Image Captioning - Evaluation layer

Core metrics (what they measure & when to use)

- **BLEU-1/2/3/4 (n-gram precision)**
    - *Use when:* You want quick, legacy comparability (older papers, small ablations).
    - *Insight:* Precision-oriented; penalizes missing common n-grams; weak on synonyms/paraphrases.
- **METEOR (unigram F, stem/synonym matching)**
    - *Use when:* You care about recall and soft-matching (morphology, WordNet synonyms).
    - *Insight:* More tolerant to paraphrase than BLEU; historically correlates better with humans than BLEU.
- **ROUGE-L (longest common subsequence)**
    - *Use when:* You want order-aware recall; often reported alongside BLEU/METEOR.
    - *Insight:* Captures sequence overlap without strict n-gram windows.
- **CIDEr / CIDEr-D (tf-idf weighted n-grams, consensus-based)**
    - *Use when:* **Primary COCO leaderboard metric**; multiple references available.
    - *Insight:* Rewards phrases common among human captions; down-weights generic words.
- **SPICE (scene-graph F1)**
    - *Use when:* You want **semantic quality** (objects, attributes, relations).
    - *Insight:* High correlation with human judgment on semantics; slower to compute.
- **SPIDEr (SPICE + CIDEr average)**
    - *Use when:* Balanced single score mixing **semantic** (SPICE) and **consensus fluency** (CIDEr).
    - *Insight:* Good overall selection metric for model picking.
- **BERTScore (semantic similarity via contextual embeddings)**
    - *Use when:* Paraphrases are common; domain captions vary lexically.
    - *Insight:* Token-level cosine similarity; robust to wording differences.
- **CLIPScore / RefCLIPScore (reference-free / reference-aware)**
    - *Use when:* You need **reference-light** evaluation or to detect image–caption alignment.

    - *Insight:* Measures vision–text alignment in a joint embedding space; complements text-only metrics.

- **Diversity/Novelty (Distinct-n, Self-BLEU)**
    - *Use when:* Checking mode collapse or repetitive outputs across a dataset.
    - *Insight:* Higher distinct-n → more lexical diversity; low Self-BLEU across corpus → diverse set.

> **Rule of thumb**
>
> - Report **CIDEr, SPICE (or SPIDEr)** as primary; include **BLEU-4**, **METEOR**, **ROUGE-L** for comparability; add **BERTScore/CLIPScore** for semantic/alignment checks. Always evaluate against **all references** per image.

Visualization methods (to understand *why* a caption was produced)

- **Grad-CAM / Grad-CAM++ on the vision encoder**
    - *What:* Localizes image regions influencing the encoder output.
    - *Why:* Verify that nouns/attributes mentioned are grounded in visual evidence.
- **Attention rollouts / cross-attention maps (encoder–decoder or Q-Former)**
    - *What:* Aggregate attention to visualize **token ↔ region** focus.
    - *Why:* Inspect which image patches support each generated word.
- **Token-level relevance overlays**
    - *What:* Show heatmap per generated token (e.g., “dog”, “red ball”).
    - *Why:* Helpful for error analysis (hallucinations vs grounded mentions).
- **Corpus-level dashboards**
    - *What:* Length histograms, novelty (distinct-n), per-category CIDEr (COCO categories), and hard-set breakdowns (e.g., NoCaps in/near/out-of-domain).
    - *Why:* Exposes brittleness and domain shift.
