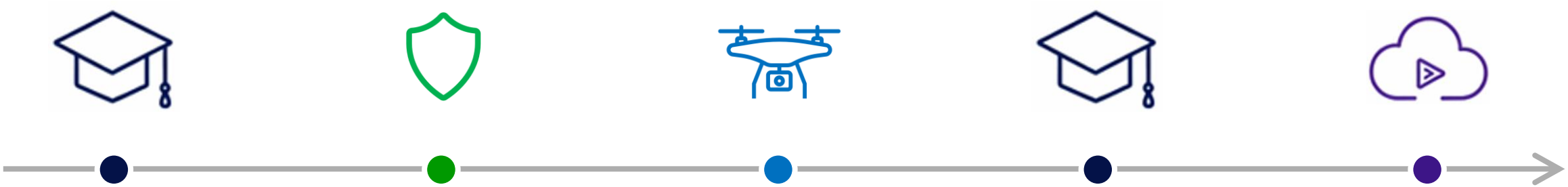
Microsoft

Tracking Detailed Descriptions in Live Video Streams

Tom Hirshberg | Edge-AI, Microsoft

Joint work with Zvi Figov, Moti Kadosh and Yonit Hoffman

About me



B.Sc. In CS

2016-2019

TECHNION Israel Institute of Technology

Algorithm Researcher

2018-2019

Defense company

Research Intern

Fall 2019

Microsoft

Redmond, USA

M.Sc. In CS

2020-2022

Cum Lauda

