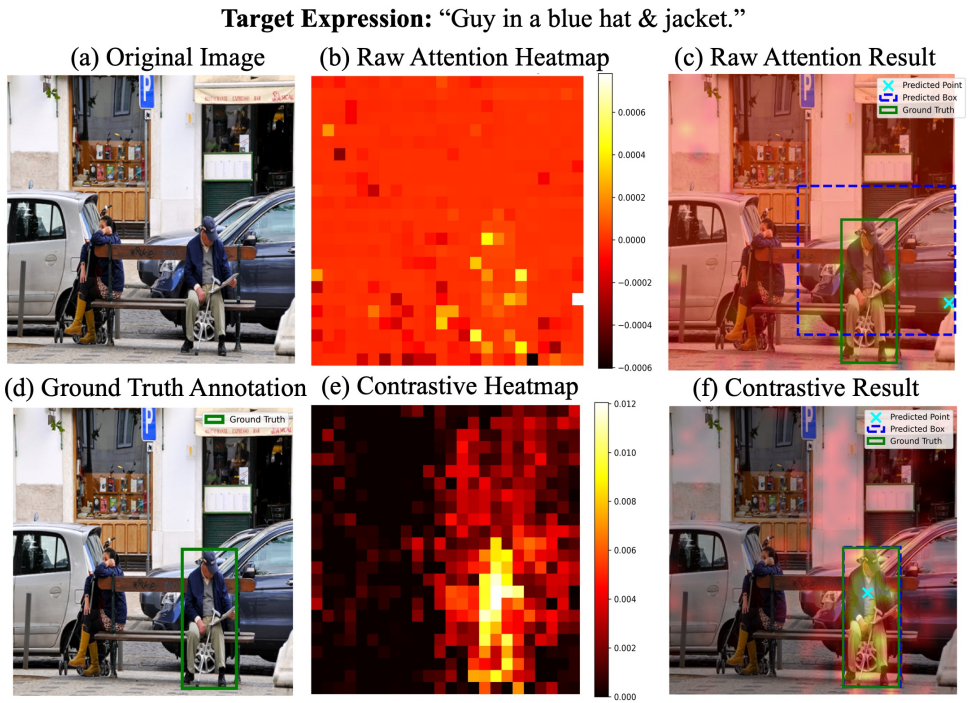
Few Contrastive Attention Heads Enable Visual Grounding in Large Vision-Language Models:

Discovering Localization Heads Without Supervision

Neha Sengar

Vehicular Intelligence Lab |
Ph.D. Candidate | *KAIST Robotics Program*
Email- neha.sengar@kaist.ac.kr

Discovering Localization Signals in Frozen Vision-Language Models



The Limitation of Naive Attention Aggregation

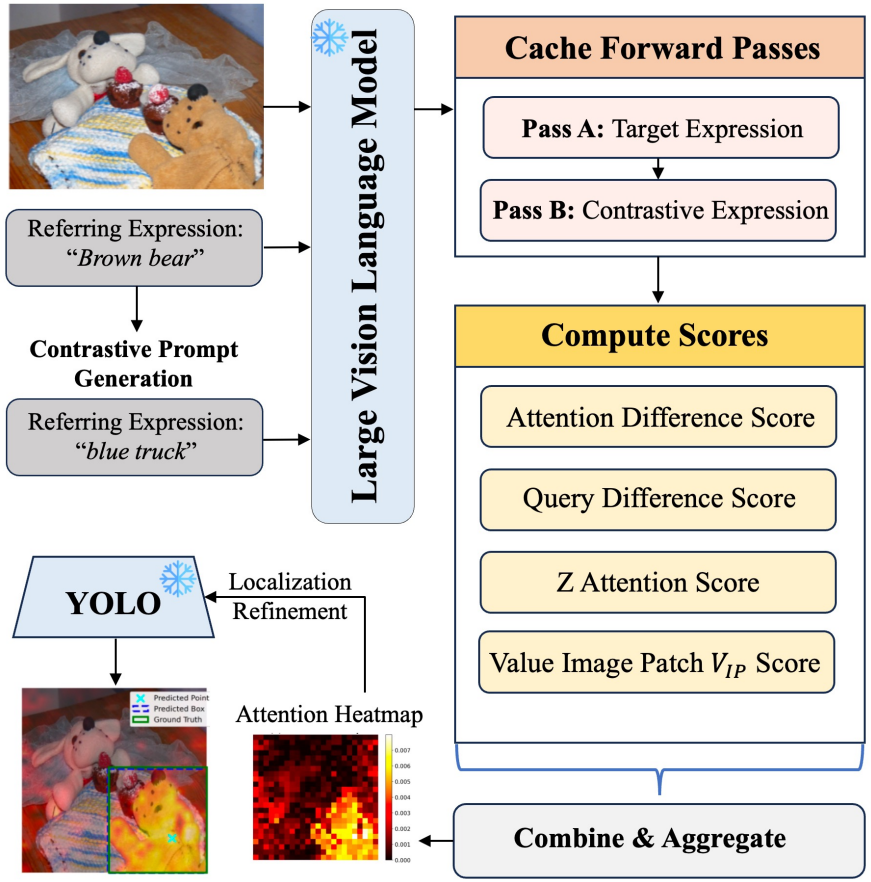
- LVLMs **describe objects** from text query → but **attention maps are diffuse** (fig. b) → miss the target (fig. c).
- **Averaging across all heads** → the spatial signal is **destroyed.**

Our Insight: The Signal Lives in a Few Heads

- Localization is encoded in a **small subset of attention heads** (fig. e) → extract via **contrastive dual prompting**.
- Aggregating only **top-10** heads → precise grounding (fig. f) with **no grounding fine-tuning** required.

