



University of the
Philippines
Diliman

Samsung
R&D Institute
Philippines

Interpreting Open-Vocabulary Referring Object Detection with Reverse Contrast Attention

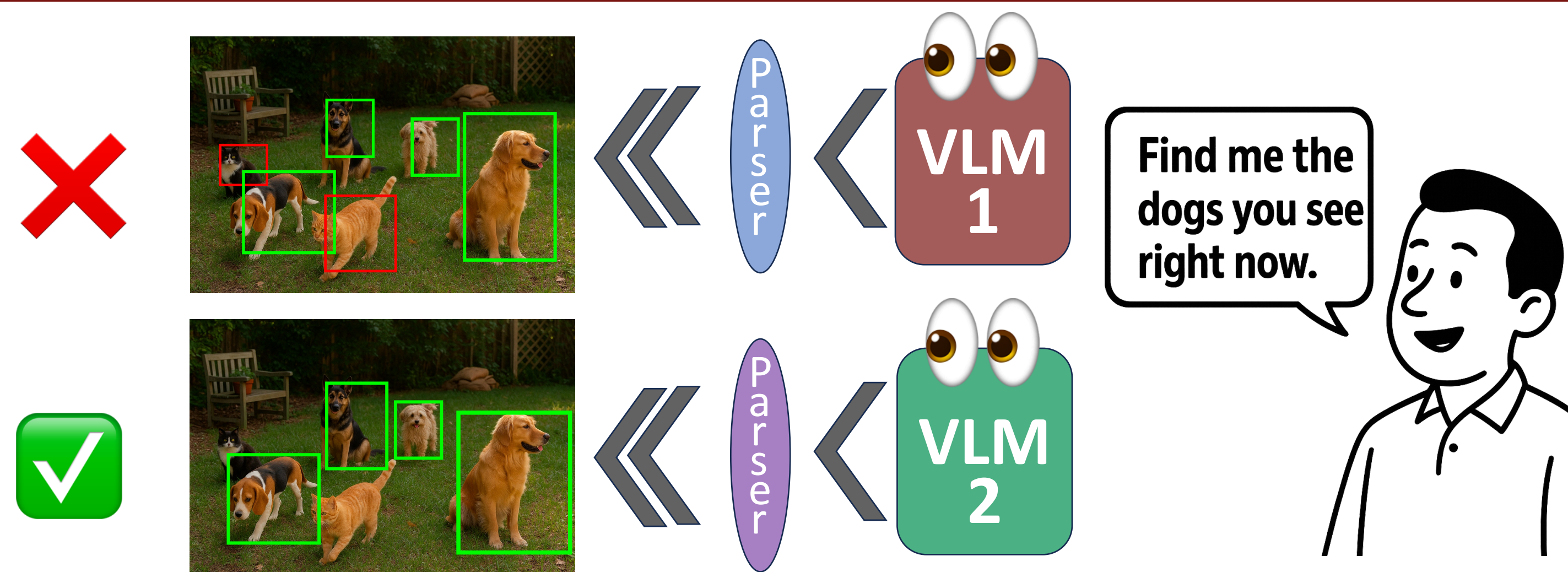
Drandreb Earl Juanico*, Rowel Atienza, Jeffrey Kenneth Go



