CVPR JUNE 3-7, 2026 DENVER COLORADO

AVA-Bench: Atomic Visual Ability Benchmark for Vision Foundation Models

Zheda Mai[1*], **Arpita Chowdhury**[1*,]
Zihe Wang[1], Sooyoung Jeon[1], Lemeng Wang[1], Jiacheng Hou[1], Jihyung Kil[2], Wei-Lun Chao[3]

[1]The Ohio State University, [2]Adobe Research, [3]Boston University

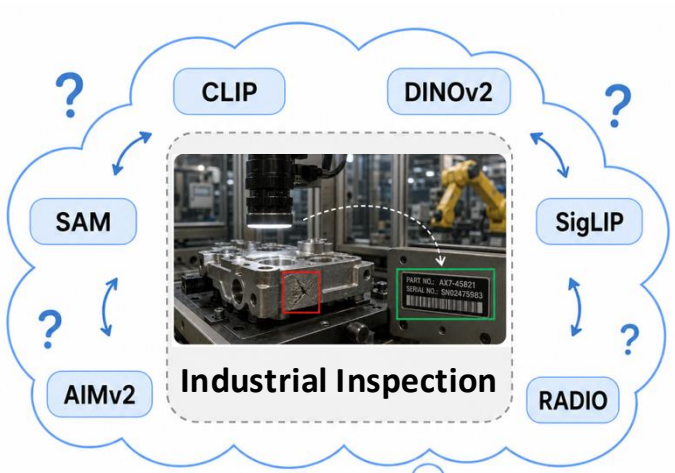
Motivation

- Vision Foundation Models (VFMs) with a variety of pre-training **objectives** and **supervision signals**

- A proliferation of **specialized** VFMs with different **strengths** and **weaknesses**

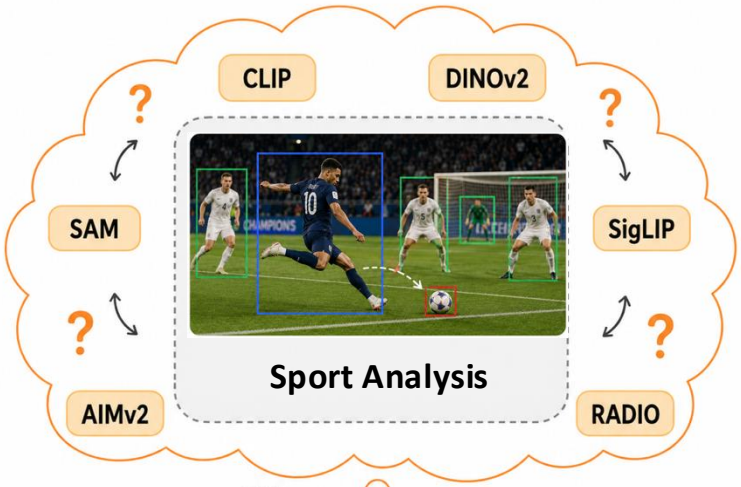
Motivation

- As **specialized** VFMs has **different** **strengths** and **weaknesses**.
- This diversity creates a practical problem: *which VFM should I use for my specific task?*
- A systematic and effective **evaluation protocol** for VFMs is increasingly crucial!