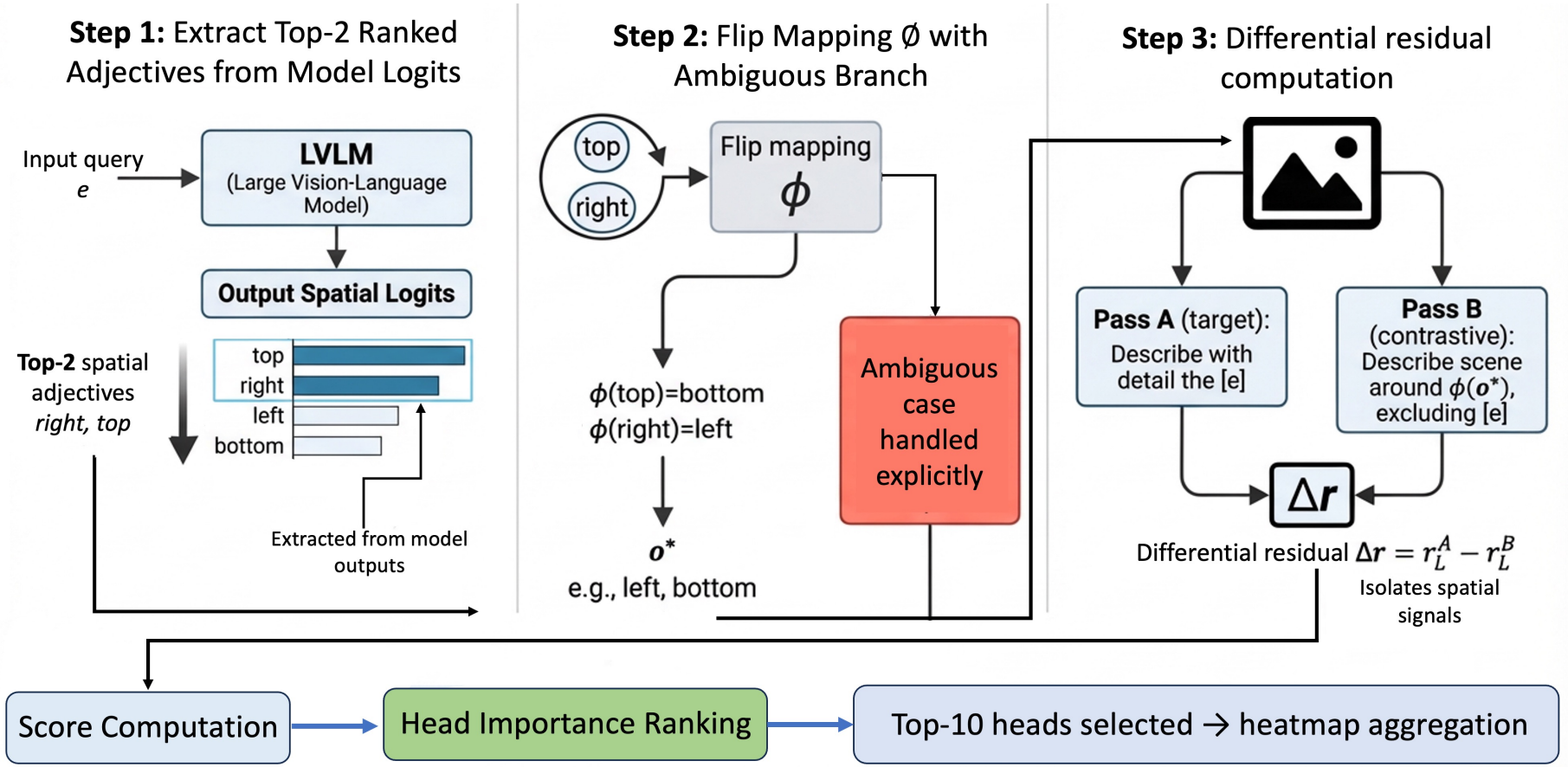
Contrastive Attention-Head Discovery: A Training-Free Framework

- **Objective:** Extract grounding from a frozen LVLM by identifying the few heads that encode localization.
- **Contrastive Dual Prompting:** Target vs. contrastive passes cancel shared semantics, **isolating the spatial signal**.
- **Three-Way Head Scoring:** (1) Project the differential residual onto each head's output (causal attribution) → (2) compare per-patch attention (attention shift) → (3) measure how target queries re-weight image keys (retrieval signal).
- **Top-k Head Selection & Aggregation:** Ranks heads, picks top-10, aggregates into a spatial heatmap.

