Оценка способности компактных визуально-языковых моделей интерпретировать указательные жесты в режиме zero-shot

Zero-shot · VLM · ERU

Гладкова Ксения Николаевна, аспирант 2 курса
Демяненко Яна Михайловна, к.т.н., доцент

ЮФУ · Институт математики, механики и компьютерных наук им. И.И. Воровича

СИТО 2026

Мотивация

Зачем понимать указательный жест?

- **Клиника и реабилитация —** пациенты с афазией или моторными нарушениями общаются через жест; система, понимающая указание, становится интерфейсом
- **Образование и тьюторинг —** учитель или ребёнок указывает на объект на экране / доске; модель должна понять что именно имеется в виду
- **Умный дом и ассистивные технологии —** «включи вот это», «принеси вон то» — жест устраняет неоднозначность без длинных команд

Указание пальцем — универсальный человеческий сигнал. Он появляется раньше речи и используется всю жизнь.

Постановка задачи

Embodied Reference Understanding (ERU)

Классическая задача
(Referring Expression Comprehension)

Текстовое описание + изображение

⬇

локализация объекта

✗ **Текстовый запрос о нужном предмете**
✗ **Жест и взгляд лишь уточняет запрос**
✓ **Реалистичные сцены взаимодействия**

Наша задача
(Embodied Reference Understanding)

Изображение с жестом

⬇

назвать объект указания

✓ **Без речевой подсказки**
✓ **Жест как пространственный фильтр**
✓ **Реалистичные сцены взаимодействия**

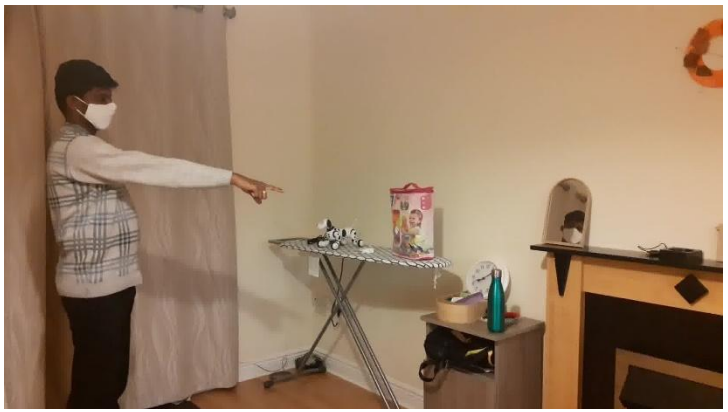
Датасет YouRefIt

Воплощённое понимание указаний в реалистичных сценах

YouRefIt

- ≈ 4 000 кадров из видеозаписей
- Помещения: кухни, офисы, гостиные
- Жест пальцем + речевая инструкция
- Аннотации: объект + bounding box
  - Загромождённые, реалистичные сцены
  - Много визуально схожих объектов
- Тестовая выборка: 1 251 кадр

Цель: «bag»



Цель: «water bottle»

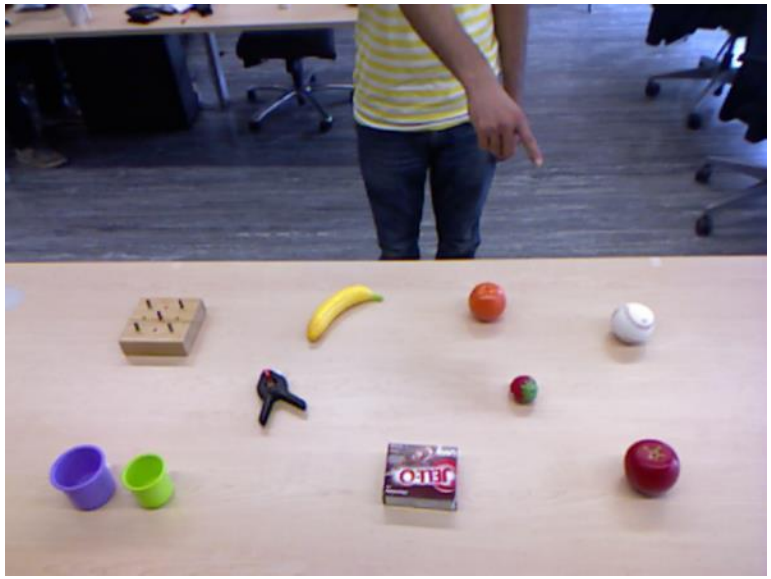
Цель: «cup»

