What Do Visual Tokens Really Encode? Uncovering Sparsity and Redundancy in Multimodal Large Language Models

Supplementary Material

A. Evaluation Setups

General Tasks We assess general visual question answering ability on four standard multimodal benchmarks: GQA, MME, $\text{MMBench}^{\text{en}}_{\text{dev}}$, and MMStar [6, 12, 17, 32]. GQA contains compositional reasoning questions over real-world images. MME perception is a comprehensive evaluation suite covering 14 perception subtasks. $\text{MMBench}^{\text{en}}_{\text{dev}}$ is a multiple-choice benchmark with about 3k questions over 20 ability dimensions; MMStar is a vision-indispensable benchmark with 1.5k carefully curated samples that cover 6 core capabilities and 18 axes. The *General VQA* score reported in the main text is the unweighted mean of the normalized scores on these four datasets, the score for MME is divided by 20.

OCR Tasks To evaluate text-centric perception, we use $\text{TextVQA}_{\text{val}}$, OCRBench, and DocVQA [33, 36, 44]. TextVQA requires reading scene text and answering open-ended questions.OCRBench is a composite OCR benchmark covering text recognition, scene-text VQA, document VQA, key information extraction, and handwritten math expression recognition. DocVQA measures document understanding via question answering on document images. The *OCR* score is the unweighted mean of the normalized TextVQA, OCRBench, and DocVQA scores.

CV Centric Tasks We use referring expression comprehension (REC) benchmarks to measure fine-grained, localization-centric vision ability. Specifically, we evaluate on RefCOCO+ REC val and RefCOCO REC testA/B [22], which localize a target region in a COCO image given a natural-language referring expression. Following common practice, we use the Acc@0.5 metric, which is the standard detection metric for REC. The *CV-Centric* score is the average accuracy over the three REC splits (RefCOCO+ val, RefCOCO testA, and RefCOCO testB).

Hallucination Hallucination robustness is evaluated on POPE and HallusionBench [14, 24]. POPE is a polling-based object probing benchmark that tests whether a model correctly judges the existence of objects using yes/no questions under random, popular, and adversarial sampling strategies, we report F1 scores. HallusionBench is an image-context reasoning benchmark designed to disentangle language hallucination and visual illusion, we use DeepSeek-V3.2-Exp Chat [10] as the evaluator and we report the All Accuracy(aAcc) . The overall *Hallucination* score reported in the paper is the mean of the normalized POPE and HallusionBench scores.

A.1. Detailed Results

This section provides the full experimental results corresponding to the three summary tables in the main text.

- For Tab. 1 (impact of sink pruning and sublayer skipping), detailed metrics are reported in Appendix Tab. a.
- For Tab. 2 (effects of pruning dead tokens), complete results are provided in Appendix Tab. b.
- For Tab. 3 (training-based visual processing analysis), detailed results appear in Appendix Tab. c.

B. Implementations of Clustering Analysis

B.1. Intra-image Visual Embeddings Clustering

Anchor-based Clustering. We perform an anchor-based clustering over normalized post-projection visual embeddings. For each anchor $\mathbf{v}_a$, a cluster C_a is formed by grouping embeddings whose cosine similarity with the anchor exceeds a threshold τ:

$$C_a = \{\, j \mid \frac{\langle \mathbf{v}_a, \mathbf{v}_j \rangle}{\|\mathbf{v}_a\|_2 \|\mathbf{v}_j\|_2} \geq \tau \,\}.$$

Dominance of a Single Visual Cluster. With the similarity threshold set to $\tau = 0.9$, we rank clusters by their token counts for each image. Fig. 1(left) shows the top five clusters ranked by size. The largest cluster, denoted as C_0, consistently contains far more tokens than all remaining clusters combined. This indicates that, for each image, around 30% of post-projection visual embeddings encode highly repetitive information. We next investigate whether this redundancy is image-specific or reflects image-agnostic representational patterns shared across inputs.