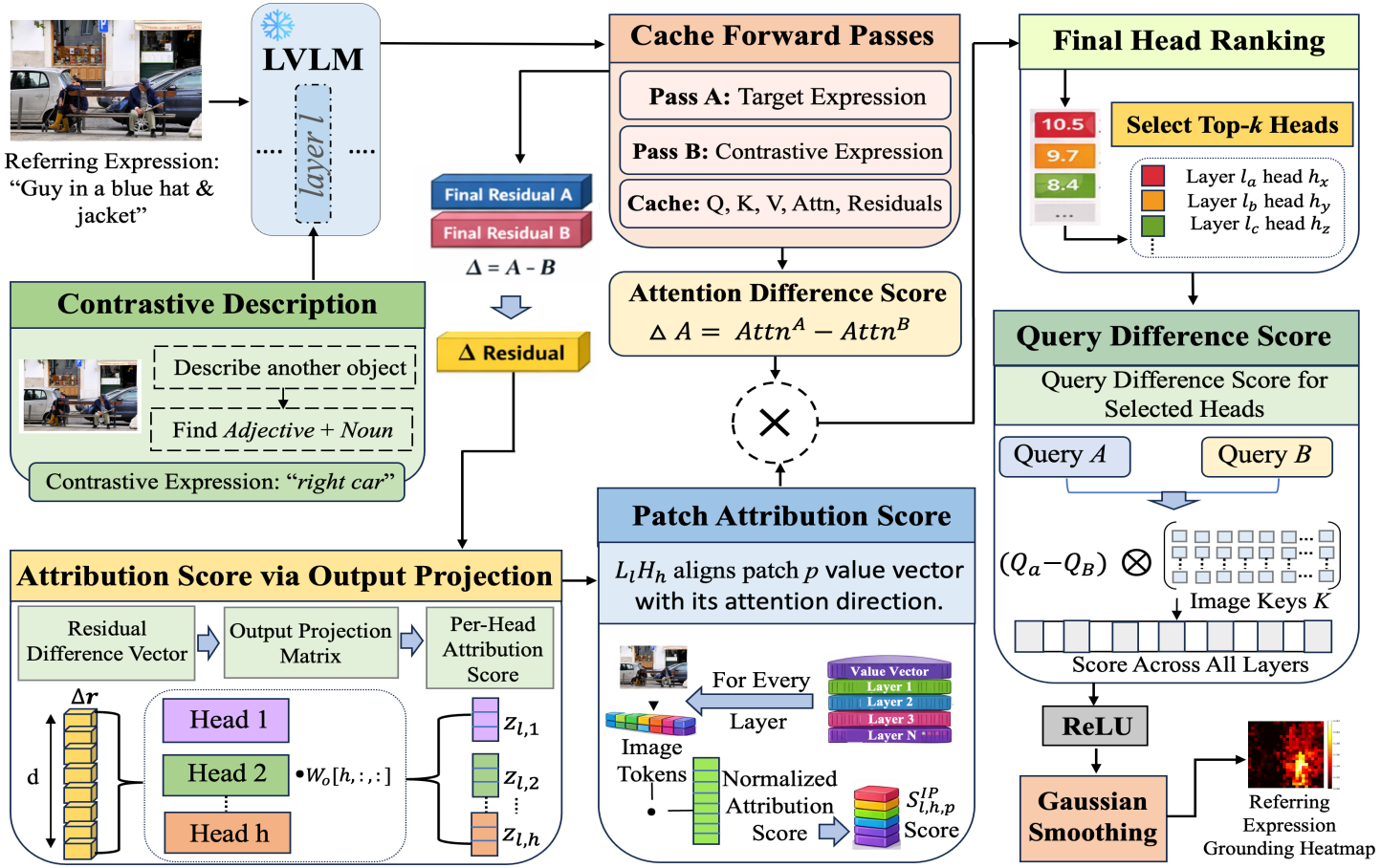


Contrastive Prompt Generation

- Contrastive prompts are generated per sample from the **model's own logits**, with no manual input.



Grounding-Head Discovery Criteria



- **Attribution (S^{OP}):** Does this head's output contribute to the target direction? → the head's causal role.
- **Attention Difference (Δa):** Which patches did the head shift attention toward? → the head's spatial focus.
- **Query Difference (S^{Q-diff}):** Which patches would the target query retrieve? → the head's retrieval signal.
- **Combined Ranking ($r^{l,h}$):** Head attends to target-relevant patches → final importance score.

Referring Expression Benchmarks

Input: Image + Text expression

Output: Localize the referred object

Benchmark	Images	Expressions	Splits	Focus
RefCOCO	~20K	~142K	Val, testA, testB	**Spatial reasoning**
RefCOCO+	~20K	~142K	Val, testA, testB	**Appearance attributes**
RefCOCOg	~26K	~95K	Val, test	**Long compositional expressions** (~8.4 words)
Flickr30K Entities	~30K	~244K	Val, test	**Cross-domain generalization** (non-COCO)

For RefCOCO and RefCOCO+: testA (person-focused), testB (object-focused)