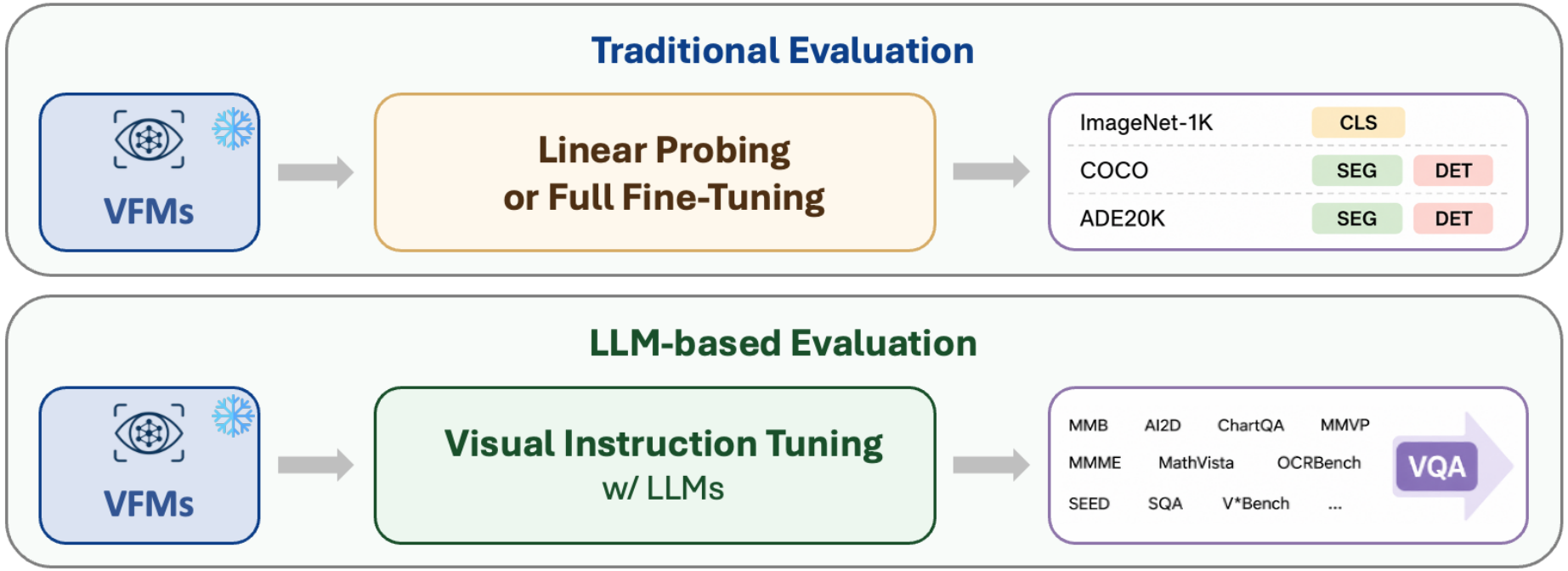
Industrial Inspection Engineer

Retail Shelf Analyst



Sports Analyst

VFM Evaluation



- Traditional paradigm: different tasks necessitate **task-specific heads** for linear probing or full fine-tuning

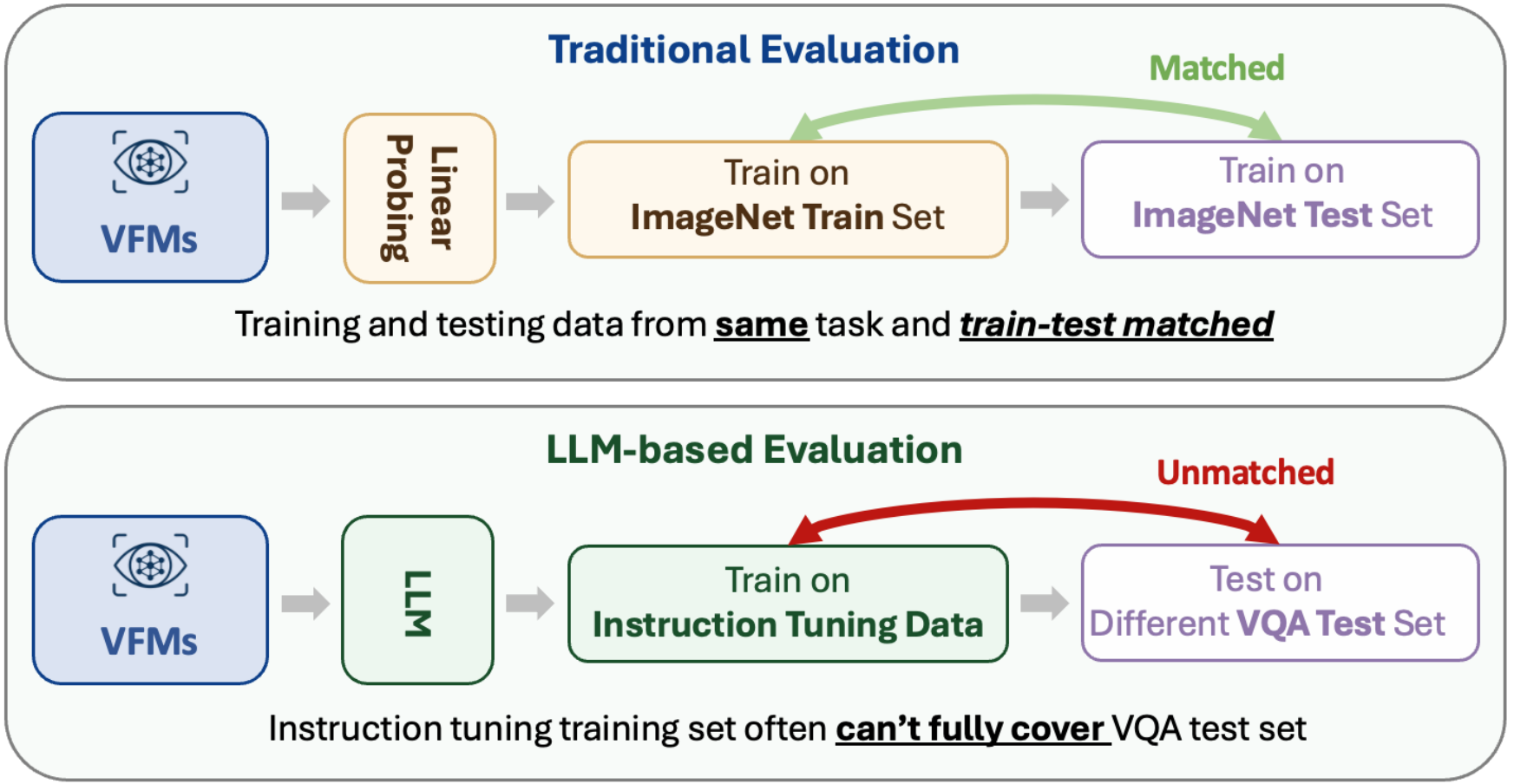
- LLM as a general-purpose heads:
  - **Visual Instruction Tuning** to train MLLM
  - Diverse Visual Question Answer (**VQA**) to evaluate VFMs

Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. Tong, etc. NeurIPS24.