Датасет Innsbruck Pointing (IBK)

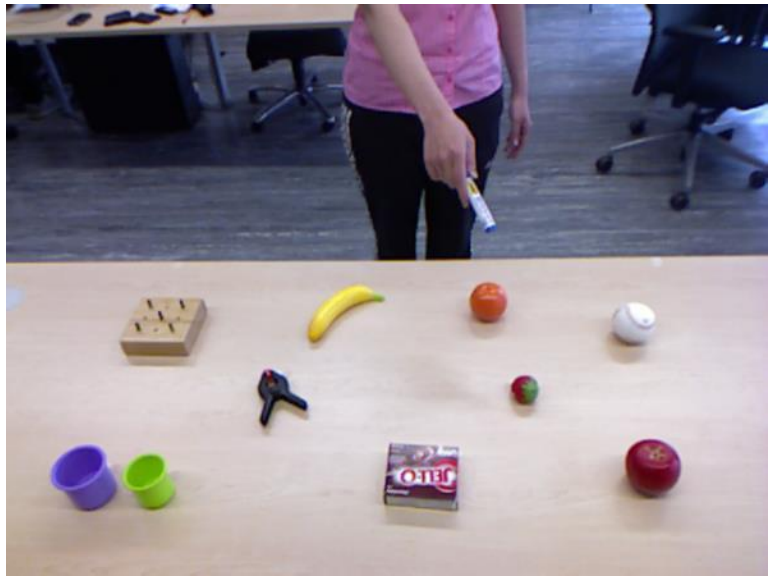
Лабораторный датасет с двумя типами указания

Innsbruck Pointing-at-Objects Dataset

- Статическая RGB-D камера
- 10 предметов на столе
- Два типа указания:
  - Естественное — указательный палец
  - Инструментальное — маркер
- Лабораторные условия:
  все объекты видны
  человек виден частично
- Выборка: 178 кадров, 10 сцен



Естественное указание



Инструментальное указание

Методология

Zero-shot тестирование компактных VLM

Тестируемые модели

Qwen

Gemma

moondream 2

Метрики

Exact Match (Accuracy Exact)
Точное текстовое совпадение с меткой

Semantic Match (Accuracy Sem)
Косинусное сходство ≥ 0.7

Zero-shot Prompt

What object is the person pointing at in this image?
Answer with only the object name, nothing else.

Результаты

Точность моделей на двух датасетах (zero-shot)

Модель	IBK Pointing Dataset		YouRefIt Dataset	
	Exact Match	Semantic Match	Exact Match	Semantic Match
Moondream2 1.8B	0.6%	0.6%	0.1%	0.5%
Gemma4 2B	12.9%	12.9%	3.0%	7.3%
Gemma4 4B	**21.4%**	**21.4%**	3.8%	7.6%
Gemma3 4B	9.6%	9.6%	3.8%	8.6%
Gemma3 12B	12.4%	12.4%	4.6%	9.2%
Qwen3-VL 2B	12.4%	12.4%	**7.3%**	**16.6%**
Qwen3-VL 4B	12.4%	12.4%	**7.1%**	16.2%
Qwen3-VL 8B	**14.0%**	**14.0%**	**7.3%**	**17.6%**

