Finding Distributed Object-Centric Properties in Self-Supervised Transformers

Supplementary Material

This supplementary material provides additional experimental details and ablation studies to support our work. We show: (1) Evaluation of the optimal number of clusters K for our head selection algorithm using the Davies-Bouldin Score, (2) Ablation study on the guidance weight α in our MLLM hallucination mitigation method, (3) Analysis of the temperature parameter τ from Equation 1 and its impact on object-centric signal quality, and a visualization demonstrating how visual patterns evolve across different temperature values, (4) Visualization of individual head behaviors in the final layer, (5) Ablation study of ensemble weight W_q, W_k, W_v in eq. 2, (6) Computational efficiency analysis for MLLM hallucination mitigation, (7) Analysis of object-centric heads across ViT architecture and reconstruction-based self-supervised models, (8) Evaluation Metrics and Datasets used. (9) Qualitative comparison of MLLM-generated captions.

8. Optimal Number of Clusters K

A critical hyperparameter in our Object-DINO algorithm is the number of clusters K used in the k-means clustering step (Algorithm 1, line 13). To determine the optimal value, we evaluate multiple candidate values of K using the Davies-Bouldin (DB) Score, a standard metric for assessing clustering quality.

Davies-Bouldin Score The Davies-Bouldin Score measures the average similarity between each cluster and its most similar cluster, indicating cluster separation and compactness. The equation is given by:

$$\mathrm{DB}(K) = \frac{1}{K}\sum_{i=1}^{K}\max_{j\neq i}\left(\frac{s_i + s_j}{d_{ij}}\right) \tag{6}$$

where, s_i is the average distance between each point in cluster i and the cluster centroid d_{ij} is the distance between cluster centroids c_i and c_j

Experimental Setup. We evaluate K ∈ 3, 4, 5, 6, 7, 8, 9, 10 on 500 randomly sampled images from the COCO dataset using DINO-V3 features. For each value of K, we compute the DB Score.

Results As shown in Fig. 8, the DB Score remains relatively stable across the evaluated range, varying between 1.717 and 1.737. The minimum score occurs at K = 5, suggesting this configuration achieves the best balance between cluster separation and compactness. Based on this, we use K = 5 for all experiments in the main paper.

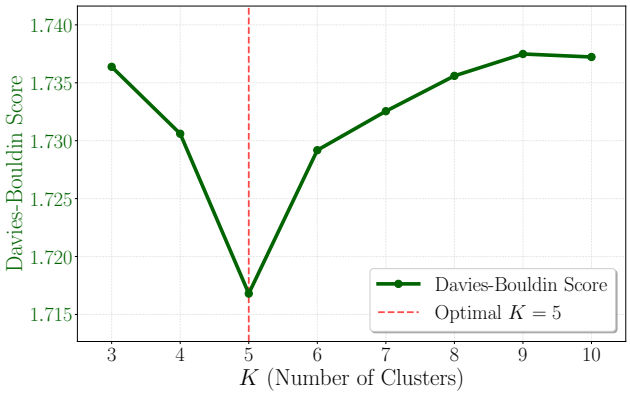


Figure 8. Evaluation of optimal number of clusters K using Davies-Bouldin Score (lower is better). We select $K = 5$ for all our experiments.

9. Ablation on Guidance Weight for MLLM Hallucination Mitigation

Our MLLM hallucination mitigation method (Sec. 4.2) introduces a guidance weight α that controls the strength of the visual grounding signal (Eq. 5):

$$L = \alpha\,\mathrm{Logits}(y|T_u, R, u) + (1-\alpha)\mathrm{Logits}(y|T_v, R, v) \tag{7}$$

When we set α = 0, it is equivalent to the original MLLM output with regular decoding, while a high alpha ensures grounding information from the Object-DINO map. We perform an ablation by varying α across a range ([0.3, 0.4, 0.5, 0.6, 0.7]) and evaluate performance on the CHAIR benchmark. Based on the Fig. 9, we select α = 0.4 for our experiments as it demonstrates the lowest score for both C_s and C_i.