Two Blind Spots of Evaluation

LLM-based Evaluation

A wrong prediction might arise from train-test mismatch rather than **genuine visual deficiencies** in a VFM.



Two Blind Spots of Evaluation

How many yellow dogs are facing backward on the left-hand side of the stop sign?

Atomic Visual Abilities

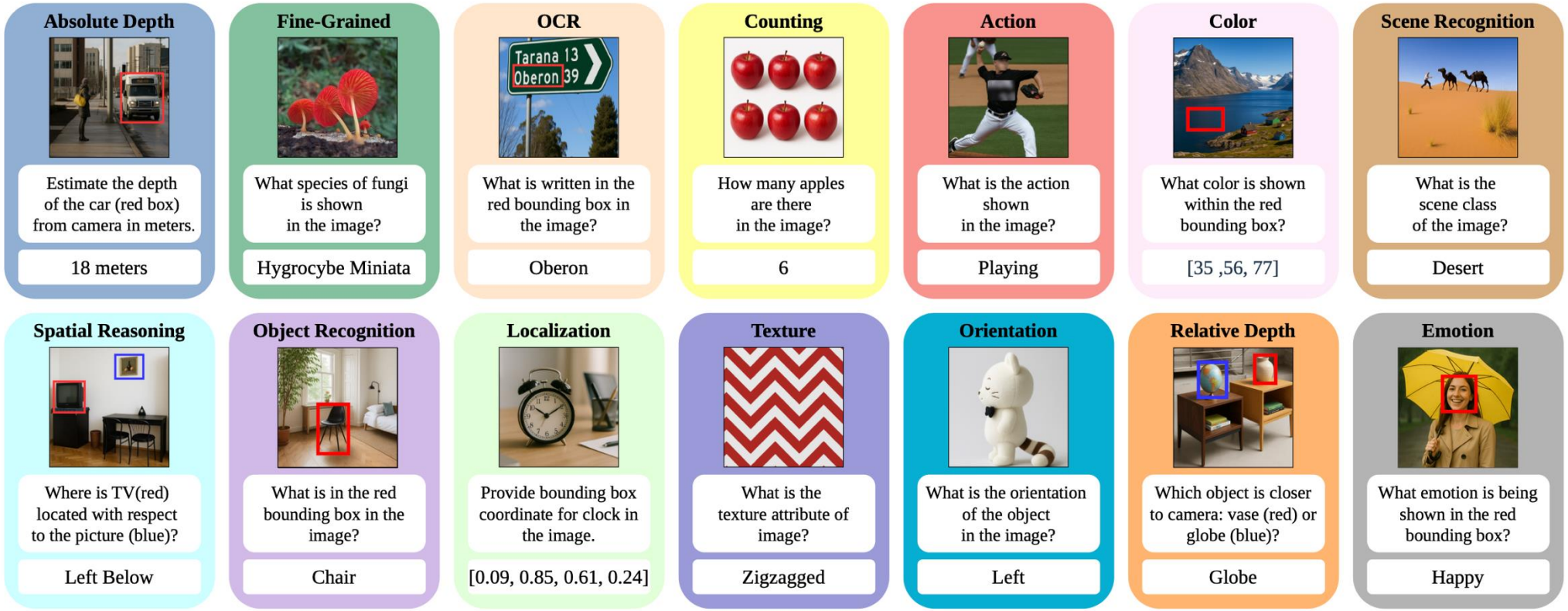
Orientation	Fine-Grained	OCR	Relative Depth	Counting	Object Recognition	Action
Texture	Color	Emotion	Absolute Depth	Localization	Spatial	Scene Recognition

VQA questions often require **multiple** visual abilities simultaneously, making it difficult to determine whether a failure arises from the **absence of multiple abilities or merely a single critical one**

AVA-Bench

AVA-Bench

- First systematic evaluation explicitly **disentangling** 14 Atomic Visual Abilities (AVAs) for VFMs.
- AVAs are **fundamental perceptual capabilities** that can be combined to address more complex visual reasoning tasks.

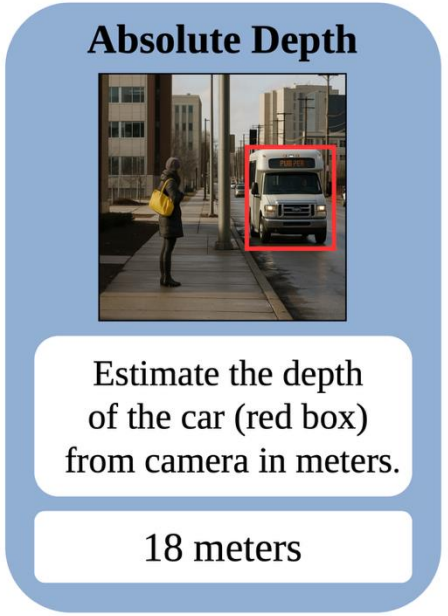


AVA-Bench

- Each AVA comes with **train-test-matched** data

- Given an AVA, image–question pairs were carefully designed to **test only on that AVA**

Eliminate the two blind spots, allowing AVA-Bench to **pinpoint exactly where a VFM excels or falters**



- Involves both localization and depth estimation;

