Finding Distributed Object-Centric Properties in Self-Supervised Transformers

CVPR 2026 (Highlight)

Samyak Rawlekar[1] Amitabh Swain[1] Yujun Cai[2] Yiwei Wang[3] Ming-Hsuan Yang[3] Narendra Ahuja[1]

[1]University of Illinois Urbana-Champaign [2]University of Queensland [3]UC Merced

CVPR JUNE 3-7, 2026 DENVER COLORADO

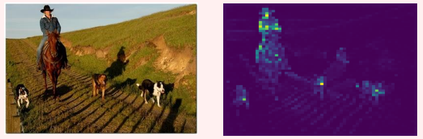
WHERE DO OBJECTS LIVE IN A ViT?

Self-supervised ViTs (DINO) discover objects, visible in [CLS] attention of the final layer.

But **[CLS] token** is noisy:

- **Spurious activations** on background regions
- **Missed or partial** object coverage

[CLS] Attention



Noisy & Incomplete

WHERE IS THE OBJECT-CENTRIC INFORMATION?

Prior works used Keys (K) from Final-layer

But two questions remain:

1. Do only Keys carry object-centric information?
2. Is this information limited to the final layer?

We answer both in our Findings

CONTRIBUTIONS

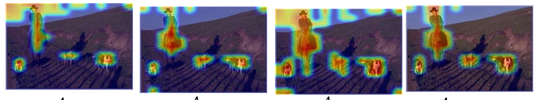
1. Object-centric info encoded in Q, K, and V similarity
2. Object-centricity is distributed across layers
3. Object-DINO: training-free head clustering to extract distributed object-centric information
4. Validated on unsupervised object discovery and MLLM hallucination mitigation

FINDINGS

FINDING 1: Q, K, V ALL ENCODE OBJECT STRUCTURE

For each attention component $r \in \{Q, K, V\}$, we compute **patch self-similarity matrix** :

$$A_{\ell,h_r} = \text{softmax}\left(r_{\ell,h} \cdot (r_{\ell,h})^{\mathrm{T}} / \tau\right)$$

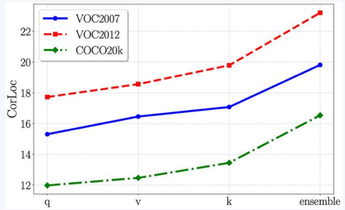


A_q A_k A_v A_{ens}

Each component captures a **complementary view** of object structure

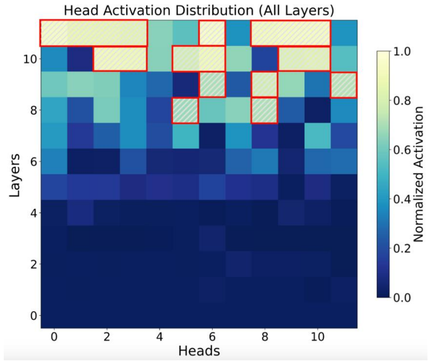
ENSEMBLE SIMILARITY $A_{ens} = w_q A_q + w_k A_k + w_v A_v$

Component Ablation (CorLoc on VOC2007) : Q < V < K < **Ensemble**



FINDING 2: OBJECT-CENTRICITY IS DISTRIBUTED

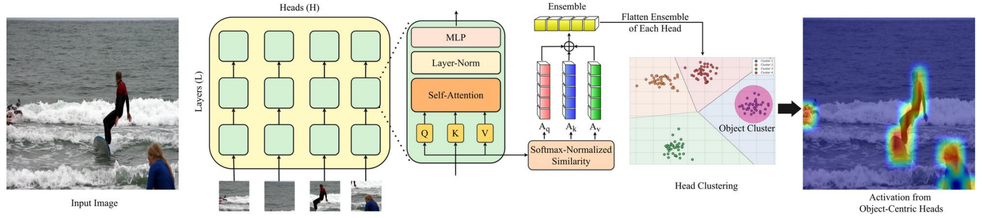
We cluster head ensemble maps over 4,000 COCO images and find object-centric heads span **multiple layers**:



4/12 final-layer heads are **noisy** (low frequency)

Many object-centric heads are in **Layers 8–10**

OBJECT-DINO OVERVIEW

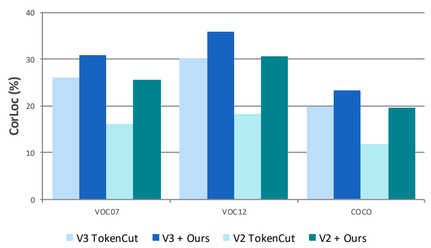


Q, K, V Similarity → **Ensemble** → **Flatten** → **Cluster Heads** → **Object Map**

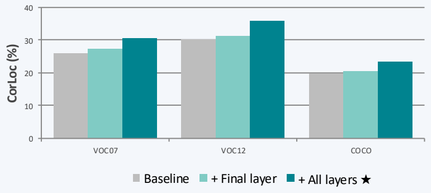
APPLICATIONS

APPLICATION 1: UNSUPERVISED OBJECT DISCOVERY

We integrate Object-DINO into **TokenCut** by replacing its final-layer key features with our distributed object-centric head ensemble.
Metric: **CorLoc** (% images IoU > 0.5).



LAYER ABLATION: WHERE DOES THE GAIN COME FROM?



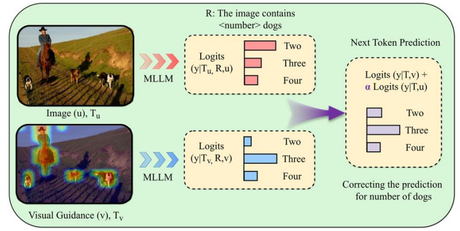
Intermediate layers contribute **+3.3 / +4.6 / +2.9** beyond final-layer-only

CONCLUSION

- **Object-centric info** is in Q, K, and V, not just K
- **Distributed across layers** — intermediate heads carry signal
- **Object-DINO** discovers this set — no training, no labels
- Gains: +3.6–12.4 (discovery), best POPE/CHAIR (hallucination)

APPLICATION 2: MITIGATING MLLM HALLUCINATION

Object-DINO provides explicit **visual grounding** to reduce object hallucination:



POPE BENCHMARK (PRECISION / F1)

Method	LLaVA-1.5	InstructBLIP	Qwen-VL
Regular	73.3 / 79.2	71.2 / 76.4	80.1 / 79.7
VCD	72.1 / 79.4	74.2 / 78.4	80.6 / 81.5
M3ID	73.5 / 80.2	73.6 / 79.1	81.4 / 82.1
RITUAL	74.4 / 80.5	74.5 / 80.3	83.1 / 82.7
DeGF	80.5 / 81.9	80.9 / 80.1	84.4 / 82.9
Ours	**87.4 / 82.7**	**87.7 / 81.6**	**89.2 / 86.1**

CHAIR BENCHMARK (LOWER = LESS HALLUCINATION)

Method	LLaVA-1.5 (s/i)	InstructBLIP (s/i)
Regular	26.2 / 9.4	31.2 / 11.1
DeGF	18.4 / 6.1	24.0 / **7.7**
MARINE	20.8 / 6.0	27.5 / 8.8
Ours	**18.4 / 5.9 ↓**	**21.4** / 8.8

Personal Page

Project Page