

PaddleOCR Deployment Checklist

PP-OCRv5 — карта frame-rate, железа и языков под правильную модель и обход шести failure-mode.

1. Выбор правильной модели

- ☐ Только CPU или edge? → PP-OCRv5_mobile_det + PP-OCRv5_mobile_rec (~8 MB, ~80 ms / кадр 1080p)
- ☐ GPU (T4 / A10 / L4)? → PP-OCRv5_server_det + PP-OCRv5_server_rec (~100 MB, ~30-60 ms / кадр)
- ☐ Мультиязычный фид? → Использовать 106-языковой multilingual-чекпоинт, не per-language.
- ☐ Экспорт в ONNX один раз → ONNX Runtime / OpenVINO / Triton рядом с другими моделями.

2. Закрывать шесть failure-mode до прод-релиза

- ☐ 1: Сэмплинг каждого кадра. Фикс: сначала motion detector / tracker; OCR каждый 10-й кадр в tracked регионе.
- ☐ 2: Server-модель на CPU. Фикс: модель по железу. Mobile — CPU, server — GPU. Не смешивать.
- ☐ 3: Отключён direction classifier. Фикс: оставить use_textline_orientation=True. +2 ms, +3 пп точности.
- ☐ 4: Мелкий текст в нативном размере. Фикс: 2x bicubic upscale до детекции. +8 ms, +5-10 пп F-score.
- ☐ 5: Per-frame чтения без агрегации. Фикс: confidence-weighted majority vote по всем кадрам одного track.
- ☐ 6: Хардкод языка. Фикс: multilingual-модель + per-track language-ID head; re-detect на смене сцены.

3. Pre-flight перед выводом в прод

- ☐ Per-frame latency помещается в real-time-бюджет (33 ms при 30 FPS, 16 ms при 60 FPS).
- ☐ Утилизация GPU держится ниже 80% на пике (запас на всплески и другие AI-стадии).
- ☐ Точность строки после агрегации $\geq 95\%$ character-level на hold-out с реальной камеры цели.
- ☐ Точность language-ID $\geq 98\%$ на multilingual hold-out; ошибки логируются на ревью.
- ☐ ONNX-экспорт воспроизводим из PaddlePaddle-чекпоинта; CI пересобирает на каждый refresh модели.
- ☐ Лицензия PP-OCRv5: Apache 2.0 — коммерческое использование без платы за страницу или кадр.

4. Проверка стоимости (числа 2026)

- ☐ Self-hosted PP-OCRv5 на T4 GPU при ~\$0.35/час: ~\$0.0006 на минуту видео 30 FPS, 10 регионов/кадр.
- ☐ Google Document AI: \$1.50 за 1000 страниц (ок для batch документов, дорого для непрерывного видео).
- ☐ AWS Textract: \$1.50 за 1000 страниц (тот же резон, что и Document AI).
- ☐ Frontier VLM (Gemini / Claude Opus 4 / GPT-4o): ~\$3-15 за 1000 кадров (ок там, где нужен контекст).