AI P14: Object Detection Model Comparison

Benchmarking Speed & Stability in Real-World Video Scenarios

TEAM MEMBERS

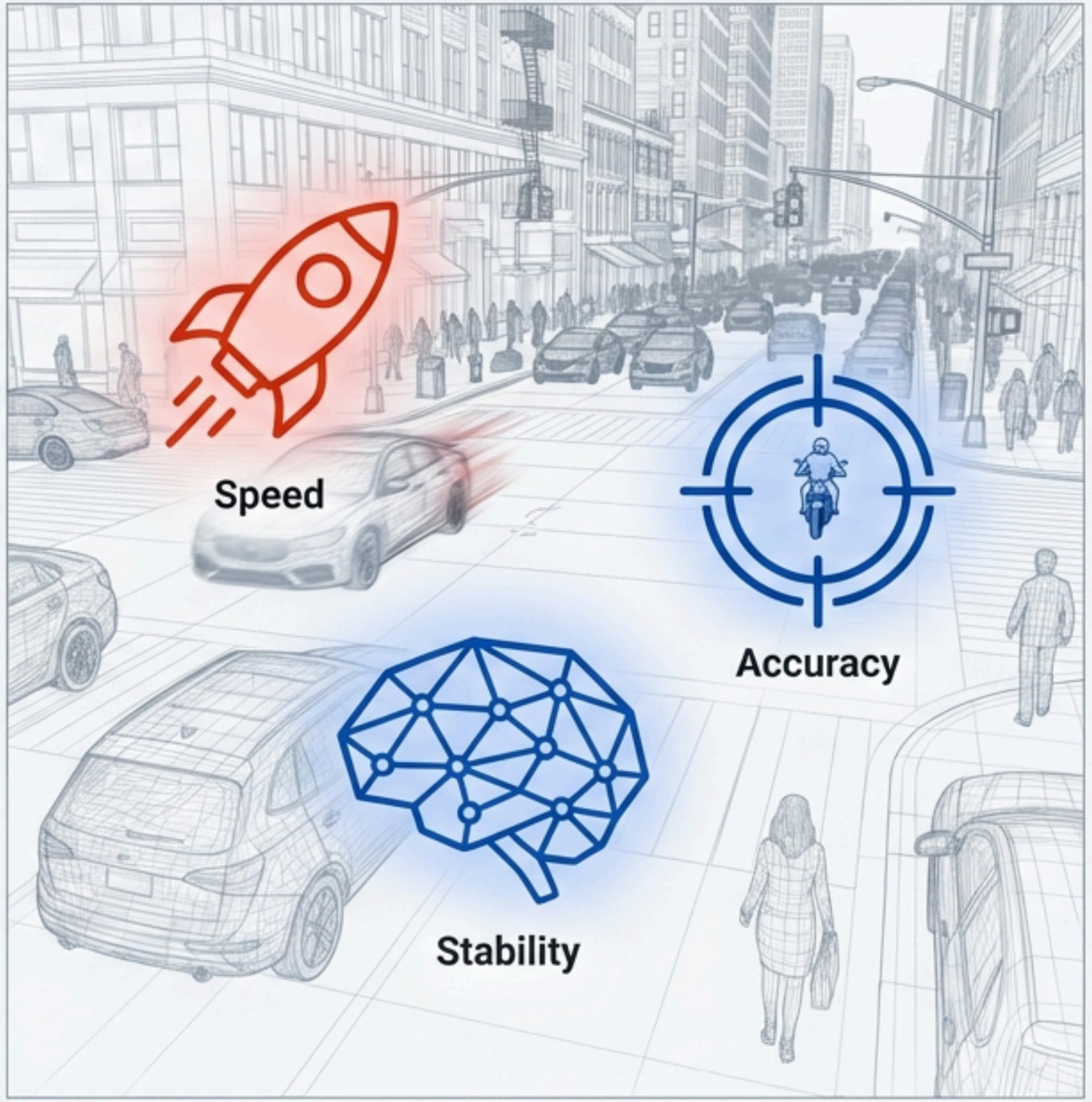
Lokesh Sohanda	Sai Surya Pratap Reddy Dumpa
Mohan.B	Valaboju Anil chary
KOMMIRI PRASANNA KUMAR	Logavanan A

🎯 **Project Objective:** To deploy and benchmark diverse computer vision architectures—including **YOLO, MobileNet, and DETR**—on video data to determine the optimal trade-off between **Inference Speed (FPS)** and **Temporal Stability** for real-world applications.

