ChatGPT

Free Multimodal AI Models and APIs

Below are key open models and APIs that take image (+ optional text) inputs and generate text (captions, OCR, VQA answers, etc.). All listed models are free to use (open-source or with free API access) and support both image and text input to text output. Some (like Hugging Face models) require a free account/API key for hosted inference; others can be self-hosted with no signup. Quality summaries (where available) cite recent benchmarks or examples.

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 [1].
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 [2].
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 [3].
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% [4]).

Model (Provider/ Platform)	Input → Output	Access	Key/ Signup	Free Limit	Quality/Notes
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 [5] .
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 [6] . 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.) [7] . Useful for captioning, VQA, UI parsing, etc.

Model (Provider/ Platform)	Input → Output	Access	Key/ Signup	Free Limit	Quality/Notes
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 [8]. Very high accuracy (≳95%) on standard benchmarks; far above legacy OCR (e.g. TrOCR ≈96% vs Tesseract ≈57% on receipt data) [9] [10].

Each model above supports image-to-text tasks (captions, OCR, VQA, etc.) and also handles text prompts. All can be downloaded and run locally (no signup/fees) given enough compute. Hugging Face "Spaces" or Inference API offers hosted use (requiring a free account/API key) with modest free-call limits. In summary, **BLIP-2, LLaVA, MiniGPT-4, OpenFlamingo, GIT, and Qwen2.5-VL** stand out as powerful free vision-language models [1] [6]. For example, BLIP-2 demonstrates impressively accurate zero-shot captions [1], while Qwen2.5-VL's 72B model matches state-of-art multimodal performance [6]. For OCR-specific tasks, TrOCR provides top-tier recognition [8] [9]. All of these can be accessed via open-source code or free API tiers (with usage limits on hosted services as noted).

Sources: All information above is drawn from the cited documentation and papers for each model [1] [2] [3] [4] [5] [6] [7] [8] [9], which detail their capabilities and performance. (The models' pages and papers were used to determine modalities, access methods, and benchmarks.)

[1] Zero-shot image-to-text generation with BLIP-2
https://huggingface.co/blog/blip-2

[2] LLaVa
https://huggingface.co/docs/transformers/en/model_doc/llava

[3] Minigpt-4
https://minigpt-4.github.io/

[4] GitHub - mlfoundations/open_flamingo: An open-source framework for training large multimodal models.
https://github.com/mlfoundations/open_flamingo

[5] GIT
https://huggingface.co/docs/transformers/en/model_doc/git

[6] Paper page - Qwen2.5-VL Technical Report
https://huggingface.co/papers/2502.13923

[7] Pix2Struct
https://huggingface.co/docs/transformers/en/model_doc/pix2struct

[8] TrOCR
https://huggingface.co/docs/transformers/en/model_doc/trocr

[9] [10] A Comprehensive Evaluation of TrOCR with Varying Image Effects - NHSJS
https://nhsjs.com/2024/a-comprehensive-evaluation-of-trocr-with-varying-image-effects/