Visual Question Answering (VQA) - Evaluation layer

Core metrics (what & when)

1) VQA “Soft Accuracy” (VQA v2, VizWiz, OK-VQA commonly)

- **What:** Consensus scoring with 10 human answers:

  $$ \text{acc} = \min\left(\frac{#\text{humans agreeing with pred}}{3},\ 1\right) $$

- **When:** Default for open-ended VQA benchmarks where multiple phrasings are acceptable.

- **Insight:** Rewards agreement with humans; robust to synonyms/typos after normalization.

2) Exact Match (EM) / Normalized EM

- **What:** Binary 0/1 after lowercasing, stripping punctuation/articles (“a/an/the”), collapsing whitespace.
- **When:** Tighter evaluation for closed-vocab, short answers (yes/no, numbers, colors) or internal QA checks.
- **Insight:** Precision of literal matching; great for unit tests and regression checks.

3) Token-level F1 (precision/recall on tokens)

- **What:** Overlap of predicted vs. gold tokens (SQuAD-style).
- **When:** Free-form answers with multiple tokens; complements EM.
- **Insight:** Partial credit for near-misses; sensitive to verbosity.

4) **ANLS** (Average Normalized Levenshtein Similarity) — TextVQA/DocVQA

- **What:** $\text{ANLS} = \text{avg}_i \max(0, 1 - \frac{ED(\hat{y}_i, y_i)}{\max(|\hat{y}_i|, |y_i|)})$
- **When:** OCR-heavy tasks where minor string diffs matter (menus, receipts, signs).
- **Insight:** Graded similarity score tolerant to small edit distances.

5) GQA diagnostics (beyond accuracy)

- **What: Accuracy**, **Consistency** (same reasoning → same answer), **Plausibility** (answer in vocabulary of image), **Validity** (well-formed), **Grounding**.
- **When:** Compositional reasoning or multi-hop datasets (GQA).
- **Insight:** Separates “got it right” from “understood it” (consistency/grounding).

6) Answerability / Unanswerable rate (VizWiz)

- **What:** Accuracy on “unanswerable” detection + standard accuracy on answerable subset.
- **When:** Real-world/noisy images; end-users need “I don’t know.”
- **Insight:** Avoids hallucinations; calibrates abstention.

7) Calibration metrics (ECE, Brier score)

- **What: ECE** (Expected Calibration Error) bins predicted confidence vs. empirical accuracy; **Brier** = mean squared error on probabilities.
- **When:** Human-facing systems; abstention thresholds; risk-sensitive apps.
- **Insight:** Are confidences trustworthy?

8) Efficiency (latency, throughput, memory)

- **What:** Tokens/s, images/s, peak VRAM.
- **When:** Productization & batch-serving.
- **Insight:** Sizing, cost, SLO compliance.

Visualization & inspection (how to “see” what the model used)

- **Attention rollout (ViT/CLIP/ViLT):** Aggregate self-attentions across layers → heatmap over image patches. *Use when:* Patch-based encoders; fast, model-native.
- **Cross-attention maps (BLIP-2/InstructBLIP):** Visualize Q-Former or decoder cross-attn weights onto image tokens. *Use when:* Explaining *why* an LLM concluded an answer.
- **Grad-CAM / Score-CAM (CNN backbones):** Class-activation style maps for answer logits or pre-answer heads. *Use when:* CNN-based encoders or region-feature pipelines.
- **OCR overlays (TextVQA/DocVQA):** Show recognized words + confidences; highlight words weighted by attention. *Use when:* Answers depend on text; quickly catches OCR failure.
- **Region/box visualization (legacy region-features / DETR):** Draw detected objects (labels/scores) used as inputs. *Use when:* Bottom-up attention or grounding analyses.