Expression: Blurry bench behind the man but not the one he's sitting on

Expression: 2nd lime to the right on the second row from bottom

Expression: Motorcycle, second from bottom

Expression: Partial elephant far right

Localization Results for Vision-Language Models

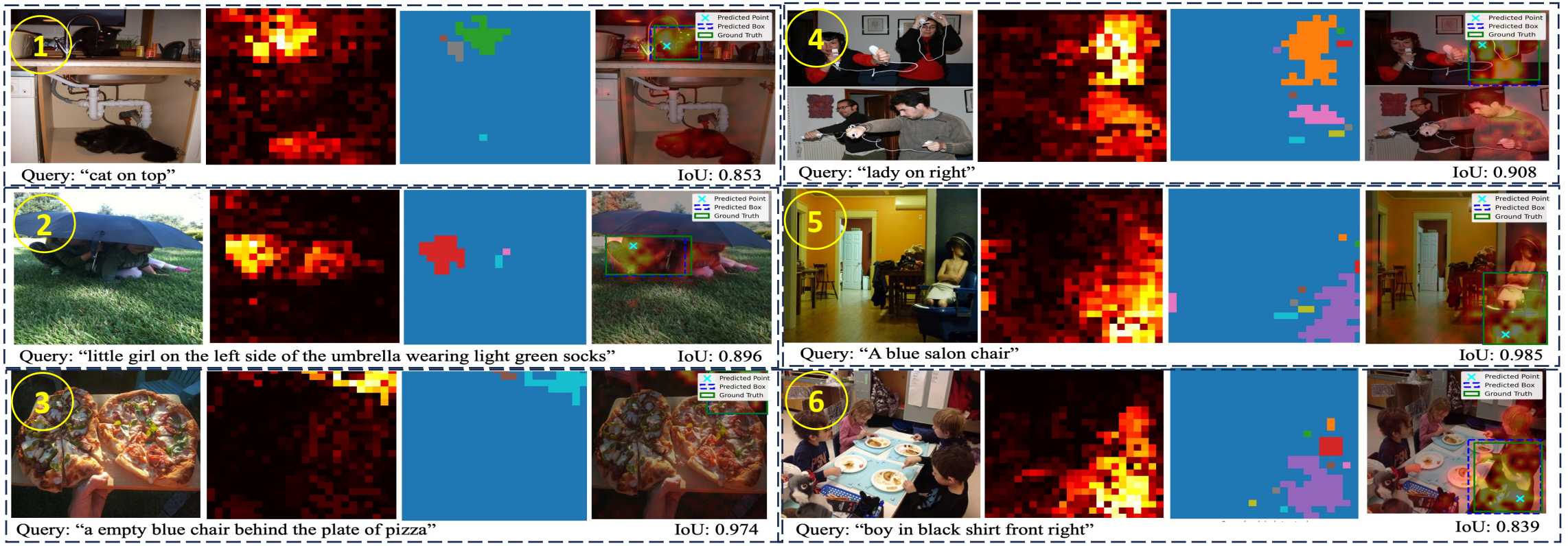
Model	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val	test
LLaVA-1.5-7B	70.17	76.35	63.04	59.11	69.52	47.27	63.38	63.07
LLaVA-1.5-13B	69.35	76.35	62.71	59.45	68.46	50.06	64.74	65.04
BakLLaVA-v1	66.96	74.03	56.57	56.45	67.44	45.27	59.26	59.75
Qwen2-VL-7B	63.57	68.40	59.60	57.59	59.28	47.40	54.20	52.40
LLaVA-1.5-7B*	79.49	84.76	74.42	67.70	77.57	58.95	71.72	71.53
LLaVA-1.5-13B*	78.82	83.80	73.59	68.15	76.91	59.07	71.81	73.64
BakLLaVA-v1*	74.91	81.67	69.94	63.39	73.27	53.56	66.43	65.50
Qwen2-VL-7B*	73.40	76.80	71.20	59.03	64.61	54.20	62.80	59.80

- The Red values indicate highest point localization within the object boundary (in %). Blue values indicate highest Acc@0.5 (in %).
- * shows point-in-box accuracy.

- Visual grounding performance of the proposed framework is evaluated across four LVLMs.
- We evaluate performance using two metrics:
  (1) Acc@0.5, the percentage of predictions with IoU ≥ 0.5
  (2) Point-in-box accuracy (point acc.), the percentage of predictions where the center point falls within the ground truth box.
- LLaVA-1.5-7B leads on spatial & attribute datasets; 13B leads on longer expressions.

Localization Results for Vision-Language Models

Each box displaying (from left) the input image → computed attention heatmap → connected component → final prediction on challenging cases involving **cluttered scenes**, **small** or **hard-to-see targets**, and **complex spatial layouts.**



Cross-Domain Evaluation

Cross-Domain Generalization

Flickr30K images are outside COCO → tests whether the method transfers to a different image distribution.

Best Training-Free Performance

LLaVA-1.5-7B reaches 64.23% (val) / 64.01% (test) → outperforms despite using no category labels → the **spatial signal is strong**.