10. Ablation on Temperature Parameter

The temperature parameter τ in Eq. 1 controls the sharpness of the patch self-similarity matrices (A_q, A_k, A_v). It scales the dot-product operation before the softmax:

$$A_r^{\ell,h} = \mathrm{softmax}\left(\frac{\tilde{r}^{\ell,h}\cdot(\tilde{r}^{\ell,h})^\top}{\tau}\right) \tag{8}$$

We investigate how τ affects the quality of object-centric signals. We evaluate $\tau \in$ 0.03, 0.1, 0.3, 1, 3, 10, 30, 100 on a subset of COCO images. Fig. 10 shows the variation of τ and how it impacts the object-centric score. Based on this, we identify that selecting a $\tau \in (10, 100)$ leads to optimal object-centric signals. We further show the evolution of τ in Fig. 11.

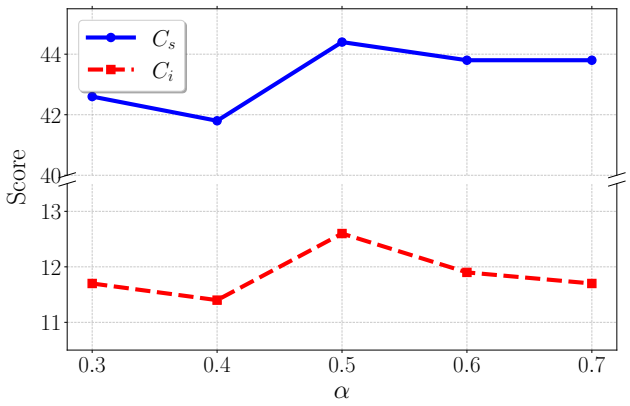


Figure 9. Guidance weight ablation for MLLM hallucination mitigation. We observe that the C_s and C_i scores are lowest for α = 0.4, making it our optimal choice

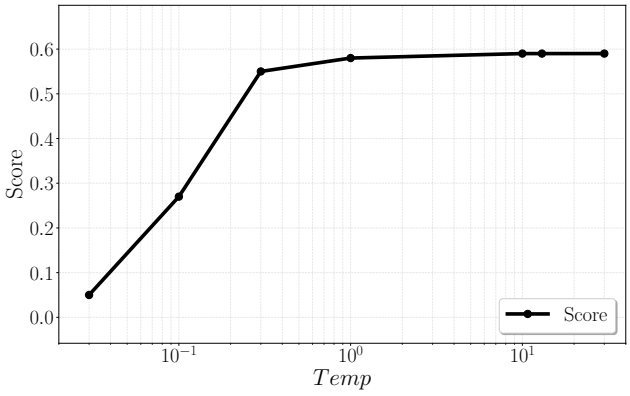


Figure 10. **Tau ablation.** We show the optimal choice of τ for object centric signal is $\in (10, 100)$

11. Final Layer Head Visualization

To illustrate the diversity of attention head specializations and support our claim that not all final-layer heads are object-centric, in fig. 12 we visualize the ensemble similarity maps (A_{ens}) for all 12 attention heads in the final layer (Layer 11) of DINO-V3. These plots are consistent with the analysis shown in Fig. 2, where heads 5, 6, 7, and 11 are not selected as object-centric because head 5 and head 6 miss regions of the man, and head 7 and head 11 include a large portion of the background.

12. Ablation of Ensemble Weights

In this section, we ablate the ensemble weights (W_q, W_k, W_v) from Eq. 2. The results in Table 6. shows CorLoc performance on VOC2007 with different weight configurations. We observe that heavily weighting the Key component $W_k = 0.7$ yields a CorLoc of 17.9. Performance progressively increases as the weights become more balanced, peaking at 19.8 CorLoc with a uniform average ($W_q = W_k = W_v = 0.33$). This confirms that all three components contribute valuable and complementary object signals. We use this average for A_{ens} in all experiments.

Table 5. **Ablation of Ensemble Weights.** We report CorLoc for different weight configurations from Eq. 2. Performance peaks at 19.8 with a uniform average ($W_q = W_k = W_v = 0.33$), confirming that all three components contribute valuable and complementary object signals. We therefore use this uniform average for A_{ens} in all experiments.