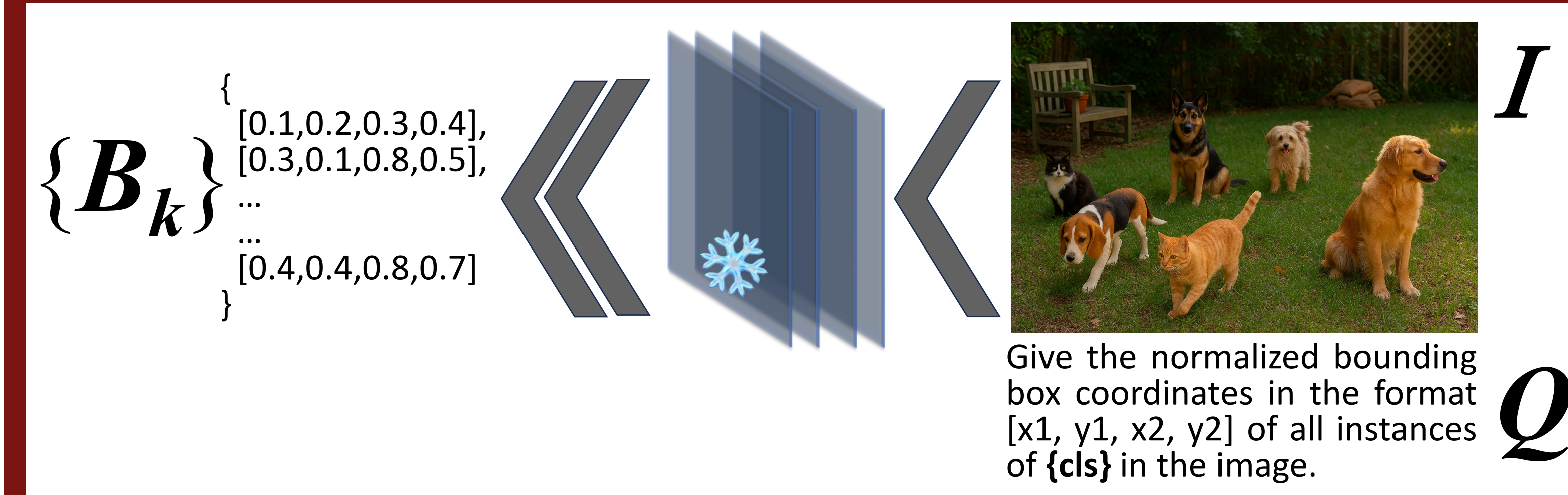
ICCV
OCT 19-23, 2025
HONOLULU
HAWAII



Open-Vocabulary Referring Object Detection



Inference



VLM Selection

Best models via:

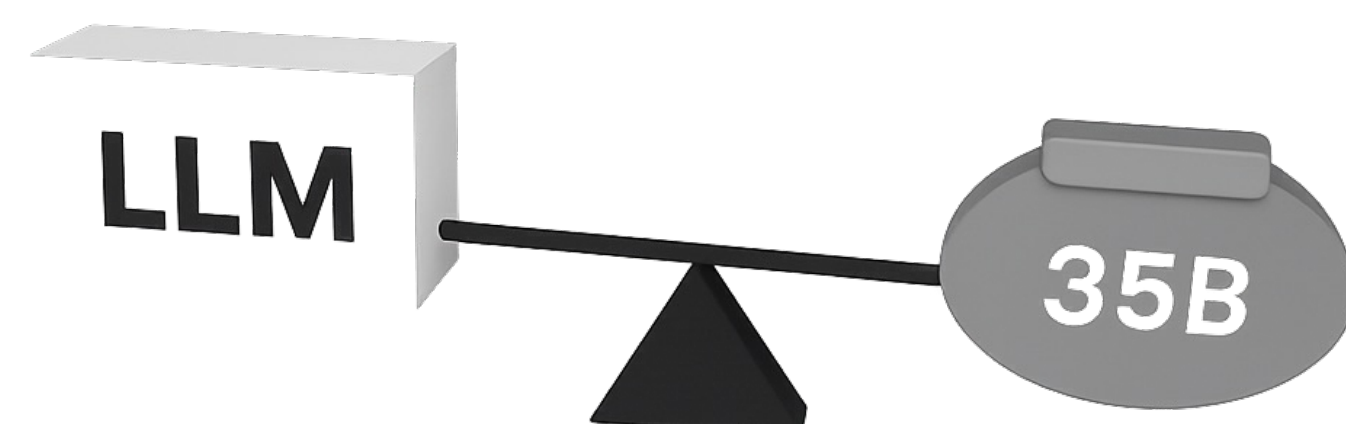


OpenCompass 司南

Checkpoints available via:



Hugging Face



bounding-box
information
parseable

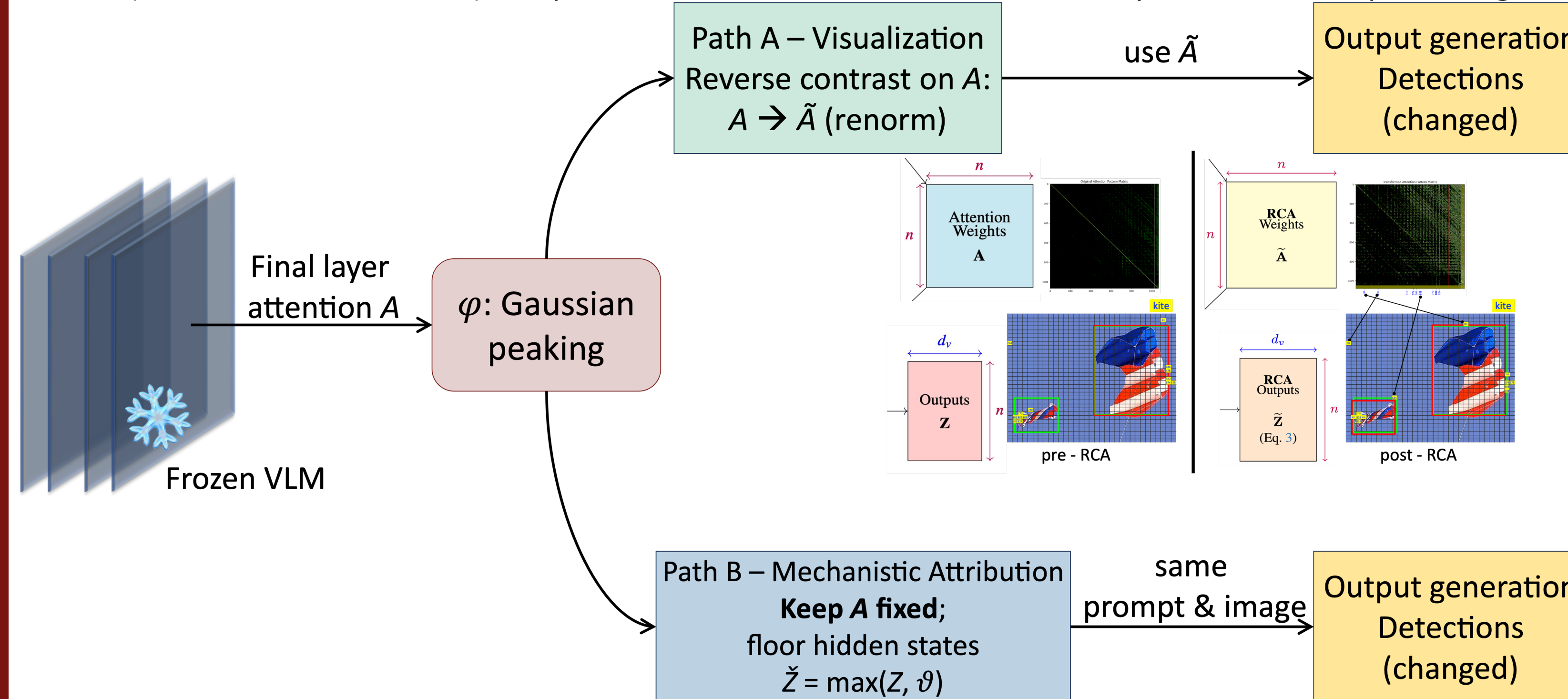
$\{B_k\}$

Reverse Contrast Attention

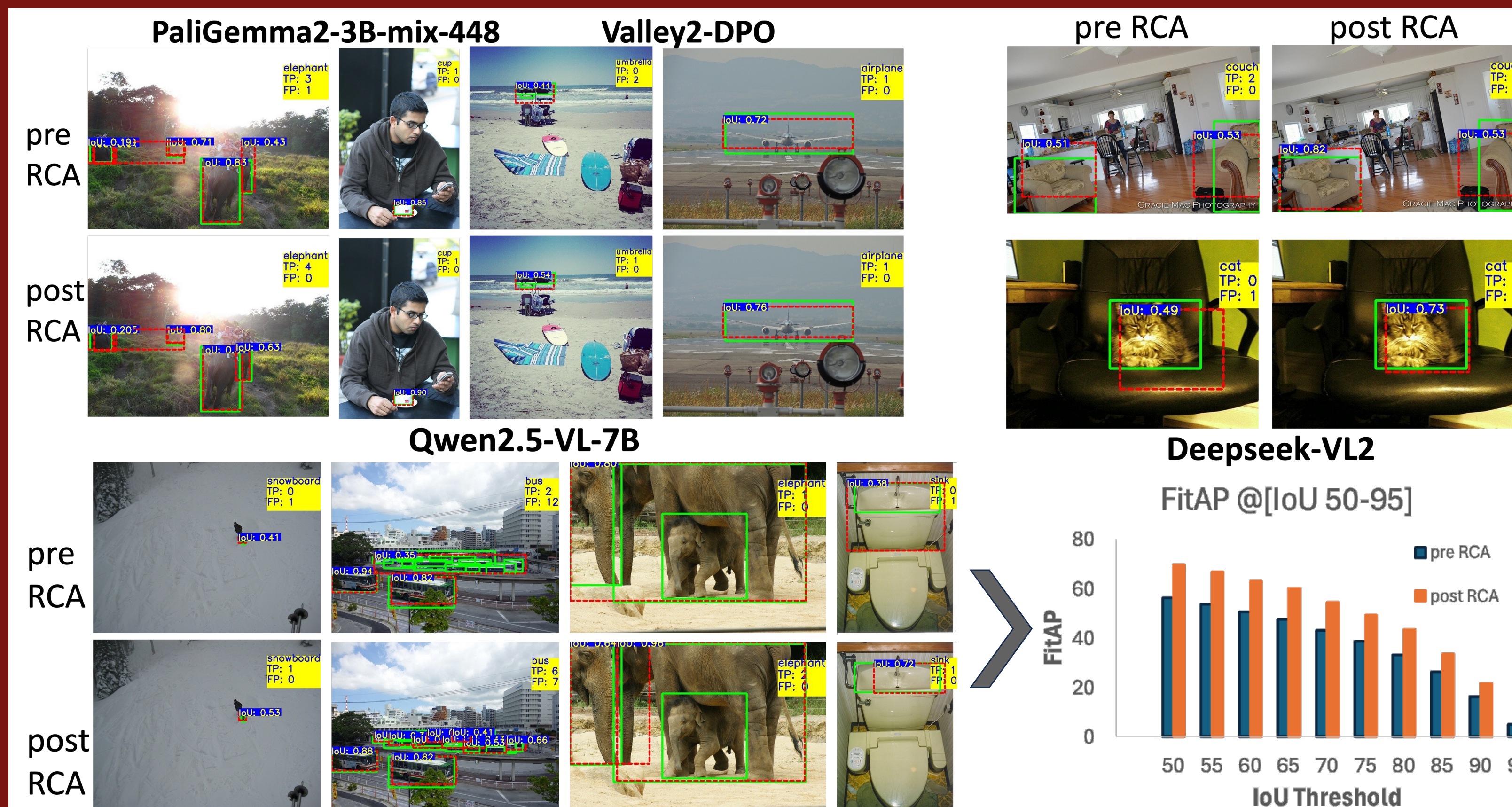
RCA: Two linked interpretations from one transformer

Path A (visualization): apply φ to attention A (reverse-contrast $\rightarrow \tilde{A}$) and feed back to outputs.

Path B (mechanistic attribution): keep A fixed; floor hidden states Z at ϑ to explain similar output change