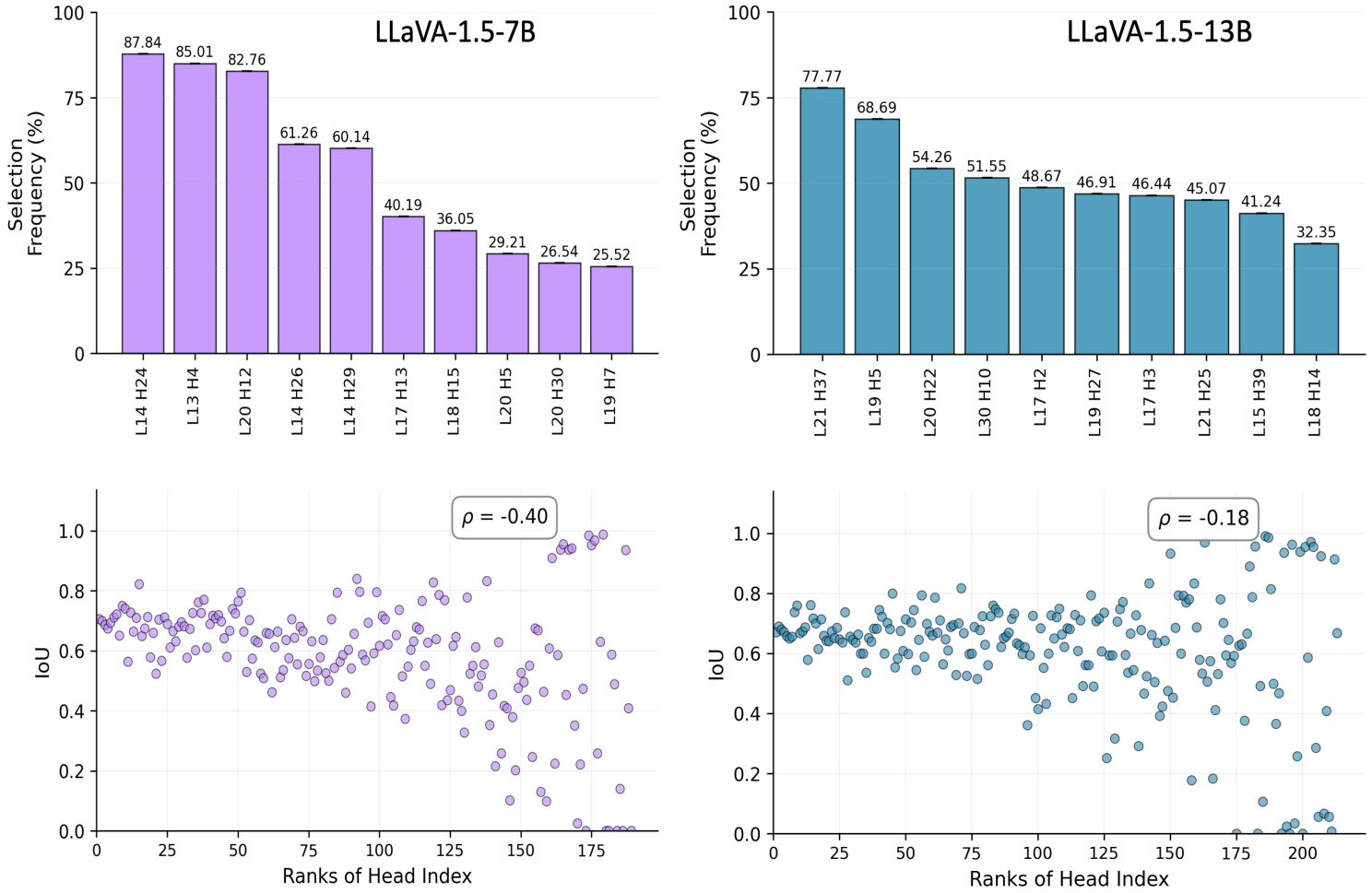
Consistent Across Four LVLMs

LLaVA-1.5-7B, 13B, BakLLaVA-v1, Qwen2-VL-7B all produce valid grounding on Flickr30K → **method is not tied to a single architecture**.

Model	Flickr30k Entities	
	val	test
LLaVA-1.5-7B	64.23	64.01
LLaVA-1.5-13B	63.39	52.58
BakLLaVA-v1	59.40	50.41
Qwen2-VL-7B	48.28	50.17
LLaVA-1.5-7B*	78.02	73.15
LLaVA-1.5-13B*	68.70	73.00
BakLLaVA-v1*	63.75	65.37
Qwen2-VL-7B*	46.94	39.85

- The Red values indicate highest point localization within the object boundary (in %). Blue values indicate highest Acc@0.5 (in %).
- * shows point-in-box accuracy.