The Architect's Choice:

A Comparative Analysis of Modern Video Object Detection Models

- This analysis provides a rigorous, data-driven evaluation of leading object detection architectures, from lightweight CNNs to state-of-the-art Transformers.
- **Objective: To identify the optimal models for two distinct, critical business needs: real-time monitoring and high-precision offline analytics.**
- The findings will inform strategic decisions on model deployment, balancing the trade-offs between speed, accuracy, and operational stability.



The Verdict Upfront: One Mission, Three Winners

Our analysis across five models and 1,512 video frames reveals that the optimal choice is mission-dependent. There is no single 'best' model. We identified clear champions for three distinct operational requirements, from edge devices to high-power analytics servers.

FOR RAW SPEED

MobileNet SSD

~66 FPS

The undisputed winner, making it the only choice for low-power edge devices like a Raspberry Pi.

FOR REAL-TIME PERFORMANCE

YOLOv11x

11.09 FPS

Provides the best balance of accuracy and speed, achieving a usable framerate for live-stream analysis.

FOR MAX ACCURACY & STABILITY

Facebook DETR

89.77%
Avg. Confidence

+28%
More Objects

Superior accuracy and stability, detecting more objects with minimal false positives.

The Proving Ground: Models, Architectures, and Data

To ensure a comprehensive test, we evaluated models across three distinct architectural paradigms. The video dataset was specifically chosen for its diversity and real-world complexity, providing a robust testbed.

Models Tested (5)

