Image Captioning - Model layer

This section covers the **model layer** for image captioning, focusing on architectures, key innovations, and practical implementations.

Core families & strong checkpoints on Hugging Face

BLIP / BLIP-2 (Salesforce)

- **What it is:** Vision encoder (ViT/CLIP) + text decoder; BLIP-2 adds a **Q-Former** to bridge a frozen LLM (OPT/FLAN-T5).
- **Why it matters:** Strong zero/few-shot, robust COCO performance, flexible fine-tuning.
- **Checkpoints:**
  - `Salesforce/blip-image-captioning-base` (fast, great starter)
  - `Salesforce/blip-image-captioning-large`
  - `Salesforce/blip2-opt-2.7b` / `Salesforce/blip2-flan-t5-xxl` (frozen LLM + Q-Former)

GIT — Generative Image-to-Text (Microsoft)

- **What it is:** ViT vision encoder + causal text decoder; trained at scale on web image–text.
- **Why it matters:** Simple, strong encoder-decoder baseline; competitive on COCO.
- **Checkpoints:** `microsoft/git-base`, `microsoft/git-large`

ViT-GPT2 (VisionEncoderDecoder)

- **What it is:** Generic **VisionEncoderDecoderModel** (ViT encoder + GPT-2 decoder).
- **Why it matters:** Small, teachable, easy to customize; great for didactics and quick PoCs.
- **Checkpoints:** `nlpconnect/vit-gpt2-image-captioning` (widely used tutorial baseline)

OFA — One For All (OFA-Sys)

- **What it is:** Unified seq2seq (BART-like) across captioning/VQA/grounding with multi-task pretraining.
- **Why it matters:** Strong multi-task transfer; good captioner with the right prompt format.
- **Checkpoints:** `OFA-Sys/ofa-base`, `OFA-Sys/ofa-large`

BEiT-3 (Microsoft)

- **What it is:** Multimodal masked pretraining unifying vision+language.
- **Why it matters:** Strong encoder for captioning/grounded tasks after fine-tuning.

- **Checkpoints:** `microsoft/beit-3-base`, `microsoft/beit-3-large` (finetune heads for captioning)

Multimodal LLMs usable for captioning (generalists)

- **Idefics2 (HF M4):** `HuggingFaceM4/idefics2-8b` — multi-image understanding & captioning/chat.
- **LLaVA (vision-chat):** `liuhaotian/llava-v1.6-vicuna-7b` — strong descriptive captions via prompts.
- **Qwen2-VL:** `Qwen/Qwen2-VL-2B-Instruct` (lightweight), `Qwen/Qwen2-VL-7B-Instruct` — good caption quality, long context.
- **mPLUG-Owl2 / InternVL (alt VLMs):** useful when you need OCR-ish or dense descriptions without task-specific heads.

> Document/receipt images: look at **Donut** (`naver-clova-ix/donut-base`) for OCR-style "captioning" of layouts.

Architectural innovations (what moves the needle)

- **Encoder–Decoder vs. Prefixing:** Classic **ViT/ResNet encoder + autoregressive decoder** (GIT, ViT-GPT2) gives controllable generation; CLIP-prefix or **Q-Former** (BLIP-2) efficiently bridges to frozen LLMs for better language priors.
- **Q-Former (BLIP-2):** Learns a small set of visual queries that summarize the image; **frozen LLM stays intact**, making training lighter and generalization strong.
- **Web-scale pretraining:** CC3M/12M, LAION, SBU drive vocabulary/rarity coverage; essential for **NoCaps**-style open-vocab.
- **Detection-augmented features:** Earlier SOTA (VinVL) fed **region features** (Faster R-CNN) into decoders—still helpful for dense/grounded captions.
- **OCR-aware fusion:** Inject detected text tokens/boxes for **TextCaps**; huge gains when scenes contain readable text.
- **Instruction-tuned VLMs:** LLaVA/Qwen2-VL/Idefics2 improve **descriptiveness & controllability** via prompts ("caption briefly", "list objects then describe").
- **Video captioning (brief):** Sample frames + **TimeSformer/X-CLIP** visual backbone + text decoder (or a VLM over frame sets). Same principles, but with temporal attention/pooling.
