Video Classification - Evaluation layer

Core metrics (what to use and why)

- **Top-1 / Top-k Accuracy (single-label)** *Use when each clip has exactly one class*. **Insight:** How often the correct class is ranked #1 (or within the top-k, e.g., k=5). Top-k is useful when classes are visually similar.

- **F1 (macro / micro) & Precision/Recall (single- or multi-label)** *Use when class imbalance exists, or errors have asymmetric cost*. **Insight:**

    - **Macro-F1:** treats all classes equally → good for skewed datasets.
    - **Micro-F1:** aggregates over all instances → good for overall performance with imbalance.

- **mAP / Average Precision (multi-label)** *Use when clips can have multiple labels (e.g., Charades, EPIC actions)*. **Insight:** Area under Precision–Recall curve per class, then averaged → robust to threshold choice and strong under imbalance.

- **Balanced Accuracy** *Use when severe class imbalance but single-label classification*. **Insight:** Mean of per-class recalls → less biased toward dominant classes.

- **AUROC / AUPRC (diagnostics)** *Use for threshold-free comparison of probabilistic outputs; especially with heavy imbalance*. **Insight:** Ranking quality across thresholds; **AUPRC** more informative than AUROC under rare positives.

- **Calibration (ECE / Reliability)** *Use in production settings where scores drive decisions*. **Insight:** Are predicted probabilities aligned with empirical accuracy?

- **Clip → Video Aggregation** *Use when you evaluate on untrimmed videos (sample multiple clips)*. **Insight:** Report both **clip-level** and **video-level** (e.g., average logits or majority vote) to reflect deployment.

Visualization methods (to understand "why")

- **Grad-CAM / Grad-CAM++ (per-frame heatmaps)** Works with 3D CNNs or by applying to spatial blocks in video transformers. Visualize *where* the model looks in frames.

- **Attention rollout / attention maps (Transformers: TimeSformer/VideoMAE)** Aggregate attention across layers/heads to get spatiotemporal importance. Great for ViT-style models.

- **Saliency → Bounding boxes (diagnostic)** Threshold heatmaps and draw **proxy boxes** around the most salient regions. (Not the same as detection, but helps sanity-check focus.)

- **Per-class Confusion Matrix (single-label)** See which classes the model confuses; pair with **per-class PR curves** for multi-label.