Lightweight CNN

MobileNet SSD

↓

Single-Stage CNNs

YOLOv5 | YOLOv8n | YOLOv11x

↓

Transformer

Facebook DETR

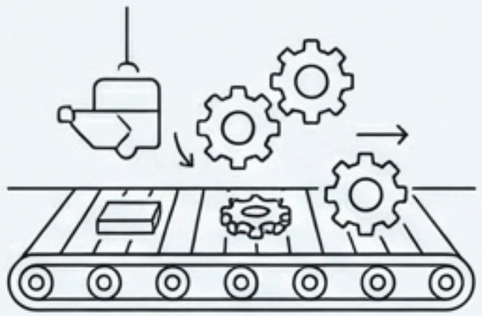
Dataset

1,512

Total Frames

Dynamic Traffic 🚗

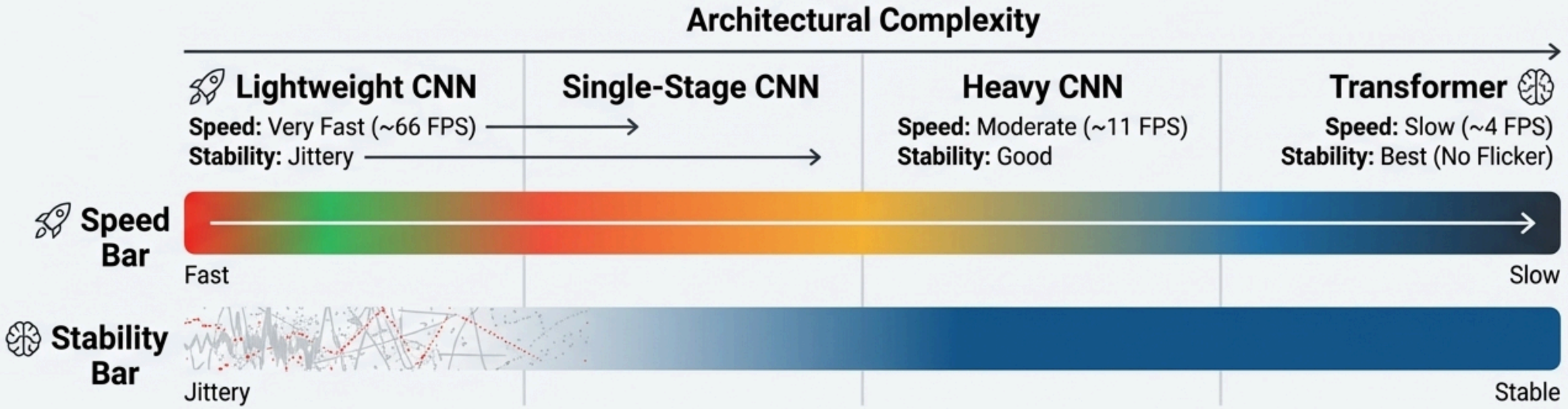
Indoor Office Surveillance 🏢



Fine-Grained Industrial Objects 🏭

The Trade-Off Matrix: Speed vs. Stability at a Glance

- This high-level comparison reveals a clear and predictable trade-off: as architectural complexity increases, inference speed decreases while stability and small-object performance improve.
- YOLO architectures represent the middle ground, but the heavier YOLOv11x variant shows a high rate of false positives, indicating over-sensitivity.
- DETR's Transformer design is fundamentally slower but its ability to avoid 'flicker' makes it ideal for offline tasks where object tracking is critical.



Quantitative Showdown: The Speed Demon vs. The Accuracy King

- Focusing on our two top-tier candidates—YOLOv11x and DETR—the performance gap is stark and clearly defines their ideal use cases.
- YOLOv11x is nearly **three times faster** in total processing time, making it the only viable option for interactive systems.
- DETR is significantly more confident and comprehensive, detecting over **2,500 more objects** in the same video sequence.

YOLOv11x

136s
Total Time

11.09 FPS
Real-World FPS

57.91%
Average Confidence

9,274
Total Objects Detected

VS

Facebook DETR

365s
Total Time

4.13 FPS
Real-World FPS

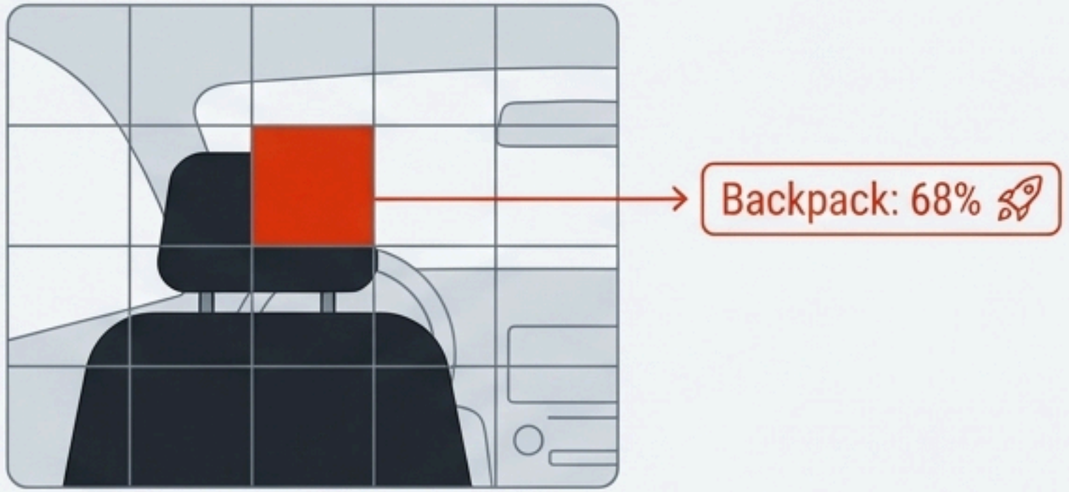
89.77%
Average Confidence

11,870
Total Objects Detected

Anomaly Analysis: Global Context vs. The Backpack Hallucination

These performance gaps are not random; they are direct results of the underlying model architectures.