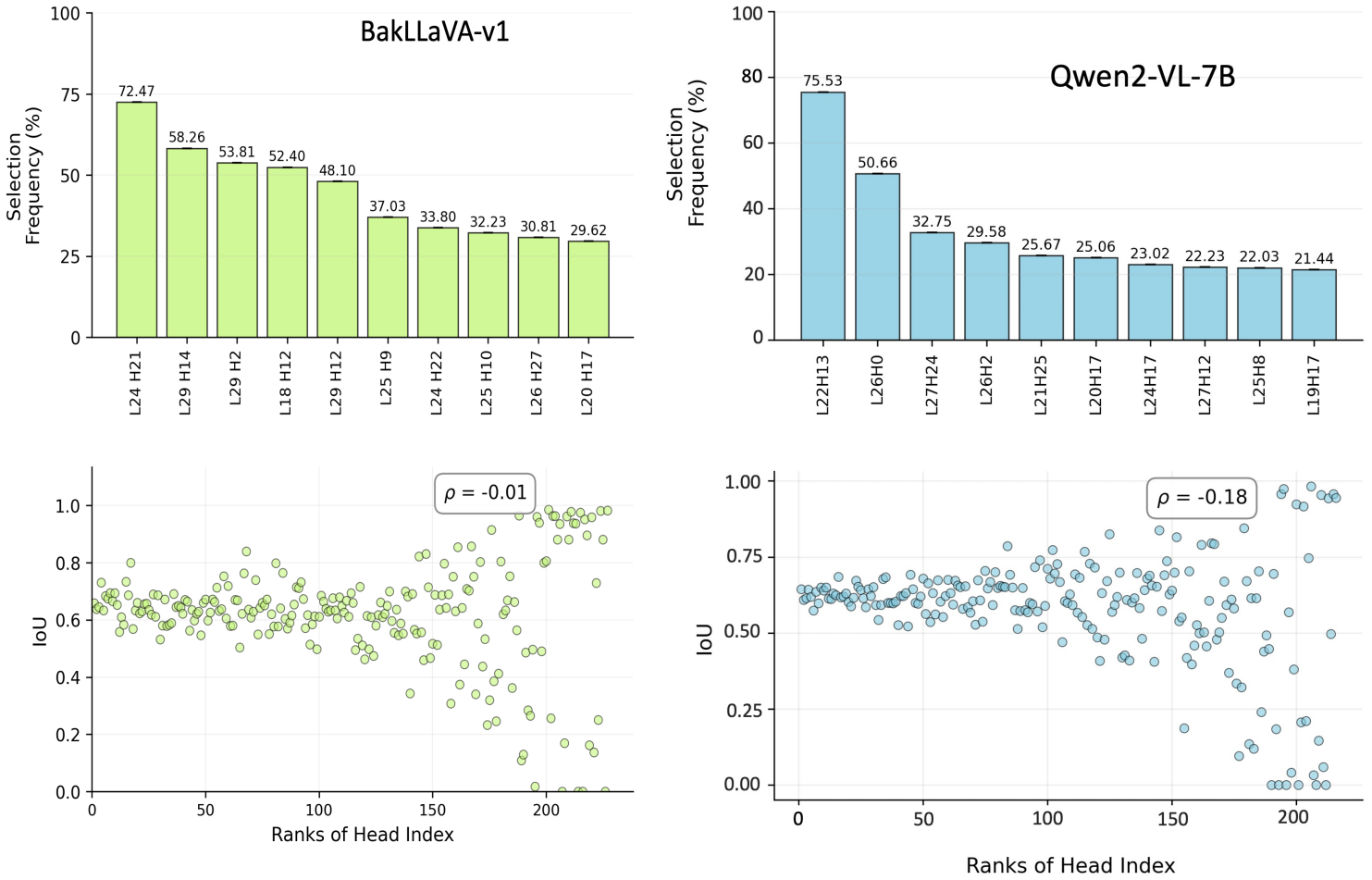
Evaluation of Attention Heads



- Top-3 heads (L14 H24, L13 H4, L20 H12) all above 80% frequency (for LLaVA-1.5-7B) .
- **Negative Spearman correlation** (ρ = −0.40 for LLaVA-1.5-7B) between selection rank and IoU → higher-ranked heads **localize better.**

(we denote layer l head h as $L_l H_h$)

Evaluation of Attention Heads

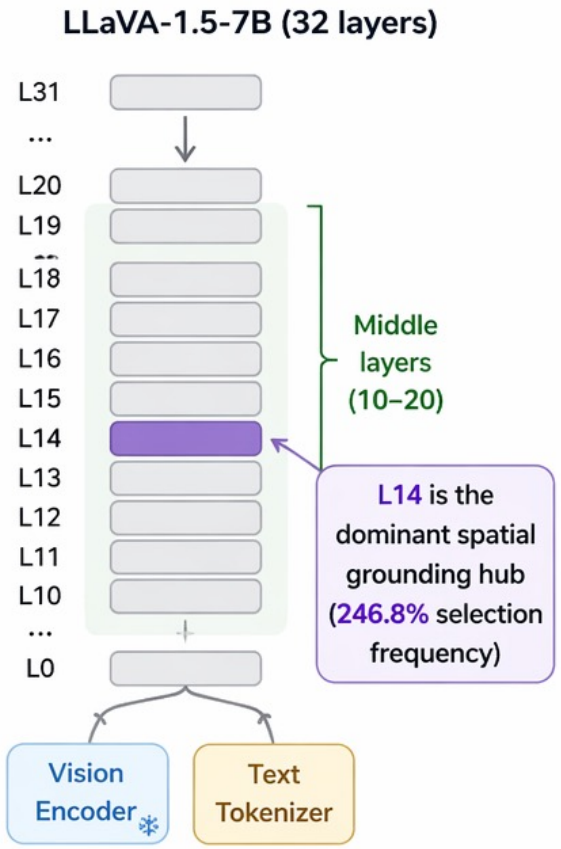


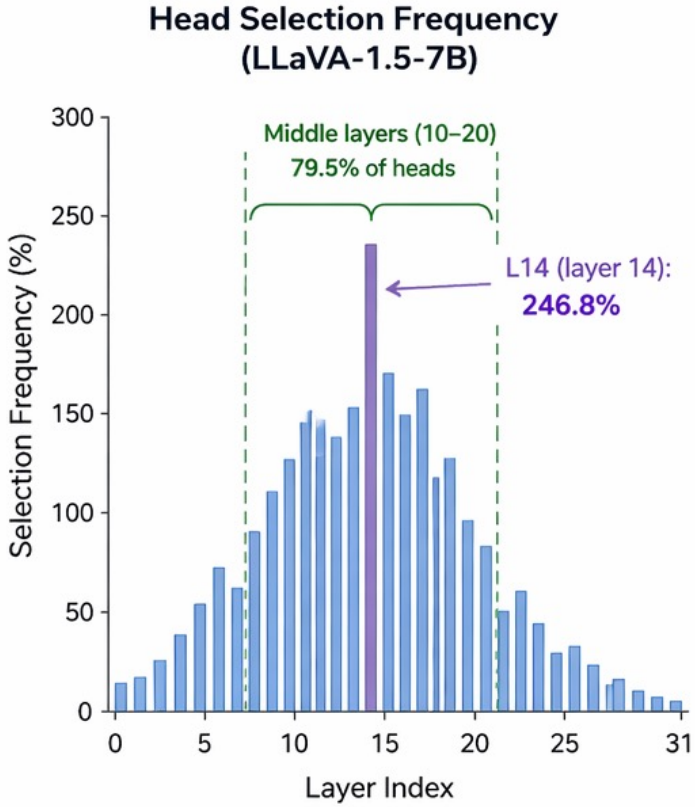
- **Head concentration holds in all four LVLMs** → a dominant head exists in each (BakLLaVA: L24 H21 at 72.5%; Qwen2-VL-7B: L22 H13 at 75.5%) → grounding heads are a general LVLM property, not a LLaVA artifact.