Результаты — Grounding-режим

Локализация объекта (bounding box) — только Qwen3-VL

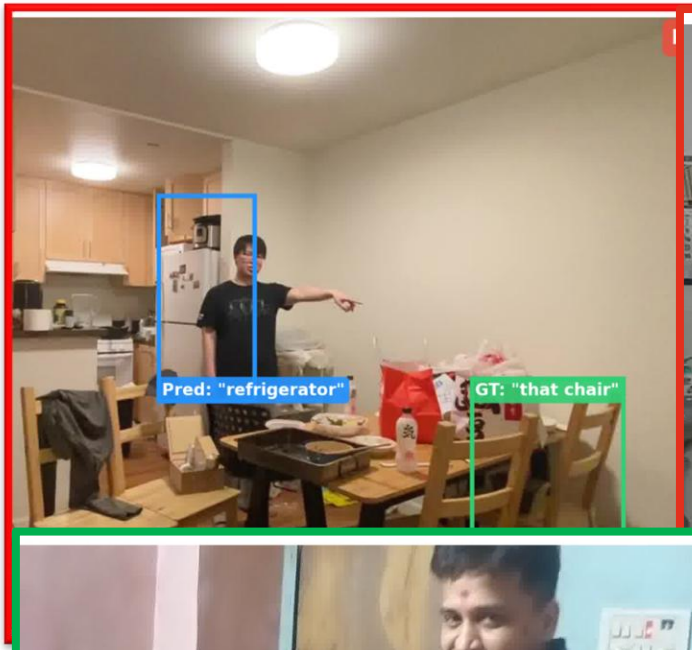
Модель	IBK Pointing Dataset		YouRefIt Dataset			
	Semantic Match	Point-in-Box Accuracy	Semantic Match	Mean IoU	Accuracy @ IoU≥0.5	Accuracy @ IoU≥0.25
Qwen3-VL 2B	5.1%	8.4%	13.5%	17.7%	17.9%	**22.5%**
Qwen3-VL 4B	**7.3%**	**13.6%**	**16.3%**	**18.4%**	**19.3%**	22.2%
Qwen3-VL 8B	**13.5%**	**20.8%**	**16.1%**	**19.5%**	**20.9%**	**23.5%**

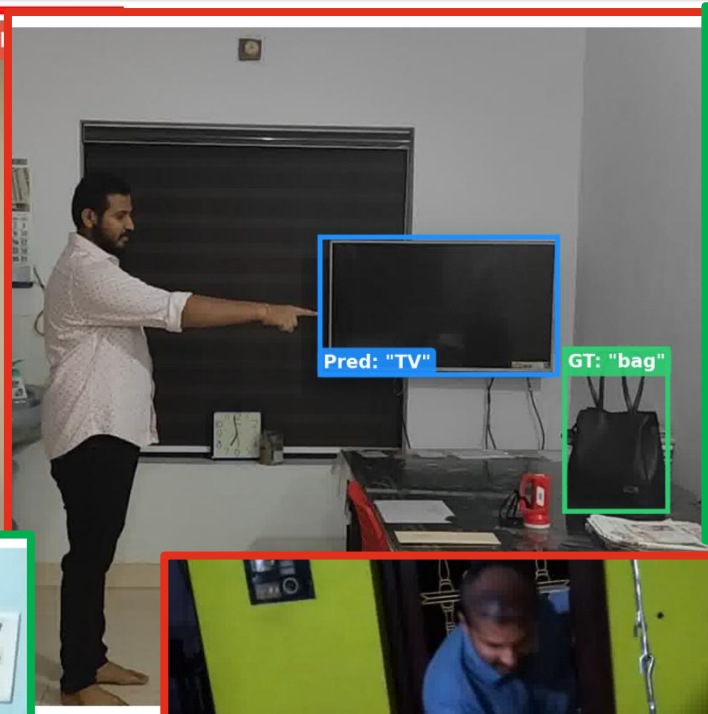
IBK: Point-in-Box — GT это 2D-координата объекта; предсказание верно, если точка попадает в predicted bbox.