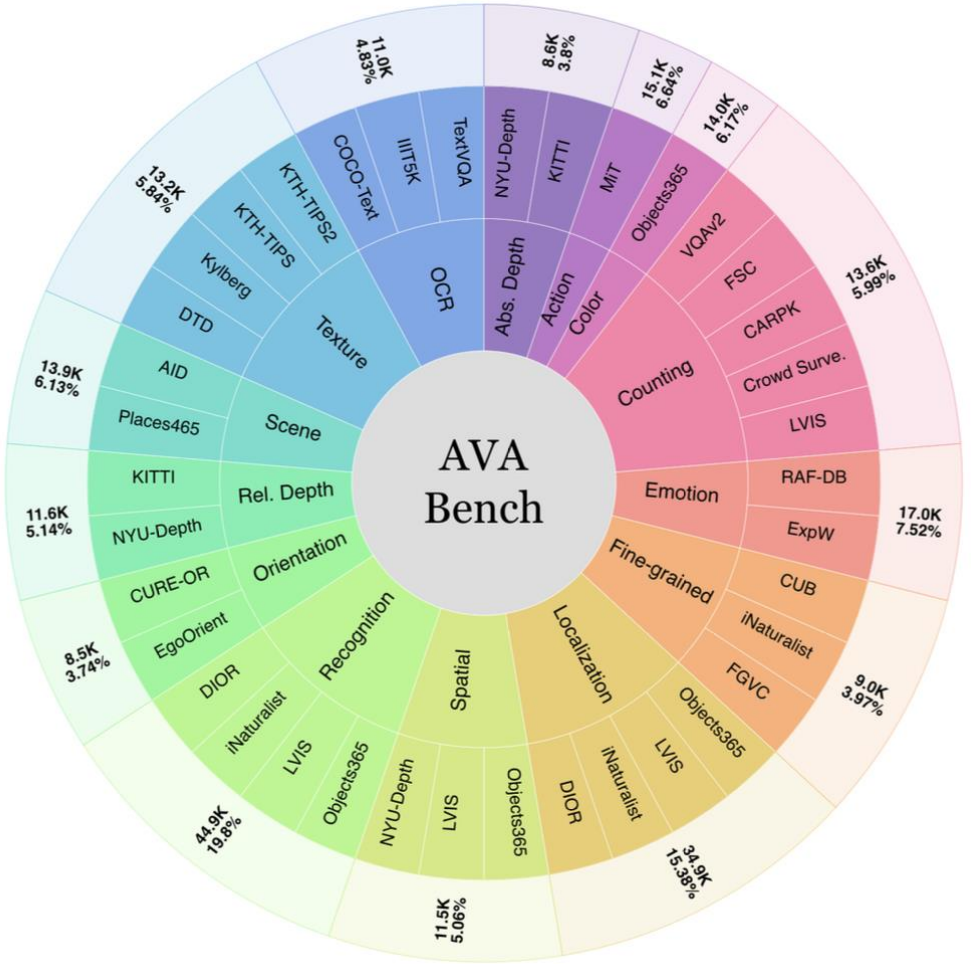
- Explicitly provide car's **bounding box** to test only depth estimation.