W_q	W_v	W_k	**CorLoc**
0.1	0.2	0.7	17.9
0.2	0.3	0.5	19.1
0.3	0.3	0.4	19.4
0.33	0.33	0.33	19.8

Table 6. **Efficiency comparison.** For each method, we show the inference latency per instance, peak GPU memory and the performance CHAIR_S on the LLaVA-1.5 model with max token 128

Method	**Avg. Latency (s)**	**GPU Memory (MB)**	**CHAIR_S (↓)**
Regular	3.44 (×1.00)	15,778 (×1.00)	55
VCD	6.91 (×2.01)	16,634 (×1.05)	54.4
OPERA	24.70 (×7.18)	22,706 (×1.44)	52.6
Woodpecker	10.68 (×3.10)	22,199 (×1.41)	57.6
HALC	22.61 (×6.51)	23,084 (×1.46)	51
DeGF	13.89 (×4.04)	19,119 (×1.21)	48.8
Ours	7.1 (×2.06)	16,966 (×1.07)	41.6

13. Computational Efficiency Analysis for MLLM Hallucination Mitigation

In Table. 6 we compare the efficiency and performance of our method against prior works for mitigating object hallucination on the LLaVA-1.5 model. Our approach, which leverages Object-DINO as a training-free visual grounding mechanism, achieves a state-of-the-art CHAIR_S score of 41.6. This significantly outperforms all competing methods, including DeGF (48.8) and HALC (51). Critically, this performance is achieved without the substantial computational overhead seen by other methods. While approaches like OPERA, HALC, and DeGF incur massive latency penalties (ranging from 4.04x to 7.18x the regular baseline) and significant GPU memory increases (1.21x to 1.46x), our method maintains a minimal computational cost. Our latency (7.1s, 2.06x) and peak memory (16,966MB, 1.07x) are only slightly above the baseline and are comparable to the much lower-performing VCD method, demonstrating a vastly superior balance of performance and efficiency

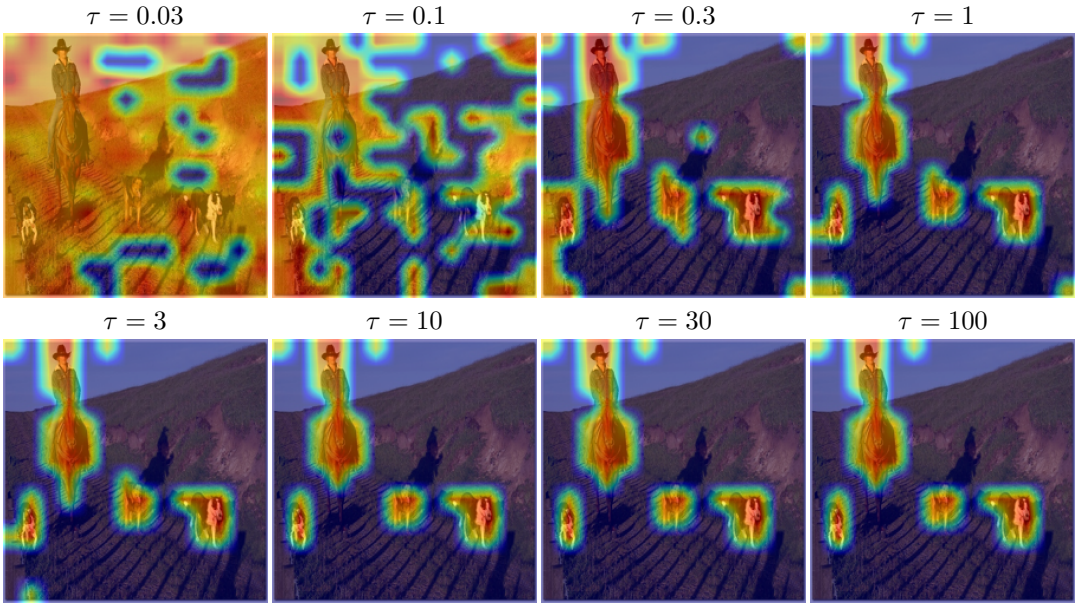


Figure 11. Evolution of localizations with temperature. For visualization, we invert the maps to show bright colors for objects.

