Image classification - Evaluation layer

Core metrics (what + when)

- **Accuracy (Top-1 / Top-k)** *Use when:* Classes are balanced or roughly so; you want a simple overall hit-rate. *Insight:* Fraction of samples where the correct label is ranked top-1 (or within top-k).
- **Precision / Recall / F1 (macro / weighted / per-class)** *Use when:* Class imbalance matters or minority classes are critical. *Insight:* Trade-off between false positives and false negatives; macro treats classes equally, weighted respects support.
- **AUROC (macro, one-vs-rest) & AUPRC** *Use when:* Severe imbalance, or you care about ranking quality (threshold-free). *Insight:* How well the model separates classes across thresholds; PR is especially informative under imbalance.
- **Log Loss (Cross-Entropy)** *Use when:* You care about *probability* quality (not just correctness). *Insight:* Penalizes overconfident wrong predictions; good for calibration checks and early stopping.
- **Matthews Correlation Coefficient (MCC)** *Use when:* Robust single-number summary under imbalance. *Insight:* Correlation between predictions and labels; balanced and informative even if classes are skewed.
- **Calibration Error (ECE / MCE)** *Use when:* Downstream decisions rely on calibrated probabilities. *Insight:* How close predicted confidences are to empirical accuracies.

> For most workshops: report **Top-1**, **Top-5**, **macro-F1**, and **log loss**. If imbalance is pronounced, add **macro-AUROC** and a **reliability diagram (ECE)**.

Visualization methods (why + when)

- **Grad-CAM / Grad-CAM++ (CNNs, ConvNeXt)** *Why:* Localize the evidence for a prediction; sanity-check spurious correlations. *When:* Explaining a single prediction; model is convolutional or has conv-like final stages.
- **Attention Rollout / Attention Maps (ViTs, DeiT, Swin)** *Why:* Trace how information flows across transformer layers/heads. *When:* Transformer backbones; global context explanations.
- **Embedding Projections (t-SNE / UMAP) of penultimate features** *Why:* See class clusters, overlap, and outliers. *When:* Dataset diagnostics; curriculum design; failure analysis.

- **Confusion Matrix** *Why:* Identify which classes get mixed up. *When:* Always—fast, high signal.

(Bounding-box plotting is a detection-specific tool; for classification, focus on CAMs, attention, embeddings, and confusion matrices.)