TECHNION Israel Institute of Technology

Data Scientist

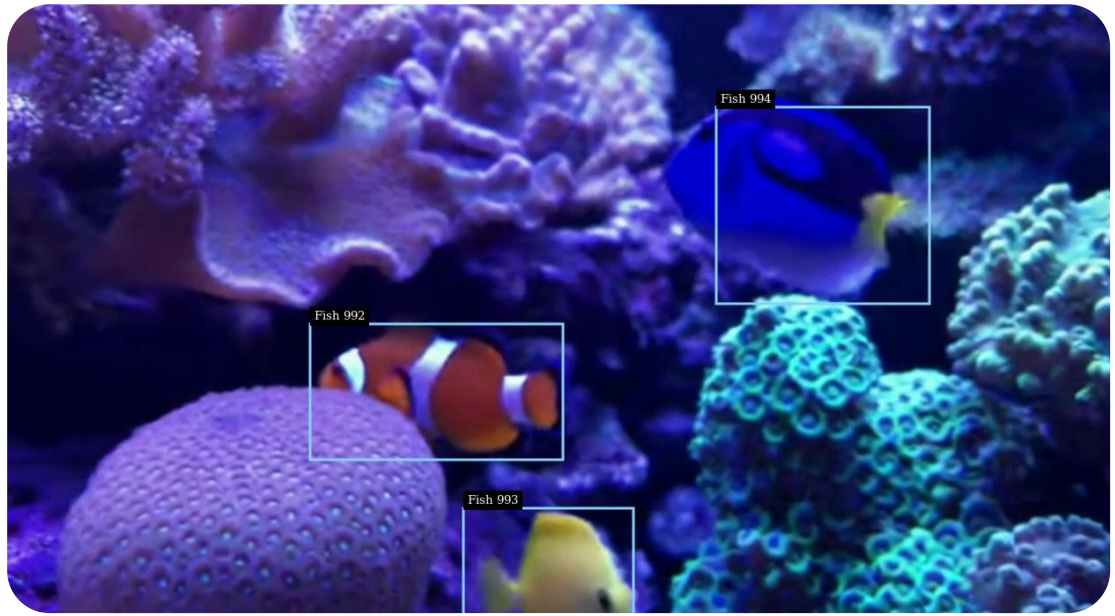
2022-Present

Microsoft

Israel

Beyond work

From object labels to live detailed description



Fish: A fish

Nemo: A clownfish with white and orange stripes

Same scene - Different problem

Live video analysis at the edge

t_0 t_1 t_2

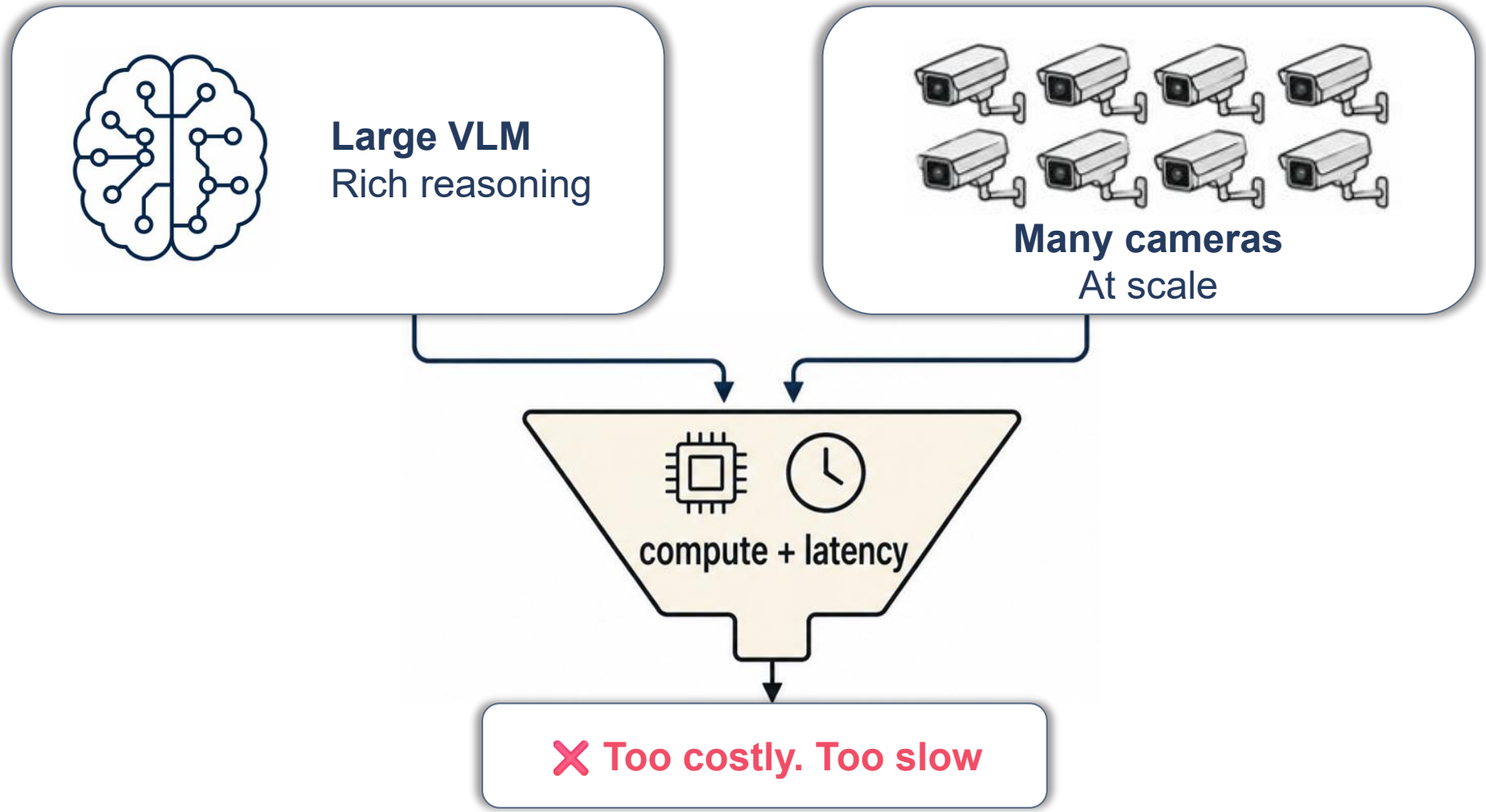
Accurate
Detailed attributes
Same identity over time