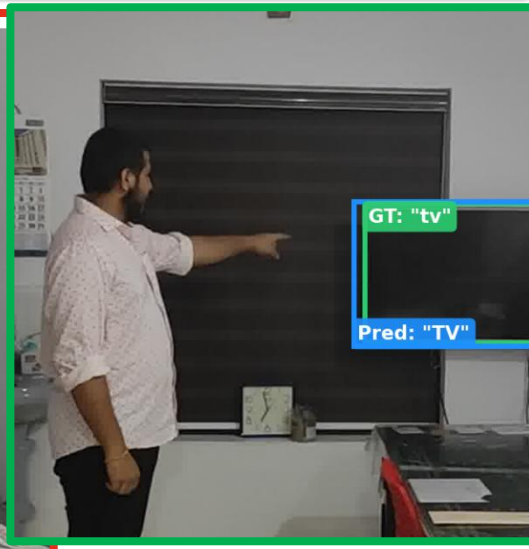
YoureIt: IoU — GT bbox vs predicted bbox. Acc@0.5 / Acc@0.25 — доля с IoU ≥ порога.

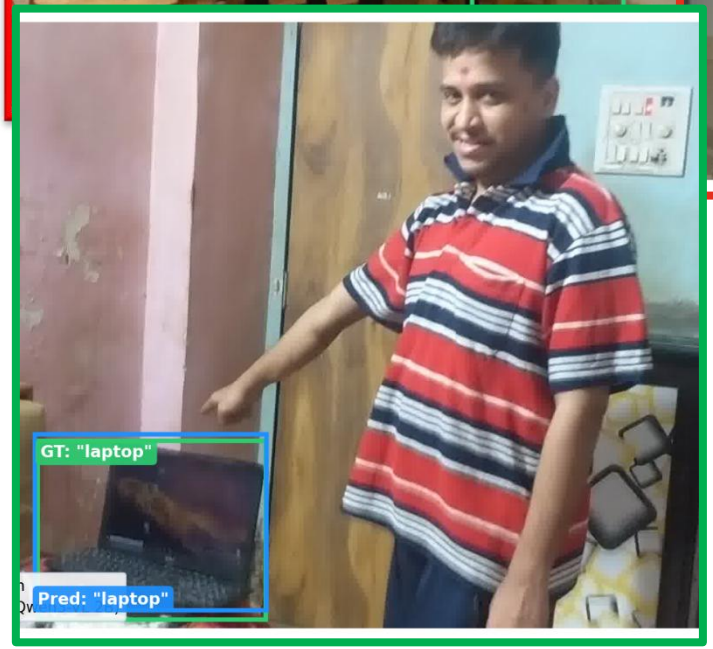
Примеры предсказаний

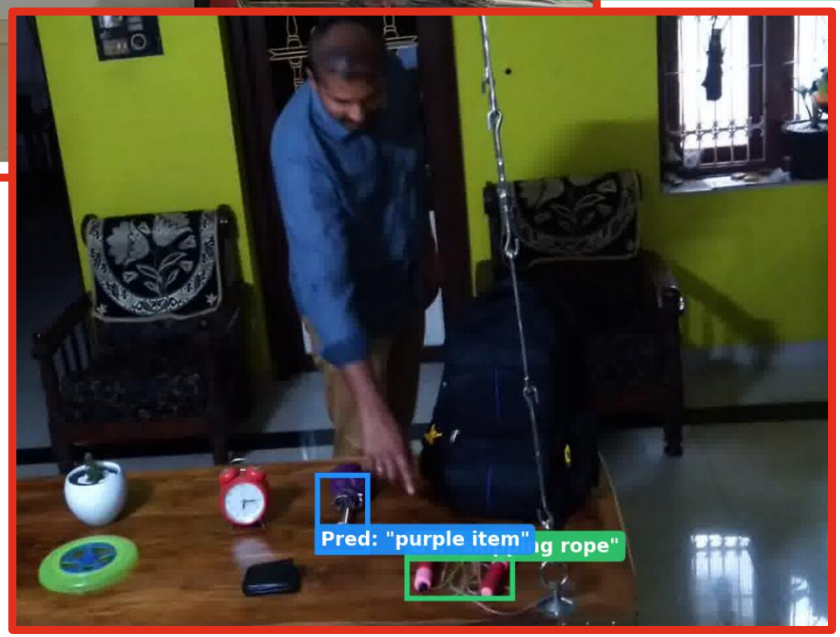
YouRefIt qwen3-VL:2B





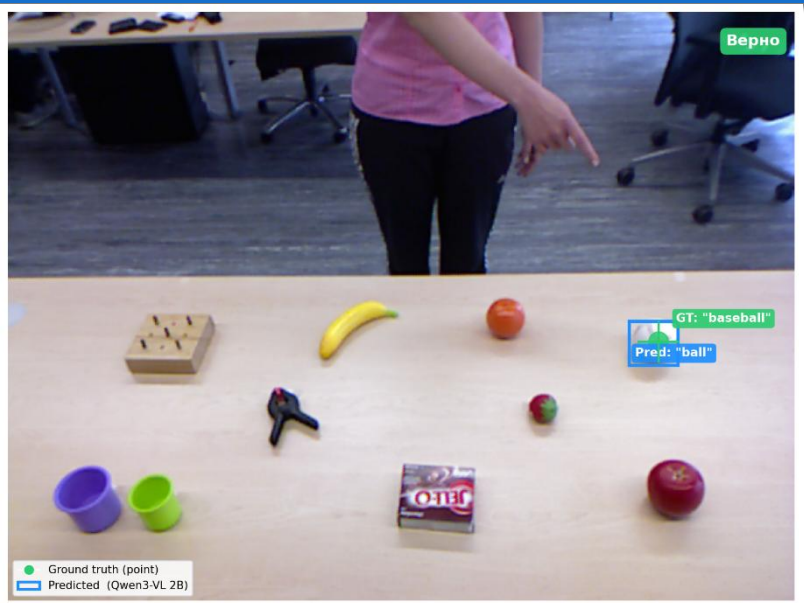


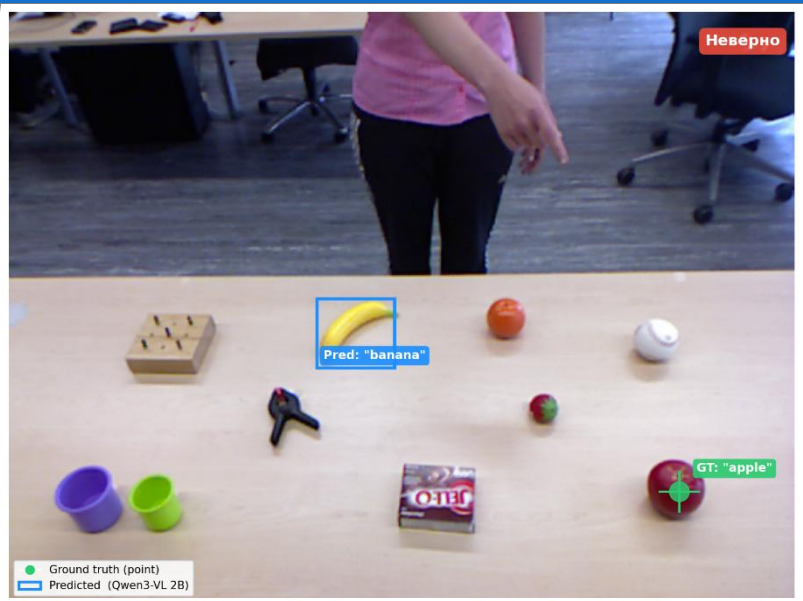


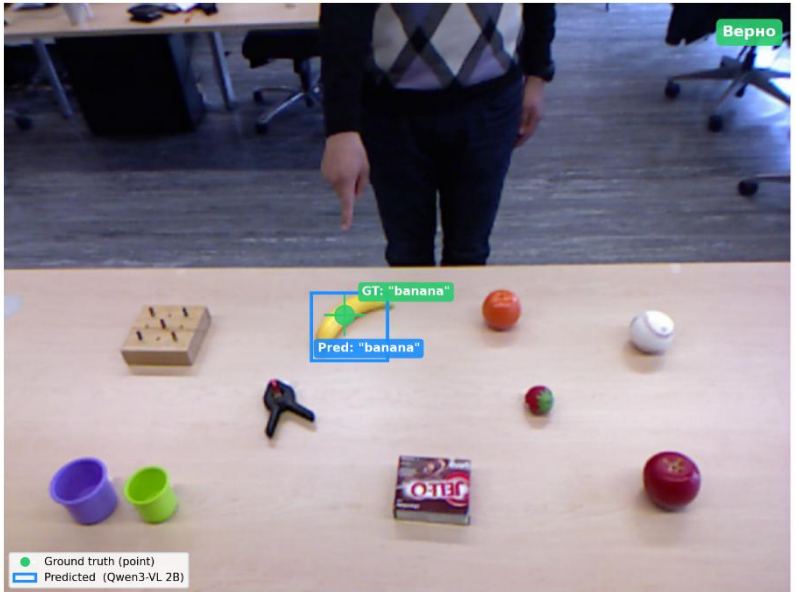


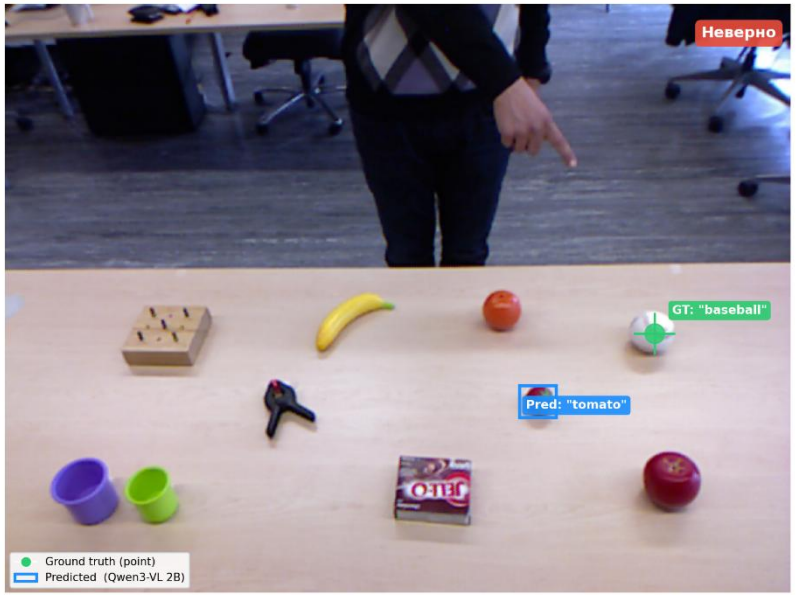
Примеры предсказаний

Innsbruck Pointing-at-Objects Dataset









Анализ ошибок

Характерные паттерны неверных предсказаний

Рука вместо цели