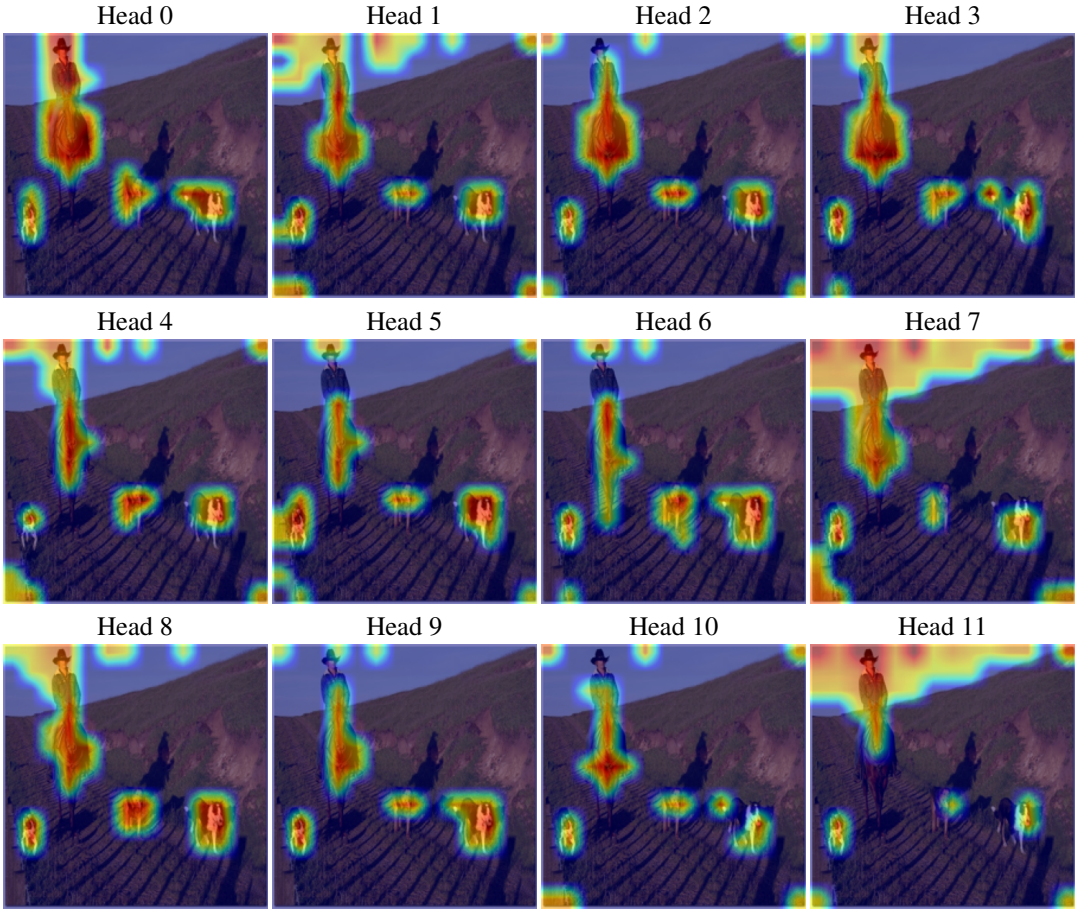
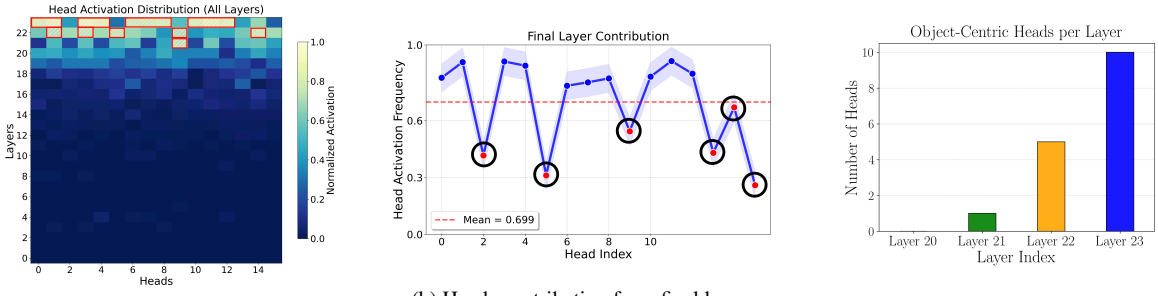


Figure 12. **Final Layer Head Visualization.** Ensemble similarity maps A_{ens} for all 12 heads in Layer 11 of DINO-V3. For visualization, we invert the maps to show bright colors for objects.



(a) Head contribution for complete network

(b) Heads contribution from final layer.