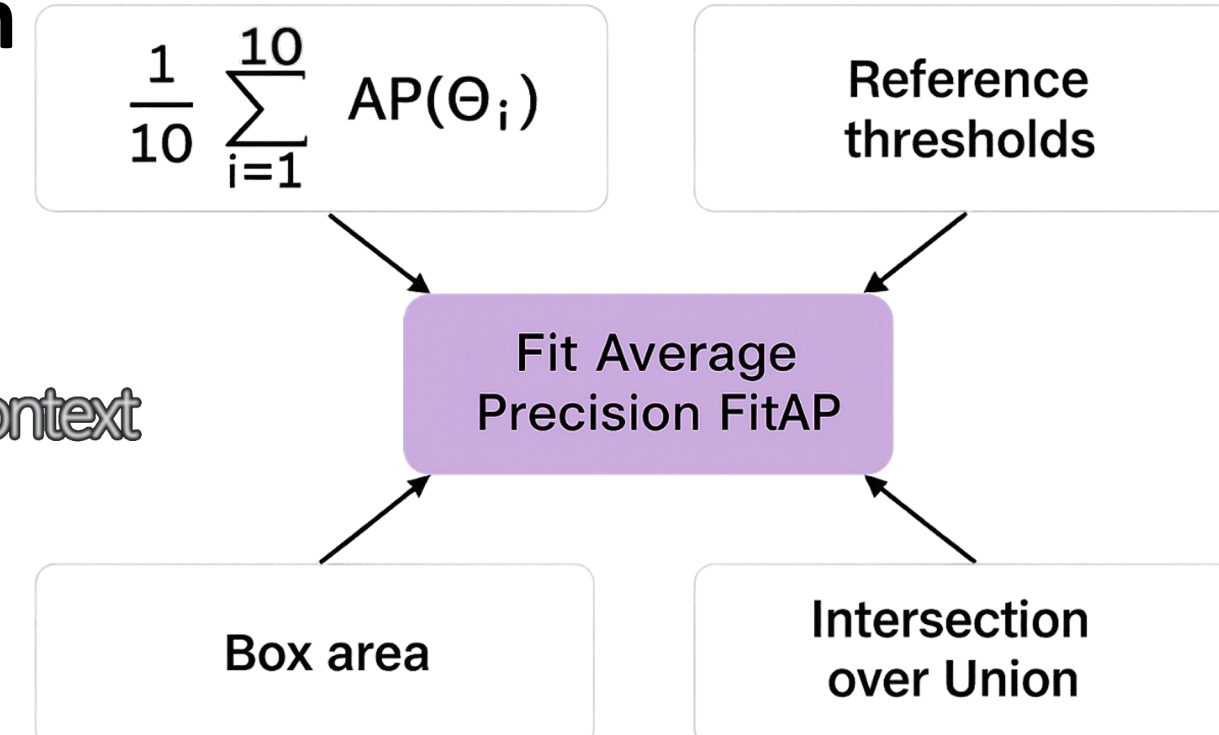


Selected Examples

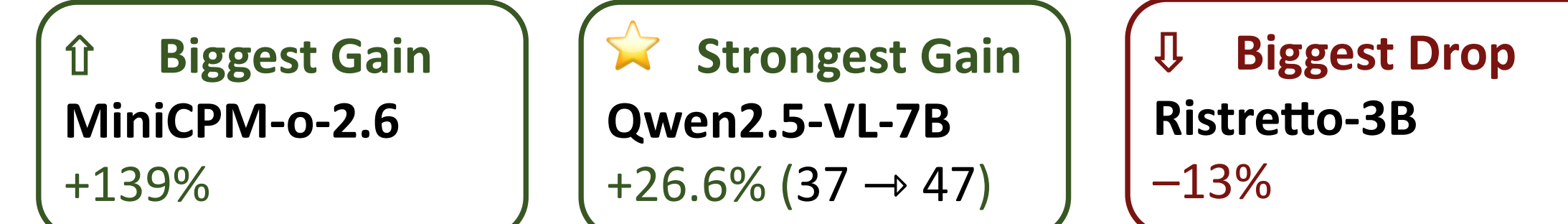


Evaluation

FitAP, average precision



Results



Top Performer
Moondream2
47.08 post-RCA

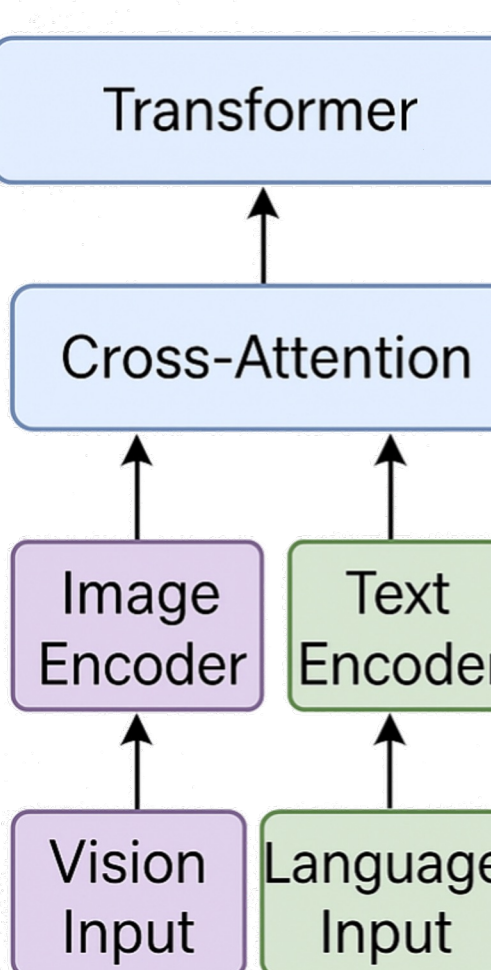
2nd Performer
Qwen2.5-VL-7B
46.85 post-RCA