AVA-Bench

14 AVAs

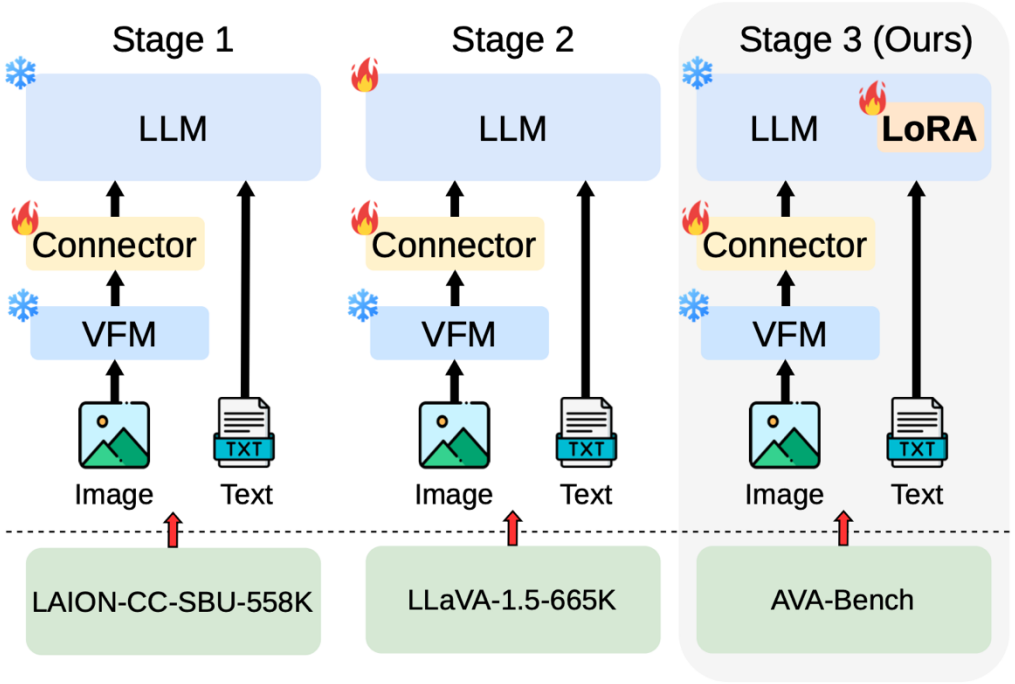
26 source datasets

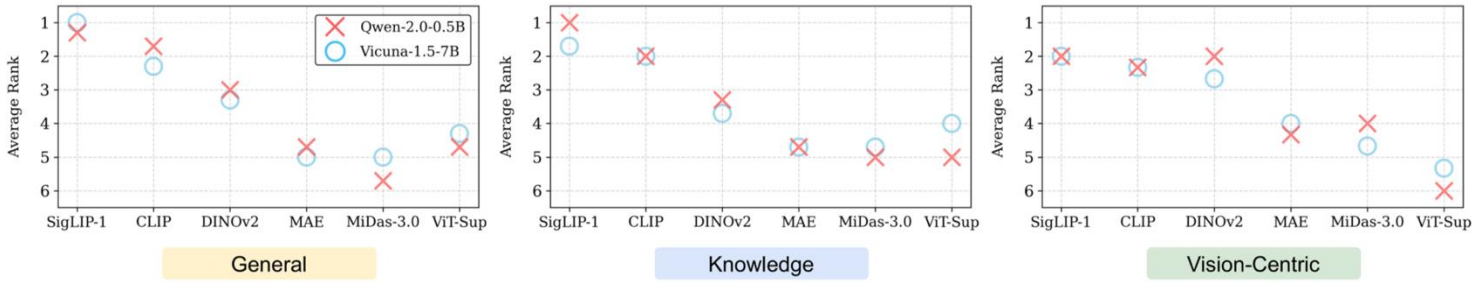
218K image-question pairs



Carefully control dataset balance, object visibility, and annotation biases.

Evaluation Pipeline





A heavyweight LLM may **NOT** be mandatory for reliable comparative evaluations.

0.5B

Qwen2 LLM

≈8×

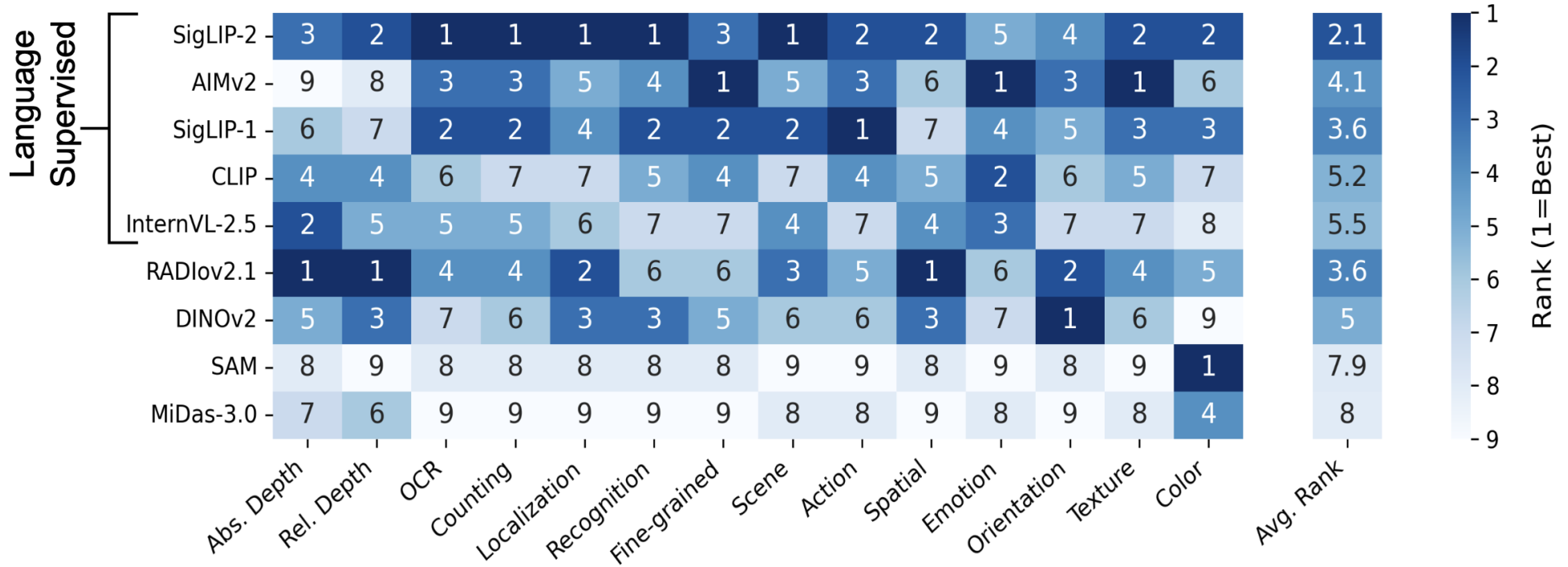
lower GPU hours

Preserves **similar relative VFM rankings** to a 7B Vicuna.

Liu, Haotian, et al. "Visual instruction tuning." NeurIPS (2023).

Each VFM Has an Ability Fingerprint

Ranked AVA performance reveals strengths that a single score would hide.