(we denote layer l head h as $L_l H_h$)

Layer Localization of Grounding Heads

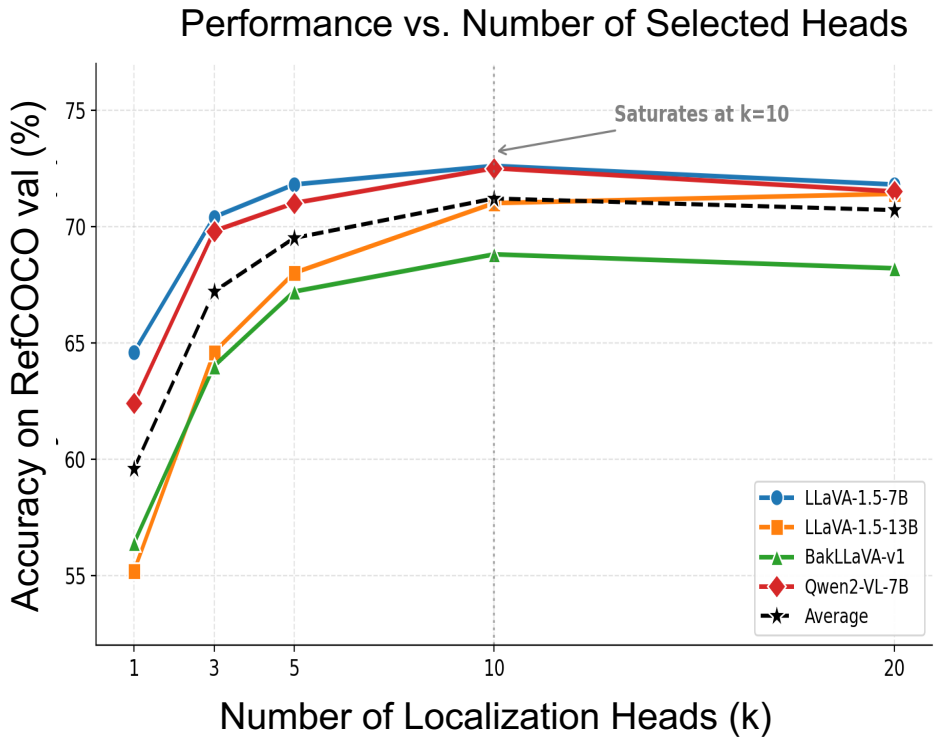
- LLaVA-1.5-7B selects 79.5% of heads from **middle layers** (10–20).
- L_{14} serves as the **dominant spatial grounding** hub (246.8% selection frequency*).
- This suggests that spatial grounding emerges at a **consistent relative depth** rather than a fixed absolute layer index.





*frequency represents the cumulative sum of selection rates across all heads within that layer, and may exceed 100% when multiple heads are each selected with high frequency.

Attention Heads Analysis



- Same heads recur across samples → grounding heads are intrinsic to the model, not per-image noise.
- **Top-10 head selection jumps from 22.5% → 79.5%**, raw attention vs. contrastive attention → the scoring identifies grounding heads.

Method	Point Acc.	mIoU	Acc@ 0.5
Raw Attention	22.52	27.95	26.59
Contrastive Attention	79.49 ↑	68.11 ↑	70.17 ↑

Causal Validation: Activation Patching

Goal: Prove the discovered heads causally carry the spatial signal not just correlate with it.

Corrupt → patch → measure restoration: Flip the image → patch the K=10 heads' clean activations into the flipped run → measure how much the decision is pulled back.

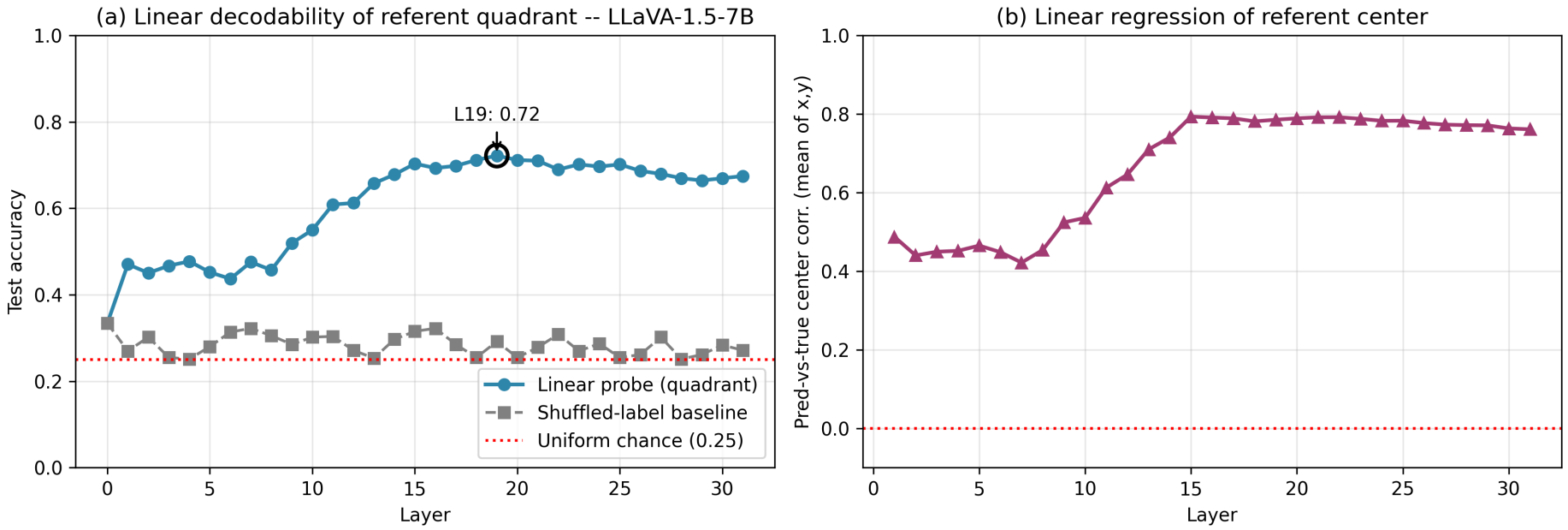
- **Selected heads restore the decision** in all settings → up to +0.88 logits (LLaVA-1.5-7B).
- Random heads restore ≈ 0 everywhere.
- One split (7B/testA): bottom heads restore +0.46.