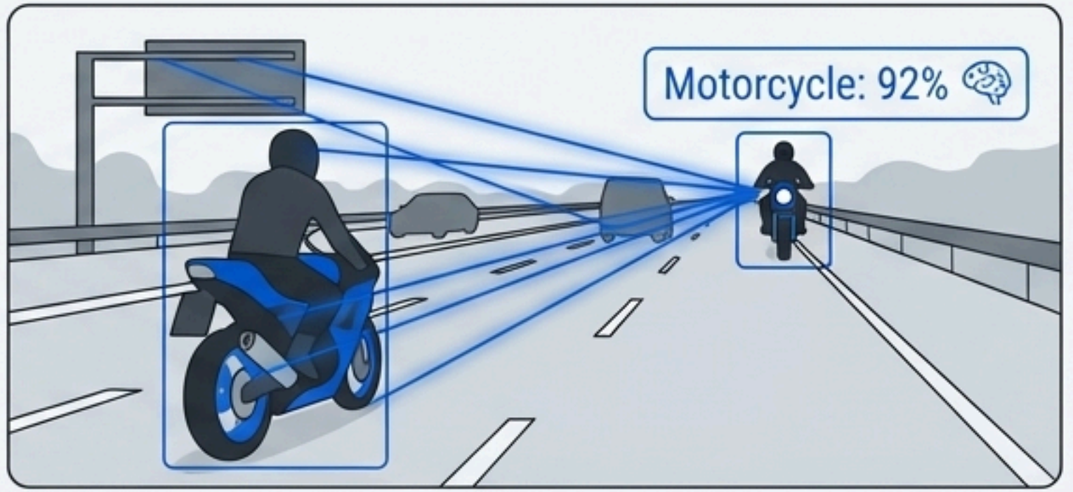
The Backpack Hallucination



YOLO's localized grid-based approach sees a shape resembling a backpack but lacks the broader context to rule it out.

617 False Positives
NEUTRAL_GRAY (#6C757D)

Global Context



The Transformer "sees" the entire image at once. This global context allows it to understand that a motorcycle-like shape *on a road* is a motorcycle.

519 Contextual Accuracy
NEUTRAL_GRAY

The Confidence Gap: Why DETR Delivers Stability

The **32% average confidence gap** between the models is the most critical finding for system stability and reliability

YOLO: Wide Net, Low Confidence

Casts a "wide net" with low confidence. This leads to flickering boxes as certainty wavers frame-to-frame and requires heavy filtering (Non-Max Suppression), which can accidentally remove correct detections.

DETR: Conservative but Sure

Is "conservative but sure." It makes fewer predictions, but with high certainty. This results in stable, non-flickering bounding boxes ideal for object tracking.

*Architectural Cause: Transformer's global self-attention vs. CNN's localized regression.

Challenges Faced During Project

01

ARCHITECTURE SELECTION

Choosing the right models

Overcame difficulties in selecting the optimal model from vast options available across platforms.

- YOLOv5, YOLOv11x
- MobileNet SSD
- Facebook DETR (ResNet50)

02

HYPERPARAMETER OPTIMIZATION

Optimizing training parameters

Struggled with tuning parameters to achieve optimal performance balance.

- Learning rates & Batch sizes
- Epoch limits
- Anchor box scaling

03

ATTENTION MECHANISM

Lack of Context & Class Errors

Architectures without global context (like basic YOLO) misclassified objects.

- 90% of motorcycles as bicycles
- Parts of cars confused
- Fine-grained gaps

Key Learning

This project revealed that object detection is not a single-metric problem. Model choice and training strategy must align with real-world scenarios—where accuracy under ambiguity outweighs speed.

The Final Verdict: Choose Your Model Based on Your Constraint

The core lesson of this analysis is that architecture dictates capability. The choice between models is not about which is "better," but which constraint—time or precision—is most critical to your application.

If your primary constraint is...

TIME

...Choose YOLO.

Roboto Slab Regular, with Slab Regular.
It is **2.7x faster** and **sufficient** for most **live** tasks.