(c) Highly active heads per layer.

Figure 13. Analysis of the object-centric head distribution in ViT-L, computed over 4,000 images from the COCO dataset. (a) The heatmap shows the frequency of heads (across all 24 layers) belonging to the object-centric cluster. The red boxes highlight that numerous heads in later layers (e.g., Layers 20-23) are consistently selected, demonstrating that object-centric information is a distributed phenomenon and not confined to the final layer. (b) The plot details the final layer's contribution, showing that several heads (circled in black) have low frequency, confirming that some final-layer heads are "noisy" (non-object-centric). (c) The histogram shows the number of strongly active object-centric heads per layer, with Layer 23 containing 10 heads, Layer 22 containing 5 heads, and Layer 21 containing 1 head.

14. Architectural and Objective Analysis of Object-Centric Heads

Architectural Analysis. To confirm that object-centric processing is a general property of Vision Transformers, we extended our analysis to ViT-L, a deeper 24-layer, 16-head architecture. The analysis was conducted on the same 4,000-image COCO dataset from the main paper. The results (Figure 13) are consistent with our primary findings:

1. Distributed Object-Centric Heads: We observed that object-centric heads are not isolated in the final layer but are distributed across the network. As with other models, these heads are most heavily concentrated in the later layers (specifically 20-23).

2. Final-Layer Heterogeneity: The final layers are not uniformly object-centric. Layer 23 (Panel b) shows significant variance in head activation frequency (mean = 0.699). Crucially, we again identified several "noisy" heads (circled in black) with substantially lower contributions.

These findings confirm that the observed distribution of specialized and non-specialized heads is a general characteristic across different ViT architectures.

Objective Analysis. Here, we extend our analysis to Masked Autoencoders (MAE) [11], which uses a reconstruction-based self-supervised objective. Unlike DINO's contrastive approach that explicitly encourages semantic similarity through teacher-student distillation, MAE learns representations by reconstructing masked image patches. This fundamental difference in training objectives raises an important question: Does object-centric information still emerge and distribute across the network in reconstruction-based models?

Experimental Setup. We apply our Object-DINO analysis framework to a MAE ViT-B/16 model pre-trained on ImageNet. Following the same protocol as our DINO experiments, we compute the ensemble similarity matrices (A_ens) for all attention heads across all layers using 4,000 images from the COCO dataset. We then perform k-means clustering (K=5) to identify object-centric heads.

Qualitative Analysis. Figure 14 shows representative examples of the patch-level similarity maps (A_q, A_k, A_v) and their ensemble from a MAE model. We observe that MAE does encode object-centric information in its attention components. The ensemble maps show recognizable foreground-background separation and partial object localization. However, we observe differences compared to DINO. First, the individual component maps (A_q, A_k, A_v) are considerably noisier, with more spurious activations in background regions. Second, the ensemble map, while identifying the general object location, does not produce sharp boundaries and often includes substantial background.

Quantitative Head Distribution Analysis. Figure 15 shows how object-centric heads distribute across MAE's architecture. (a) Shows the frequency heatmap of heads belonging to the object-centric cluster across all 12 layers. Similar to DINO, we observe that object-centric information is distributed throughout the network rather than confined to the final layer. (b) Examines the final layer (Layer 11) specifically and reveals substantial heterogeneity—several heads (circled in black) show low activation frequencies, indicating they are non-object-centric. (c) Quantifies the number of highly active object-centric heads per layer

15. Evaluation Metrics and Datasets

In this section, we provide detailed descriptions of the evaluation metrics used throughout our paper. We organize them by task: unsupervised object discovery (CorLoc) and multi-modal hallucination mitigation (CHAIR and POPE).