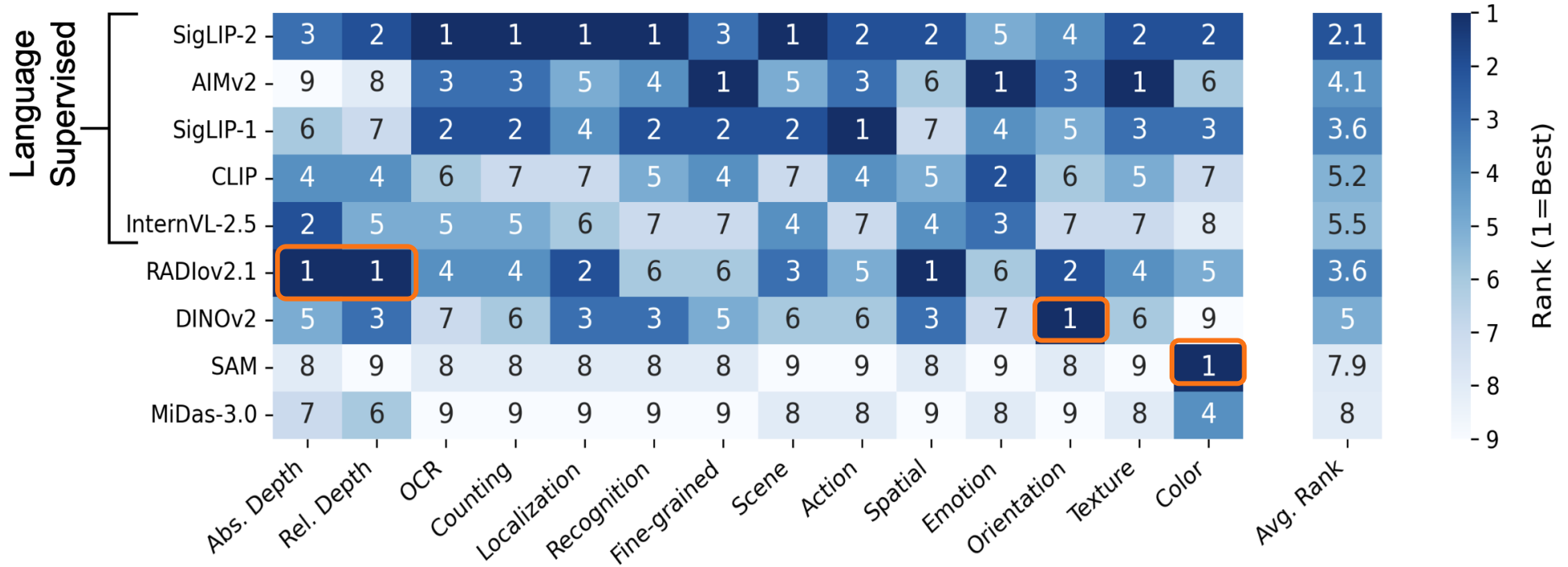


Each VFM has an **AVA fingerprint**

Language-supervised VFMs excel broadly across AVAs

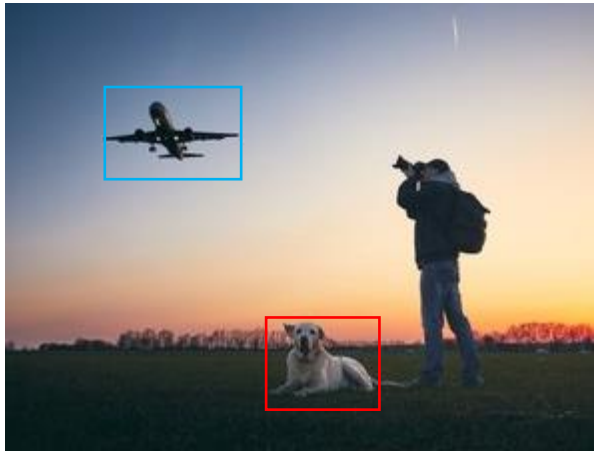
Each VFM Has an Ability Fingerprint

Ranked AVA performance reveals strengths that a single score would hide.



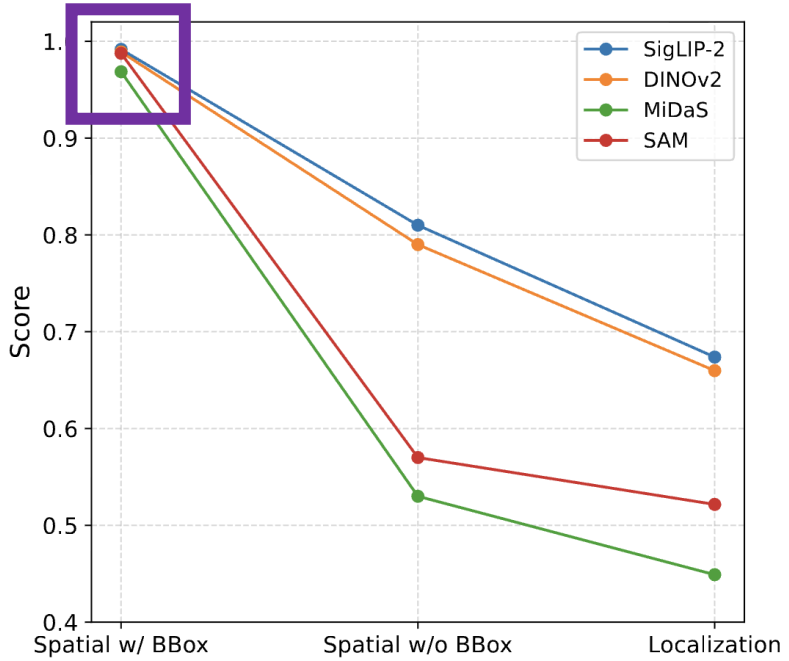
Even weaker VFMs perform well in **at least one** AVA

Diagnostics Reveal Failure Source



Where is the dog (annotated by the red box) located with respect to the plane (annotated by the blue box)?