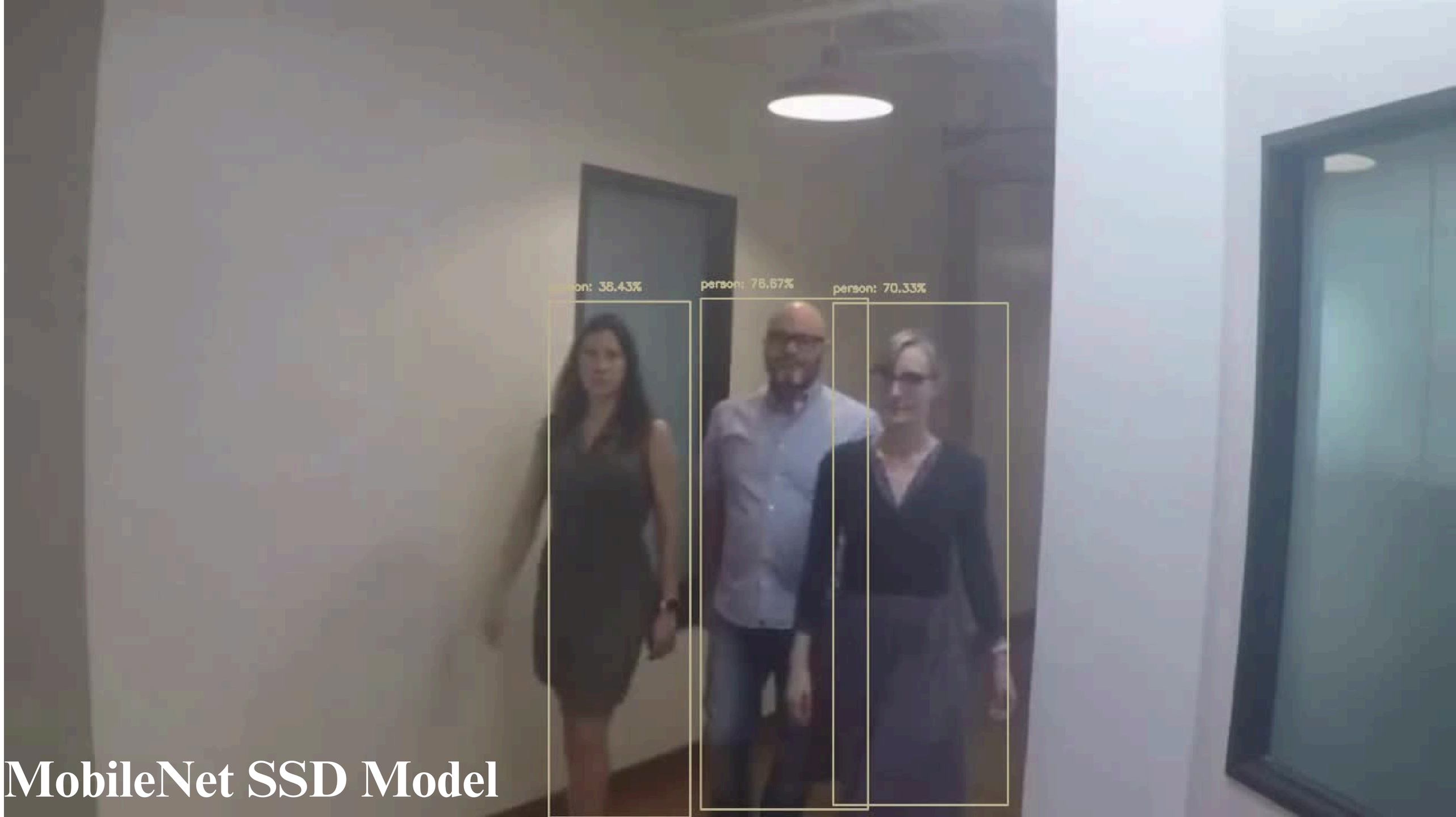
If your primary constraint is...

PRECISION

...Choose DETR.

Roboto Slab Regular, with Slab Regular.
It finds **28% more objects** and excels at **small, difficult targets** with **near-90% confidence**.

**The Unifying Strategy: Speed favors YOLO.
Precision favors DETR. Smart systems leverage both.**



MobileNet SSD Model


car 0.56
YOLO Model
person 0.92
car 0.82
car 0.82
.40
car 0.89
person 0.40


DETR Model
person 0.99
car 0.99
car 0.78
car 1.00