11 of 15 selected VLM gained post-RCA

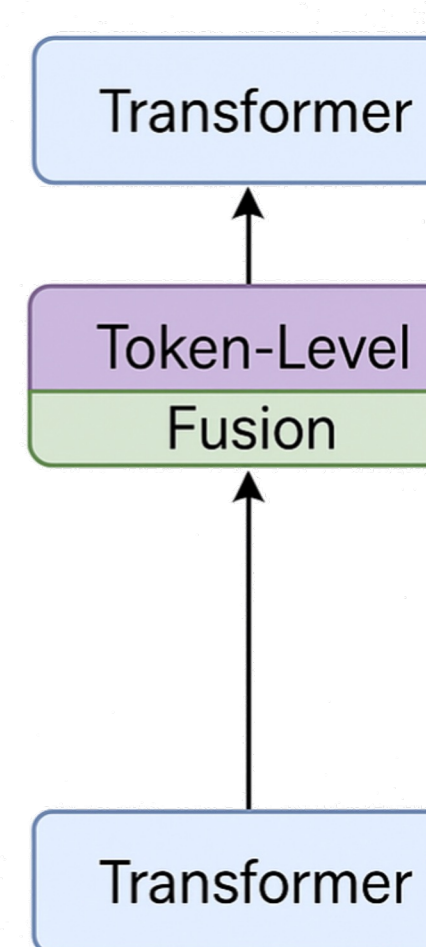
Explainability

Fusion Timing

Modular Fusion

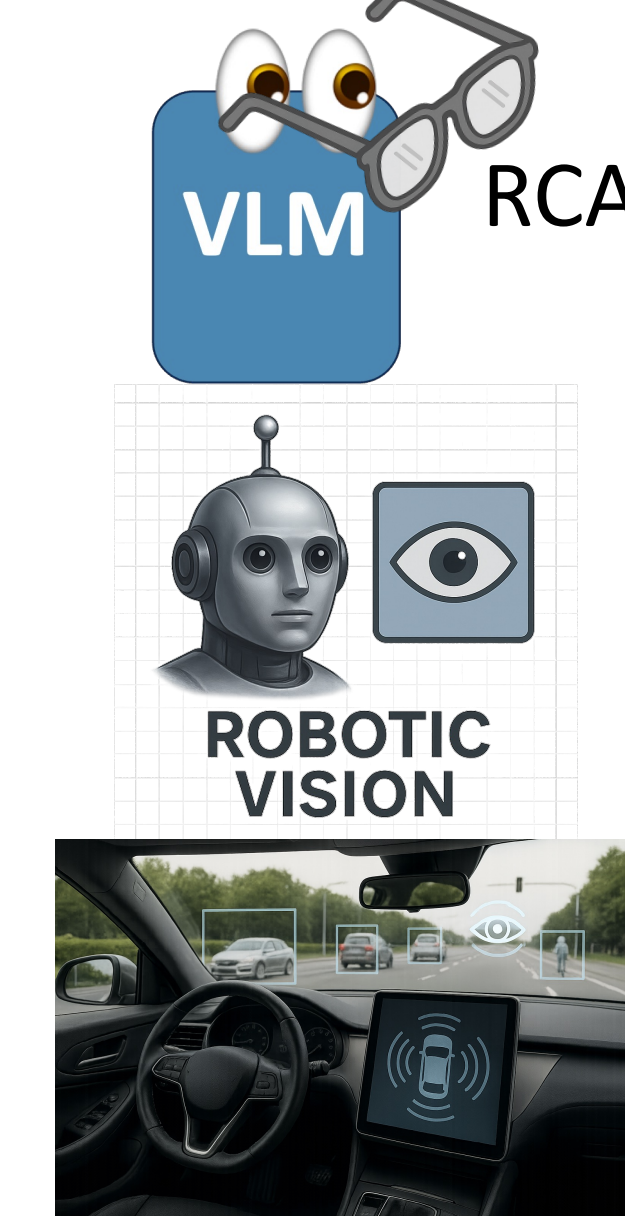


Early Fusion



other factors: training data, prompt, specific implementation choices

Applications



Acknowledgement: This work was supported by Samsung Research Philippines. DEJ acknowledges the support from the DOST-SEI, ERDT, and ICCV Broadening Participation.