Efficient
Limited compute
Many cameras

Speed
Low latency
Instant results

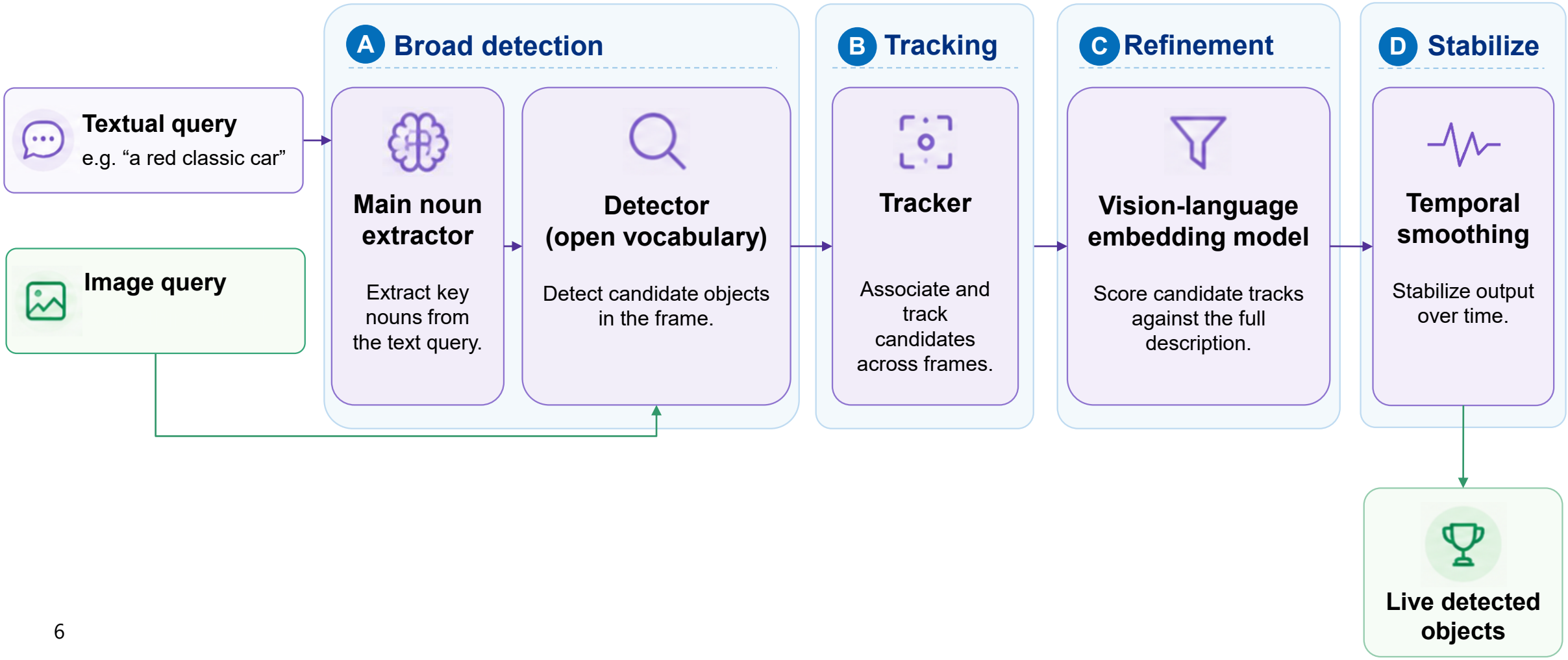
4 Query reminder: a clownfish with white and orange stripes