Модель называет объект, ближайший к руке или инструмент указания, а не объект в направлении жеста.

Визуальная заметность

Выбирается наиболее крупный или яркий объект на сцене, а не тот, на который указывают.

Путаница схожих объектов

Модель не различает несколько семантически или визуально схожих предметов в кадре.

Модели воспринимают жест частично:
они реагируют на присутствие руки в кадре, но не умеют корректно экстраполировать линию взгляда/пальца до удалённого объекта.

Выводы и перспективы

Выводы

- Компактные VLM недостаточны для надёжной zero-shot интерпретации жестов
- Лабораторные сцены (IBK) интерпретируются лучше загромождённых (YouRefIt)
- Жест несёт информацию, но модели не умеют её извлекать

Перспективы

- Дообучение лёгких VLM на пространственные намерения человека
- Создание датасетов разного уровня сложности (от лаборатории до реальных сцен)
- Совместный учёт геометрии жеста и семантики сцены

Оценка способности компактных визуально-языковых моделей интерпретировать указательные жесты в режиме zero-shot

Zero-shot · VLM · ERU

Гладкова Ксения Николаевна, аспирант 2 курса
Демяненко Яна Михайловна, к.т.н., доцент

ЮФУ · Институт математики, механики и компьютерных наук им. И.И. Воровича

СИТО 2026