A. Left above B. Left below C. Right above D. Right below.

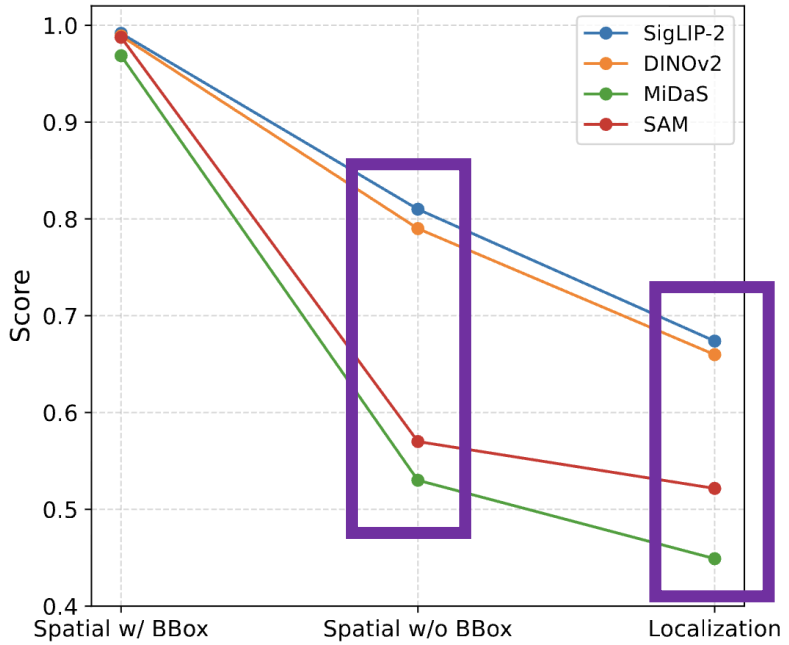


Diagnostics Reveal Failure Source

Composite task failures typically stem from **specific** AVA deficiencies rather than **general** visual incompetence.

Where is the dog located with respect to the plane?

A. Left above B. Left below C. Right above D. Right below.

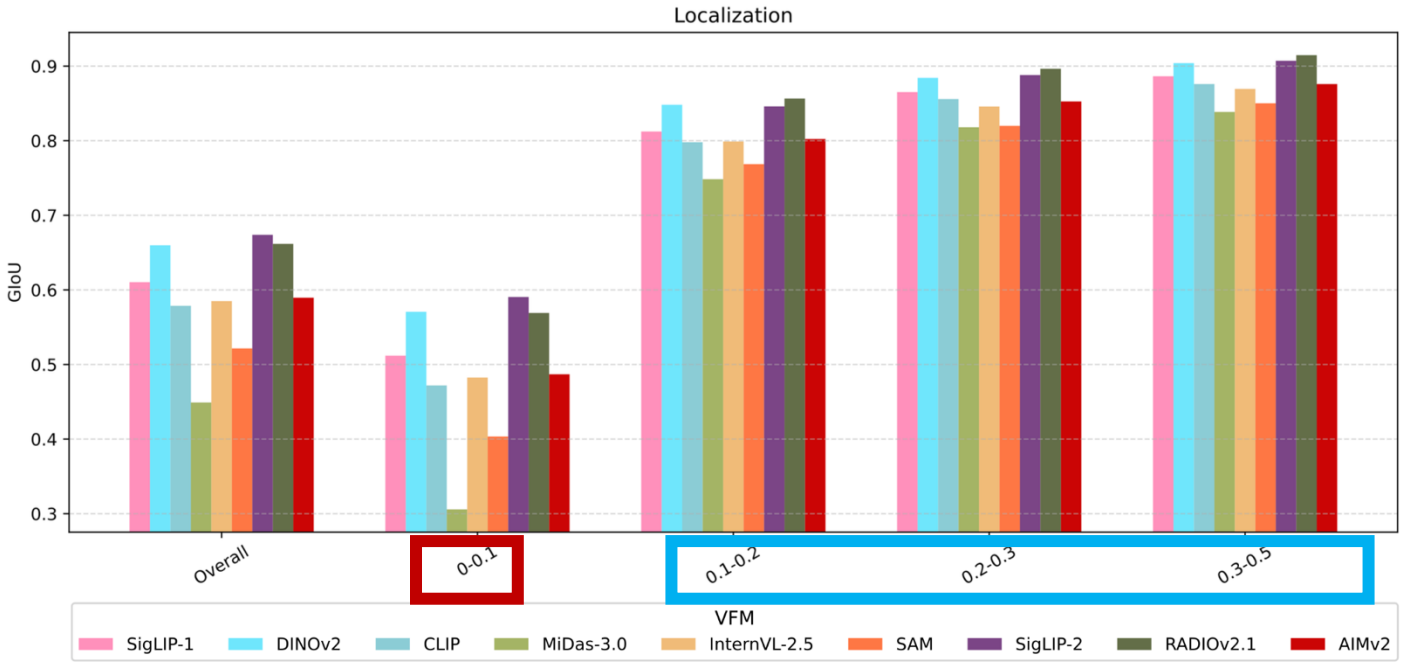


Subgroup Analyses

Subgroup analyses reveal hidden nuanced insight from aggregated metrics.