Naïve approach - run a large VLM on every frame



5 VLM = vision-language model

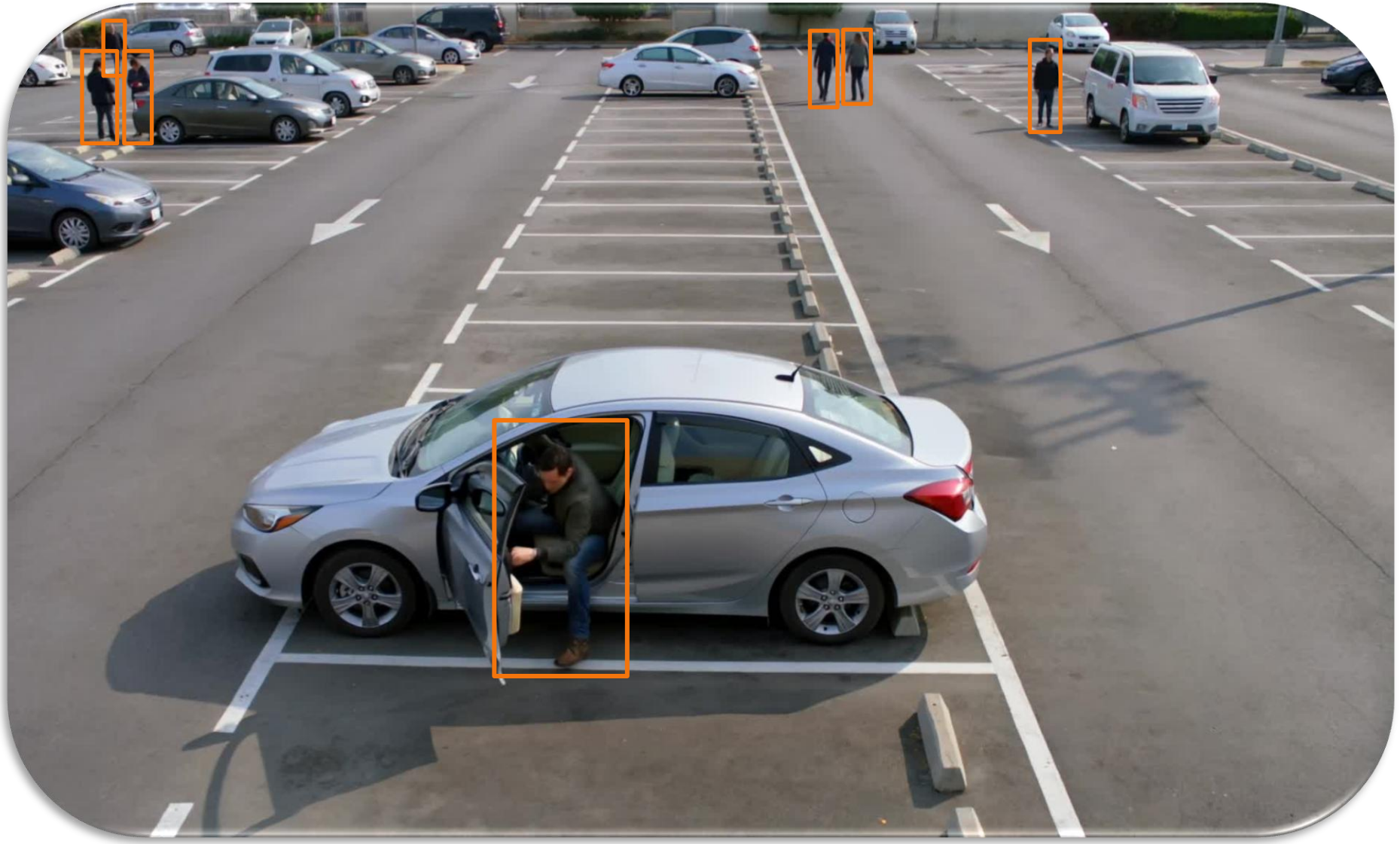
Method: from query to stable live detection



