

Worksheet Выбора Open-Vocabulary Детектора

Ответьте на восемь вопросов до выбора между Grounding DINO, Florence-2, OWLv2, RT-DETR и RF-DETR.

1. Бюджет латентности на кадр

- ☐ До 10 мс — только closed-vocab (RF-DETR-Large, RT-DETRv3).
- ☐ 10-50 мс — closed-vocab на хот-пути, OVD асинхронно.
- ☐ 50-200 мс — OWLv2 / Florence-2 на motion-гейте.
- ☐ Свыше 200 мс — допустим Grounding DINO 1.5 / 1.6 Pro.

2. Пол точности (COCO-эквивалент AP)

- ☐ Максимум (60+ AP) — RF-DETR-2XL (closed-vocab).
- ☐ SOTA zero-shot (55+ AP) — Grounding DINO 1.6 Pro.
- ☐ Хватит zero-shot (30-40 AP) — OWLv2 или Florence-2.
- ☐ Цель файнтюна — Florence-2 (мульти-задачный) или OWLv2.

3. Prompt-интерфейс — что печатает оператор

- ☐ Свободное описание / referring expression — Grounding DINO.
- ☐ Список class-фраз — OWLv2 или YOLO-World.
- ☐ Task-токен + текст — Florence-2 (`<OD>`, `<DENSE\_REGION\_CAPTION>`).
- ☐ Фиксированный класс-лист на обучении — RT-DETR / RF-DETR.

4. Лицензионный путь

- ☐ Apache 2.0 closed-source — RF-DETR core / OWLv2 / RT-DETR.
- ☐ MIT closed-source — Florence-2 (предпочтительно для коммерции).
- ☐ Grounding DINO 1.5 / 1.6 Pro — контракт IDEA Research API.
- ☐ Юридическая отмашка получена (неделя 1, не неделя 12).

5. Deployment-железо

- ☐ Cloud NVIDIA GPU (T4, L4, A10, A100, H100) — любая модель.
- ☐ Jetson AGX Orin — RF-DETR / RT-DETR или Grounding DINO Edge.
- ☐ Только CPU — RT-DETRv3-R18 или RF-DETR-Nano.
- ☐ Apple Silicon — Florence-2 (CoreML), без FlashAttention.
- ☐ Браузер (WebGPU) — YOLO-World или RF-DETR-Nano через WASM.

6. Мульти-задачный охват — за пределами детекции

- ☐ Только детекция — RF-DETR или RT-DETR.
- ☐ Детекция + сегментация — RF-DETR (встроено) или SAM 2.
- ☐ Детекция + OCR + captioning + grounding — Florence-2.
- ☐ Детекция + referring expressions — Grounding DINO.

7. План авто-разметки

- ☐ Двухуровневый пайплайн: closed-vocab + OVD-fallback.
- ☐ Слой prompt-шаблонов переписывает ввод оператора.
- ☐ Confidence-threshold тюнится для каждого шаблона.
- ☐ Авто-метки ревьюются до ретрейна (≥ 500 на класс).

8. Ритм ретрейна — refresh closed-vocab

- ☐ Месячный ретрейн на авто-размеченных данных (дефолт).
- ☐ Валидационный сет заморожен — нет утечек из авто-меток.
- ☐ A/B-тест: новая closed-vocab модель в shadow две недели.
- ☐ Rollback-путь определён при падении точности на старых классах.

Восьмиэлементная сводка

Бюджет латентности · пол точности · prompt-интерфейс · лицензия · deployment-железо · мульти-задачи · авто-разметка · ритм ретрейна.

Если хоть один из восьми не закрыт, решение по open-vocab детектору будет принято по неверной причине. Закройте все восемь до первого продакшен-промпта.