Video Eval Rig - LLM-as-Judge Build Checklist

Fora Soft Learn · AI for Video Engineering

1 - The five parts of the bench

```
Golden set      curated video test cases + known-good outcomes; stock the hard / weird / adversarial ones.
Feature test    your actual model + prompt, run against every case in the golden set.
Judge           an LLM (or, for video, a VLM) that scores each output against a written rubric.
Scorecard       a per-case pass/fail breakdown - where quality holds and fails, not one blurred average.
Gate            block the release if the score drops, a critical case regresses, or hallucinations rise.
```

2 - Why word-overlap metrics fail, and when you need a VLM judge

```
BLEU / ROUGE count shared words: they punish synonyms (a correct caption can score 0.11) and never watch.
LLM-as-judge scores meaning vs a rubric; in the seminal study a GPT-4 judge matched humans >80% (= human-human).
VLM-as-judge also reads SAMPLED FRAMES, so it can check a claim against the pixels and catch invented detail.
```

3 - Three scoring decisions

```
Shape       pointwise (score one output) to TRACK quality;  pairwise (A vs B) to CHOOSE a model or prompt.
Reference   reference-based (gold answer) offline on the golden set;  reference-free online in production.
Scale       always coarse (pass/fail or 1-5). Fine scales (1-100) add noise, not precision.
```

4 - Neutralize the three standing judge biases

```
Position bias     judges favor the first output  ->  grade twice, swap order, count a win only if it wins both.
Verbosity bias    judges reward longer answers    ->  put a length cap in the rubric; tell it to ignore length.
Self-preference   judges favor their own family   ->  use a judge from a DIFFERENT model family than the generator.
```

5 - Judge the judge (do this BEFORE you trust a single automated score)

```
Have humans label a few dozen golden-set cases; run the judge on the same cases; measure agreement.
Use Cohen's kappa (chance-corrected) + Spearman. Trust the judge only if agreement is high; precision is not accuracy.
```

6 - Before you ship the feature, ask:

- [ ] Is there a golden set of real video cases - including the hard and adversarial ones - the feature is scored against?
- [ ] Does the judge read the footage (VLM) when the output is a claim about video, so grounding is actually checked?
- [ ] Has the judge been calibrated against human labels (Cohen's kappa), not just assumed to be right?
- [ ] Are the three biases neutralized - swap-and-average order, a length cap, and a cross-family judge?
- [ ] Is the rig run offline as a release gate AND online as a 1-5% sampler of production traffic?
- [ ] Do production failures flow back into the golden set, so the rig compounds instead of decaying?
- [ ] Is the scoring mode chosen on purpose - pairwise to pick a model, pointwise to track one over time?
- [ ] Can the team show its evaluation evidence against a named framework (NIST AI RMF MEASURE / ISO/IEC 42001)?