From description to stable target

Query
a person carrying a cardboard box

Broad detection

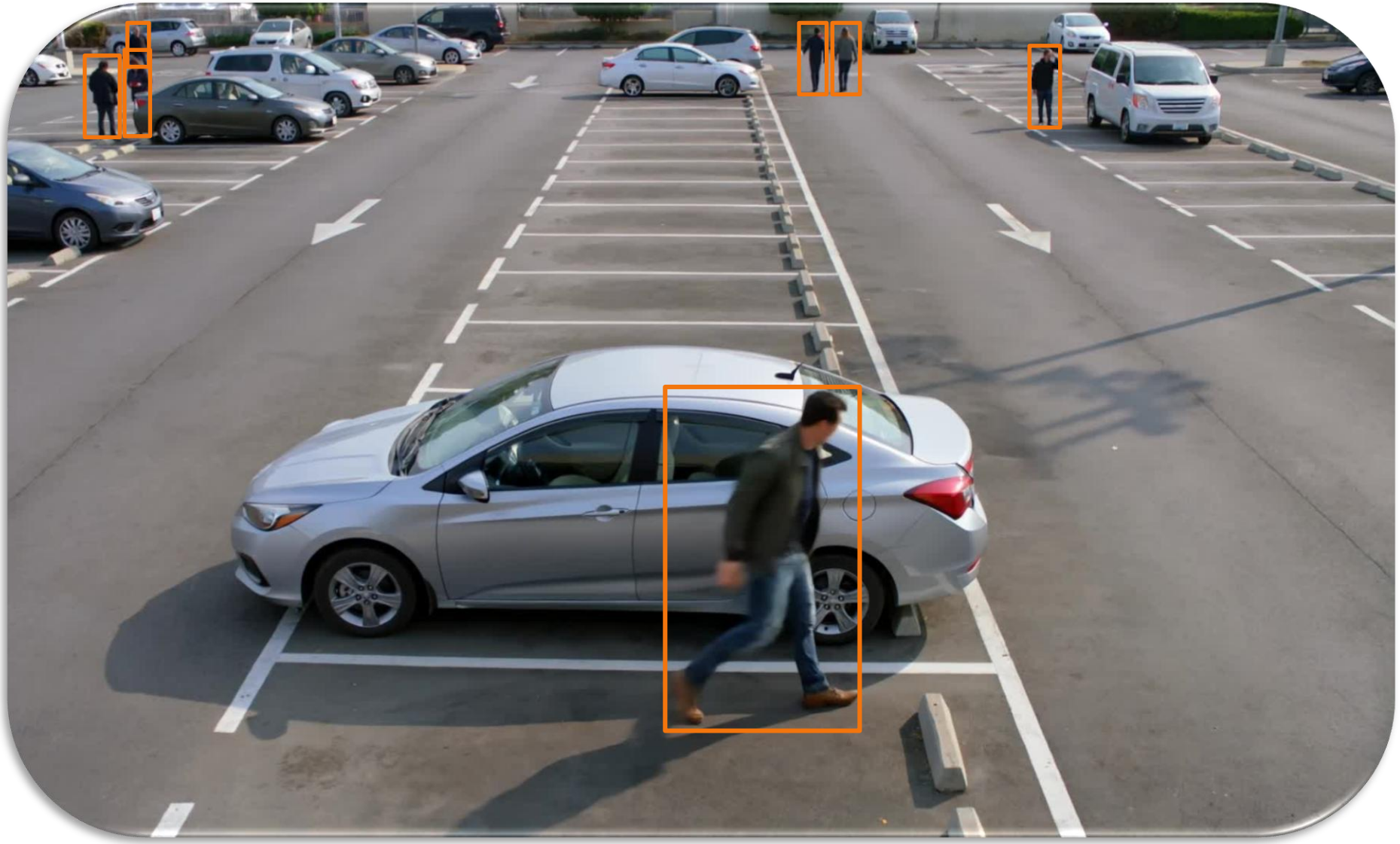


From description to stable target

Query
a person carrying a cardboard box

Broad detection

Tracking



From description to stable target

Query
a person carrying a cardboard box

Broad detection

↓

Tracking

↓

Refinement

From description to stable target

Query
a person carrying a cardboard box

- **Broad detection**
- **Tracking**
- **Refinement**
- **Stable target**

Demo: Live track from a detailed text description

Query: a person carrying a cardboard box

Negative examples define intent

True positive

Worker wearing a helmet

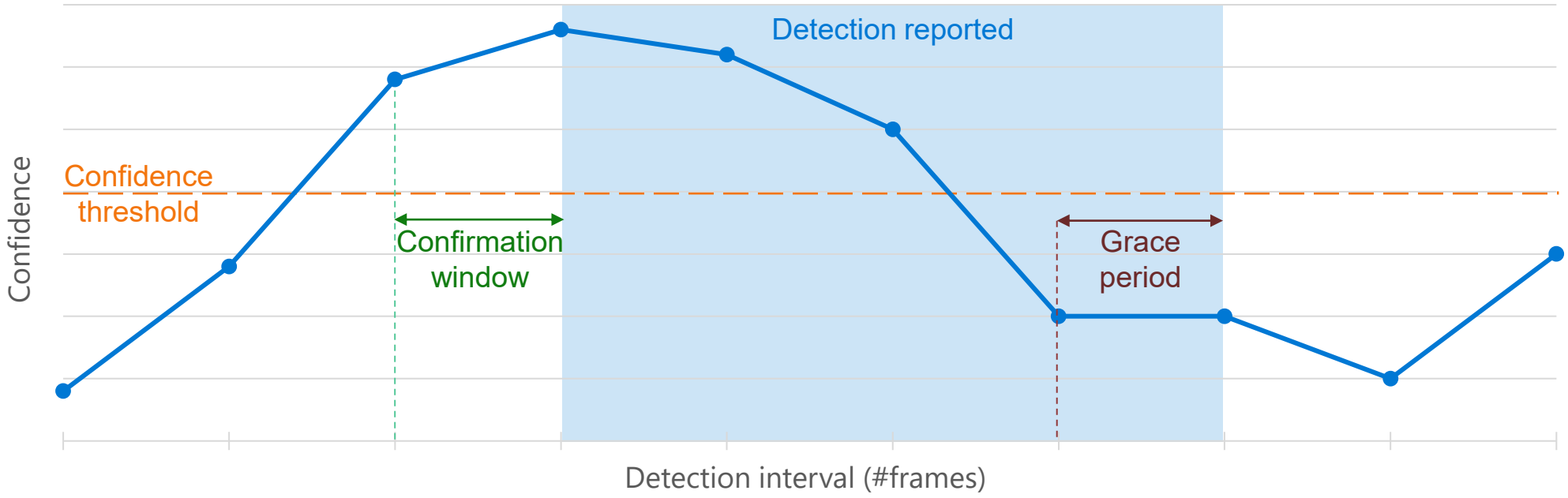
False positive

Worker carrying a helmet

Runtime contrast - Score the candidate against likely false positives. Report only when the true positive wins.

 NegBench / arXiv:2501.09425: modern VLMs often struggle with negation.

Hysteresis prevents flickering detections



Confirmation window
Sustained confidence above threshold

Grace period
Brief drops allowed before reset

Hysteresis turns noisy frame-level scores into stable live output

Demo: detailed detections are dynamic

Query: A worker wearing a yellow helmet
Negative example: A worker carrying a yellow helmet

Evaluation: selecting the refinement model

Model	F1	Precision	Recall
SigLIP2	**0.89**	**0.88**	**0.9**
MMRet	0.89	0.83	0.95
SigLIP	0.86	0.87	0.85
CLIP	0.82	0.80	0.85

ⓘ Precision = correctness | Recall = coverage | F1 = balance between them

CLIP, SigLIP, SigLIP2, and MMRet are **vision-language embedding models** used to score image-text similarity

SigLIP2 = Best practical balance of precision and recall

Dataset

7 object-attribute categories
A few dozens of positive and negative examples per category

Learning transferable visual models from natural language supervision / International conference on machine learning. PmLR, 2021.
Sigmoid loss for language image pre-training / Proceedings of the IEEE/CVF international conference on computer vision. 2023.
Megapairs: Massive data synthesis for universal multimodal retrieval / Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. 2025.
Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features / arXiv:2502.14786. 2025

Production-ready, with responsible safeguards

Real-world deployment

Production-tested in Azure AI Video Indexer.

Supporting detailed queries for live detection and tracking.

Edge-friendly live processing.

Responsible AI guardrails

Customer-controlled video locality.

No biometric identification; object and attribute matching only.

Validation and thresholds bound ambiguity.

Key takeaways

Detect broad

Open-vocabulary detection is a **strong starting point**, not the destination.

Track identity

Separate identity from state. **Track once**, score attributes independently.

Stabilize explicitly

Negative examples and hysteresis are not minor details – they **build trust**.

Live detection needs more than rich semantics.
It needs the right pipeline.

Thank you!

Any questions?

tom-hirshberg