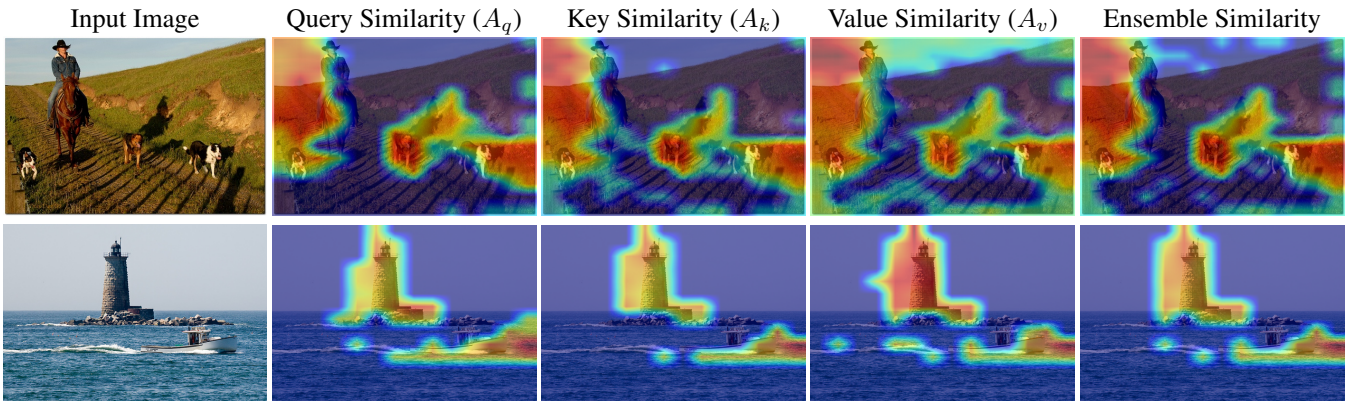


Figure 14. **Object-centric information in patch-level interactions from MAE [11].** We visualize the inter-patch similarity maps (A_q, A_k, A_v) computed from the Query, Key, and Value representations of patch tokens in a MAE ViT-B/16 model. For visualization, we invert the similarity maps so objects appear bright.

15.1. CorLoc: Correct Localization

CorLoc is the standard metric for evaluating unsupervised object discovery methods. It measures the percentage of images where the predicted bounding box correctly localizes at least one ground-truth object instance.

For a dataset $\mathcal{D}$ containing N images, let $B_{\text{pred}}^{(i)}$ denote the predicted bounding box for image i, and let $\mathcal{B}_{\text{gt}}^{(i)} = \{B_{\text{gt}}^{(i,1)}, B_{\text{gt}}^{(i,2)}, \ldots, B_{\text{gt}}^{(i,m_i)}\}$ denote the set of m_i ground-truth bounding boxes for that image. The Intersection over Union (IoU) between two boxes is defined as:

$$\text{IoU}(B_1, B_2) = \frac{\text{Area}(B_1 \cap B_2)}{\text{Area}(B_1 \cup B_2)} \tag{9}$$

An image is considered correctly localized if the predicted box has IoU > 0.5 with at least one ground-truth box:

$$\text{CorLoc} = \frac{1}{N} \sum_{i=1}^{N} \text{Correct}^{(i)} \times 100\% \tag{10}$$

We report CorLoc on three standard benchmarks:

- **PASCAL VOC 2007 & 2012** [9]: These datasets contain images from 20 object categories with bounding box annotations.
- **COCO 20k** [25]: A subset of 20,000 images from the MS-COCO dataset used for evaluating unsupervised discovery methods.

15.2. CHAIR

CHAIR [21] quantifies object hallucination in image captioning by measuring how often models mention objects that are not actually present in the input image.

For each image, we extract all objects mentioned in the generated caption and compare them against the ground-truth objects present in the image (from MS-COCO annotations). An object is considered hallucinated if it appears in the caption but not in the ground-truth. CHAIR provides two complementary metrics:

CHAIR_S (Sentence-level):

$$\text{CHAIR}_\text{S} = \frac{\text{\# captions with hallucinated objects}}{\text{\# all captions}} \times 100\% \tag{11}$$

This measures the percentage of captions that contain at least one hallucinated object.

CHAIR_I (Instance-level):