Model	Clean→Flipped margin	Selected (K=10)	Random	Bottom
LLaVA-1.5-7B	+1.5 → −1.2	+0.51 to +0.88	≈ 0	≈ 0*
LLaVA-1.5-13B	+1.5 → −0.9	+0.30 to +0.50	≈ 0	≈ 0
BakLLaVA-v1	+2.2 → −1.6	+0.09 to +0.23	≈ 0	≈ 0
Qwen2-VL-7B	+2.4 → −1.6	+0.23 to +0.36	≈ 0	≈ 0

"Random" = 10 randomly chosen heads; "Bottom" = 10 never-selected heads

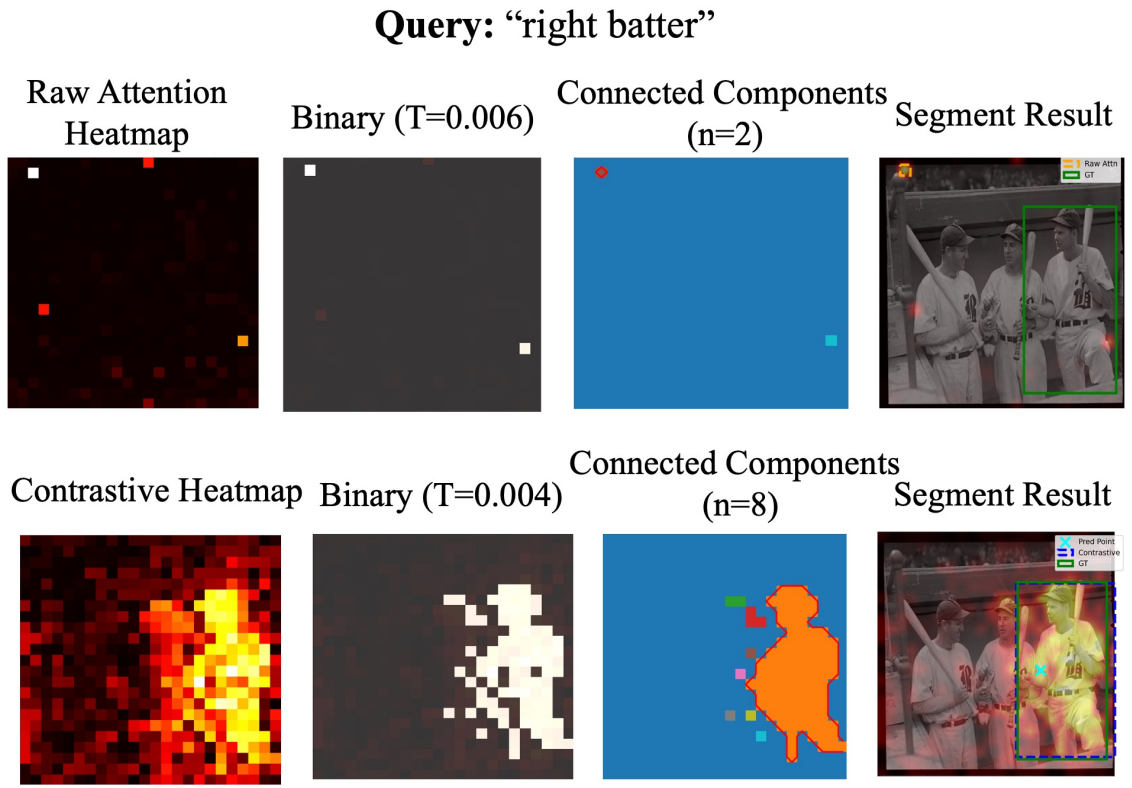
Probing Validation: Spatial Location Is Linearly Decodable

At each layer, a simple linear probe tries to read the object's location from the hidden states → it succeeds at up to 72% (vs. 29% with shuffled labels) → **spatial information is present and can be extract from the middle layers.**



Mask Analysis of Localization Heads

- Localization heads **group patches by shared object features** → all patches of ”person" activate similar representations.
- **Contrastive attention suppresses background** → preserves **spatially-coherent, mask-like patterns**.
- **Masks align closely with ground-truth boundaries** → the mechanism generalizes beyond localization.



Conclusion: **Zero-shot contrastive head discovery** gives **precise object grounding** without fine-tuning or architectural modifications.

Thank You

KAIST VIL Vehicular Intelligence Laboratory