 72B “Instruct” model rivals GPT-4V/Claude in vision tasks, excelling at document/diagram understanding. Smaller (7B/3B) models are also open.
Pix2Struct (Google)	Image (+ text prompt) → Text (structured)	Open-source (HF Transformers)	No (self-host); HF API key	Self-host: unlimited; HF free tier limited	Pretrained on screenshots. A single Pix2Struct model achieves SOTA on 6 of 9 vision-language tasks (docs, illustrations, UIs, etc.). Useful for captioning, VQA, UI parsing, etc.
TrOCR (Microsoft)	Image → Text (OCR)	Open-source (HF Transformers)	No (self-host); HF API key	Self-host: unlimited; HF free tier limited	Transformer OCR model for printed/handwritten text. Very high accuracy (≳95%) on standard benchmarks; far above legacy OCR (e.g. TrOCR ≈96% vs Tesseract ≈57% on receipt data).