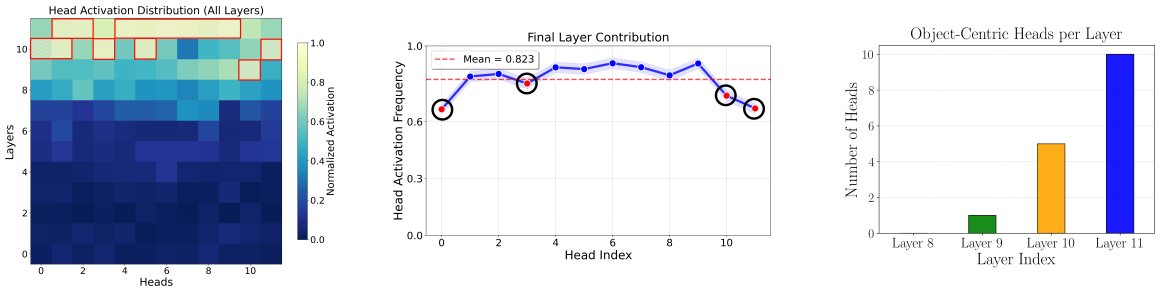
$$\text{CHAIR}_\text{I} = \frac{\text{\# hallucinated objects}}{\text{\# all mentioned objects}} \times 100\% \tag{12}$$

15.3. POPE

POPE [14] evaluates whether MLLMs can accurately determine object presence through binary yes/no questions. Unlike CHAIR, which evaluates free-form generation, POPE directly tests object recognition.

For each image, POPE generates questions in the format: "Is there a {object} in the image?" The questions include both objects that are present (positive samples) and objects that are not present (negative samples). Negative samples are created using three strategies: Random, Popular (frequently co-occurring objects), and Adversarial (objects that typically appear together with present objects).

POPE reports three standard classification metrics: Accuracy, Precision, F1-Score. F1 balances precision (avoiding hallucinations) and recall (detecting actual objects). We evaluate on the standard POPE benchmark constructed from MS-COCO validation images following the standard protocol [14].



(a) Head contribution for complete network

(b) Heads contribution from final layer.

(c) Highly active heads per layer.

Figure 15. Analysis of the object-centric head distribution in MAE [11], computed over 4,000 random images from the COCO dataset. (a) The heatmap shows the frequency of heads (across all 12 layers) belonging to the object-centric cluster identified by our k-means clustering (K=5). (b) The plot shows the final layer's contribution, showing that some heads (circled in black) have low frequency. (c) The histogram shows the number of strongly active object-centric heads per layer.

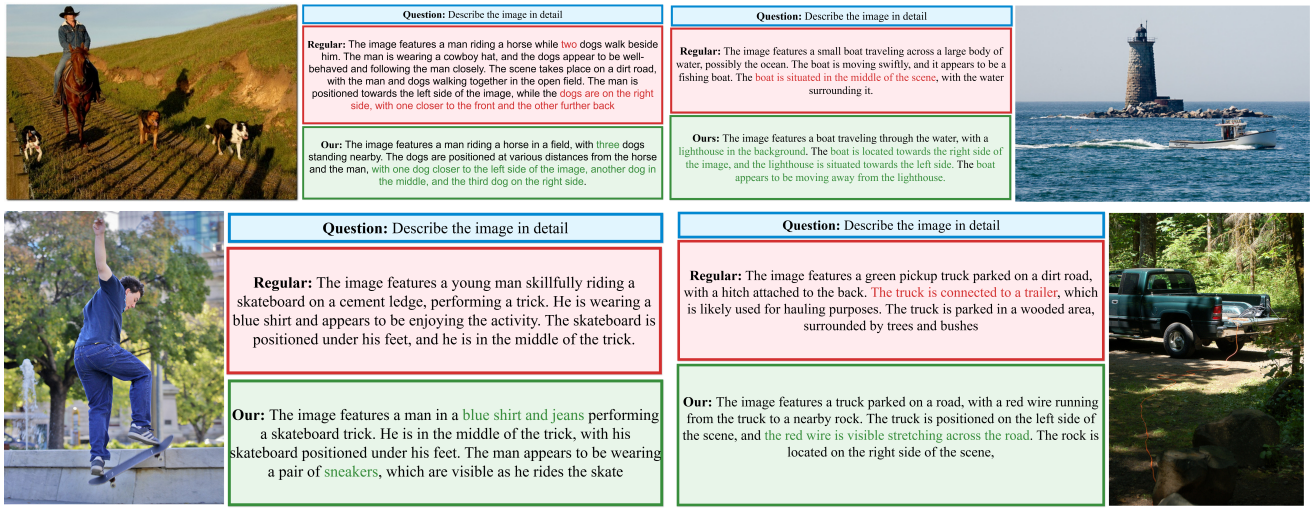


Figure 16. Qualitative comparison of captions generated by Regular decoding (Red) and Ours (Green)