

Чек-лист рецензента отчёта

Шесть вопросов, что превращают голую оценку обратно в утверждение — отметьте каждый перед решением.

Q1 · Что за метрика, чей инструмент?

Назовите метрику и реализацию, прежде чем читать число.

- ☐ Метрика названа (PSNR дБ / SSIM 0-1 / VMAF 0-100), не голое «0.95».
- ☐ Назван инструмент и версия (libvmaf, VQMT, FFmpeg) — они разные.
- ☐ PSNR/SSIM сравнивают лишь внутри одного инструмента и версии.
- ☐ Помните: 0.95 SSIM — это не 0.95 чего-либо ещё.

Q2 · Какая модель VMAF?

Модель — часть числа VMAF; её пропуск — обычный трюк.

- ☐ Модель названа — default 1080p@3H, phone или 4K@1.5H.
- ☐ Модель соответствует устройству, о котором отчёт.
- ☐ VMAF без модели — это $\pm 4-6$ пт, пока не спросишь какая.
- ☐ Телефонная модель читается на пункты выше default.

Q3 · Как сведено (пулинг)?

Одно заглавное число сводит тысячи кадровых оценок.

- ☐ Пулинг указан — среднее, гармон., медиана, мин или перцентиль.
- ☐ Нижний перцентиль (p5) или минимум показан рядом со средним.
- ☐ Среднее прячет короткую катастрофу: 59 с хорошо, 1 ужасно.
- ☐ Спрашивайте «как плох худший участок?», не только среднее.

Q4 · Какой разброс?

Оценка VMAF — прогноз; она несёт доверительный интервал.

- ☐ Указан 95% ДИ или stddev, не только точечная оценка.
- ☐ ДИ = оценка $\pm 1.96 \times stddev$ (напр. 75.44 $\pm$ 2.57).
- ☐ Для сравнения проверьте, перекрываются ли интервалы.
- ☐ «Победа» в 1 пункт внутри погрешности — шум, не результат.

Q5 · Яблоки к яблокам?

Сравнение верно, лишь когда всё, кроме теста, равно.

- ☐ Те же кадры, эталон, модель и пулинг в каждой колонке.
- ☐ То же разрешение — низкое оценено АПСКЕЙЛОМ до эталона.
- ☐ Нативная оценка низкого разрешения завышена расстоянием.
- ☐ Проверьте разрешение, на котором реально работала метрика.

Q6 · Контент или плохой эталон?

Каждая метрика где-то слепа; полнореференсной нужен исходник.

- ☐ Метрика применима? Бэндинг / зерно / темнота / потеря пакетов обманут.
- ☐ Эталон настоящий, выровнен по кадрам и верно помечен по цвету.
- ☐ У лайва или UGC нет исходника — полнореференс невозможен.
- ☐ Высокая оценка на сложном контенте — слепота, не качество.

Разбор примера — реальна ли победа в один пункт?

Энкодер А: VMAF 93.0 (95% ДИ 91.5–94.5). Энкодер В: VMAF 94.0 (95% ДИ 92.4–95.6). Интервалы перекрываются почти целиком, поэтому разрыв в 1 пункт внутри шума — честное чтение «при этом объёме выборки разница неизмерима», а не «В победил».

BD-rate — не оценка качества. «BD-rate –30%» значит на 30% меньше бит при равном качестве (отрицательный = выигрыш) — и лишь с названными метрикой, моделью, контентом и якорем.

Совет: видео, сравнённое само с собой, набирает ~98.7, не 100 — «не 100» не дефект. Запустите бесплатный линтер (vmaf-report-linter.py), чтобы оценить эти шесть вопросов автоматически.