PaddleOCR Deployment Checklist

PP-OCRv5 — map your video pipeline to the right model and avoid the six production failure modes.

1. Pick the right model

- [ ] CPU only or edge box? → PP-OCRv5_mobile_det + PP-OCRv5_mobile_rec (~8 MB, ~80 ms / 1080p frame)
- [ ] GPU (T4 / A10 / L4)? → PP-OCRv5_server_det + PP-OCRv5_server_rec (~100 MB, ~30-60 ms / frame)
- [ ] Multilingual feed? → Use the 106-language multilingual checkpoint, not the per-language one
- [ ] Export once to ONNX → serve through ONNX Runtime, OpenVINO, or Triton alongside other video AI models

2. Fix the six failure modes before going live

- [ ] Failure 1: Sampling every frame. Fix: run a motion detector or tracker first; OCR every 10th frame in each tracked region.
- [ ] Failure 2: Server model on CPU. Fix: match model to hardware. Mobile for CPU, server for GPU. Never cross the streams.
- [ ] Failure 3: Disabled direction classifier. Fix: keep use_textline_orientation=True. Costs 2 ms; saves 3 percentage points of accuracy.
- [ ] Failure 4: Small text at native size. Fix: 2x bicubic upscale before detection on burned-in subtitles. +8 ms, +5-10 points F-score.
- [ ] Failure 5: Per-frame reads without aggregation. Fix: confidence-weighted majority vote across all frames of the same tracked object.
- [ ] Failure 6: Hard-coded language. Fix: use the multilingual model with a per-track language-ID head; re-detect on scene change.

3. Pre-flight checks before going live

- [ ] Per-frame latency stays under your real-time budget (33 ms at 30 FPS, 16 ms at 60 FPS).
- [ ] GPU utilisation stays under 80% at peak load (room for traffic spikes and other AI stages).
- [ ] Aggregated string accuracy on hold-out test set is >= 95% character-level on real footage from the target camera.
- [ ] Language-ID accuracy is >= 98% on the multilingual hold-out set; misclassifications are logged for review.
- [ ] ONNX export is reproducible from the PaddlePaddle checkpoint; CI rebuilds it on each model refresh.
- [ ] PP-OCRv5 license note: Apache 2.0 commercial-use is fine; no per-page or per-frame fee.

4. Cost sanity check (2026 numbers)

- [ ] Self-hosted PP-OCRv5 on T4 GPU at ~$0.35/hour: ~$0.0006 per minute of 30-FPS video at 10 regions/frame.
- [ ] Google Document AI: $1.50 per 1,000 pages (right for batch document workloads, wrong for continuous video).
- [ ] AWS Textract: $1.50 per 1,000 pages (same reasoning as Document AI).
- [ ] Frontier VLM (Gemini / Claude Opus 4 / GPT-4o): ~$3-15 per 1,000 frames (right when context matters, slower).