Example: Localization

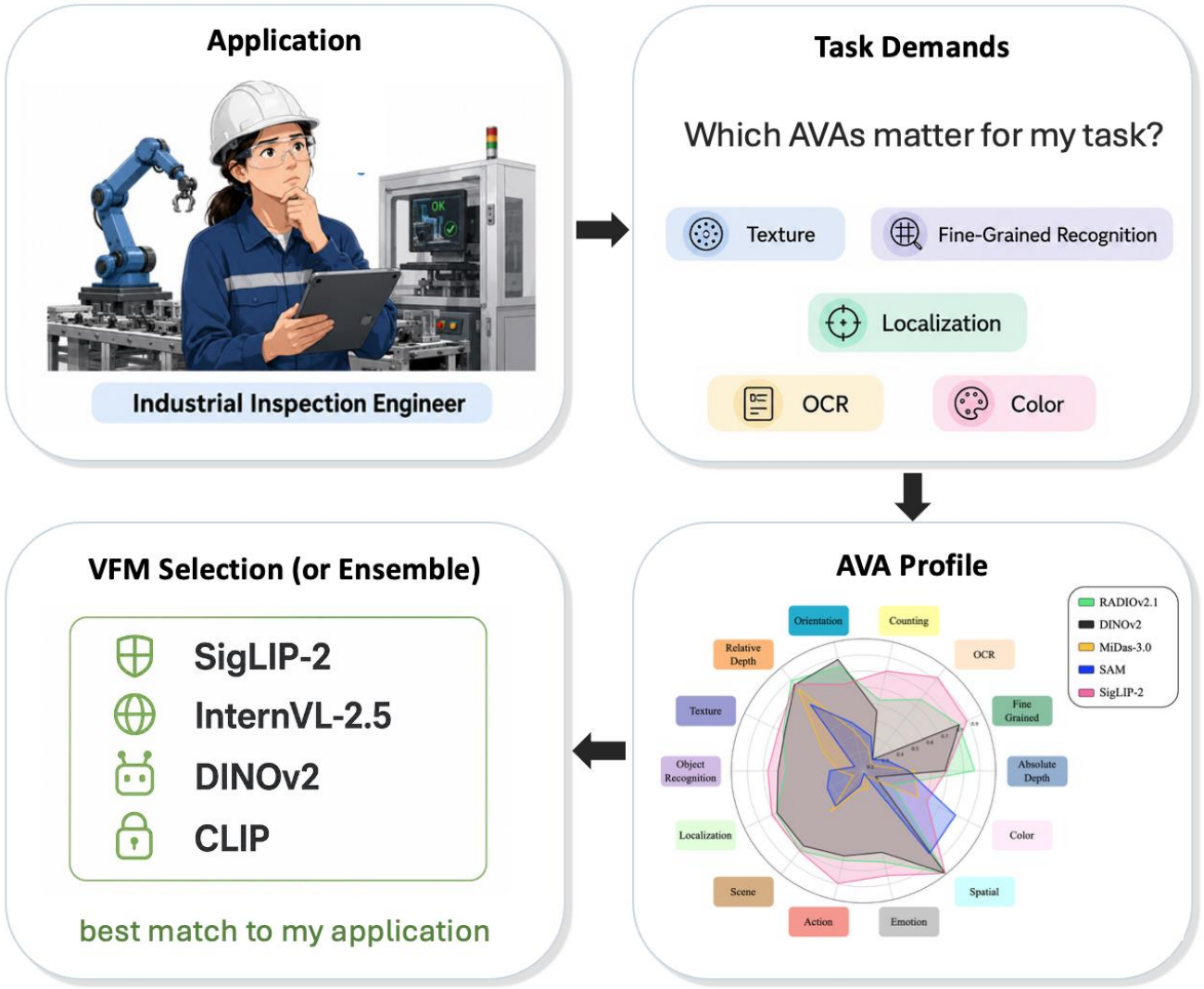
- VFM perform similarly for **large** objects
- Poor overall performance in some VFMs is mainly due to struggles in **small** objects.



0.1 means the bounding box size is 10% of the image size

VFM Selection

From heuristic choice to principled selection



Summary

AVA-Bench makes VFM choice **diagnosable, actionable, and efficient**

01

Diagnose

A systematic and diagnostic VFM evaluation benchmark

- 14 AVAs isolate visual strengths and failures
- not hidden inside a single VQA score

02

Guide selection

A detailed evaluation and insightful analysis

- Ability fingerprints reveal which VFM fits the task

03

Efficient + Open

A resource-efficient evaluation protocol & open-source

- Facilitate the development of the next generation of accountable and versatile VFM

Final message

Choose the VFM by the visual abilities your application needs — not by a single aggregate score.