B.2. Cross-image Cluster Similarity

To evaluate the consistency of visual clusters across different images, we compute the maximum cross-image similarity between cluster centroids. For each input image, we perform the same anchor-based clustering, and normalize the resulting cluster centroids. Given two sets of normalized cluster centroids $\mathbf{C}^{(1)}$ and $\mathbf{C}^{(2)}$, their cross-image similarity is defined as:

$$S_{ij} = \mathbf{C}^{(1)}_i \cdot \mathbf{C}^{(2)}_j, \quad \text{and} \quad s_i = \max_j S_{ij},$$

Table a. **Effect of sink pruning and sublayer skipping.** Performance remains stable under pruning or skipping operations, indicating weak reliance on visual sink tokens. $\notin \mathcal{I}_{S_{\text{LLM}}}$–MLP2 skips MLP-2 only for non-sink visual tokens.

Method	General				OCR			CV Centric			Hallucination		**Avg.**
	GQA	MME^{P}	MMStar	MMBench	VQA_{text}	OCRbench	VQA_{Doc}	$\text{RefCOCO+}_{\text{val}}$	$\text{RefCOCO}_{\text{testA}}$	$\text{RefCOCO}_{\text{testB}}$	POPE	Hallusion	Avg.
LLaVA-v1.5 7B	61.9	1507	33.5	64.1	58.2	31.2	21.5	50.0	64.5	47.4	85.9	36.2	52.7
Sinks Pruning													
−ViT Sinks	61.9	1502	33.2	64.5	58.2	31.5	21.9	50.4	65.1	47.6	85.9	36.3	52.8
−LLM Sinks	62.0	1488	33.5	63.8	58.2	31.3	21.5	50.1	64.2	47.0	85.9	36.6	52.7
−All Sinks	62.0	1499	33.1	64.3	58.2	31.4	21.9	50.5	64.5	47.4	85.9	36.6	52.8
Sublayer Skipping													
$\mathcal{I}_{S_{\text{LLM}}}$−MLP2	62.0	1501	32.9	64.7	58.2	31.7	21.6	50.7	64.9	47.8	85.7	36.2	52.8
$\mathcal{I}_{S_{\text{LLM}}}$−MLP1,ATT2	61.9	1517	33.5	64.3	58.2	31.6	21.5	50	64.3	47.4	85.9	36.4	52.8
$\mathcal{I}_{S_{\text{LLM}}}$−MLP1/2,ATT2	62.0	1513	33.4	64.7	58.2	31.5	21.5	50.5	64.6	47.8	85.8	36.6	52.8
$\mathcal{I}_{S_{\text{LLM}}}$−MLP1/2,ATT2	62.0	1513	33.4	64.7	58.2	31.5	21.5	50.5	64.6	47.8	85.8	36.6	52.9
$\notin \mathcal{I}_{S_{\text{LLM}}}$−MLP2	61.9	1525	33.6	64.2	58.2	31.6	21.5	50.4	64.5	48.1	85.7	35.8	52.8

Table b. **Effect of pruning dead-token clusters.** Pruning dead-token clusters yields a small boost in accuracy, whereas removing the same number of remaining visual tokens causes substantial drops across all benchmarks. This validates that dead tokens are semantically void and redundant, while non-dead tokens carry meaningful information.

Method	General				OCR			CV Centric			Hallucination		**Avg.**
	GQA	MME^{P}	MMStar	MMBench	VQA_{text}	OCRbench	VQA_{Doc}	$\text{RefCOCO+}_{\text{val}}$	$\text{RefCOCO}_{\text{testA}}$	$\text{RefCOCO}_{\text{testB}}$	POPE	Hallusion	Avg.
LLaVA-v1.5 7B	61.9	1507	33.5	64.1	58.2	31.2	21.5	50.0	64.5	47.4	85.9	36.2	52.7
−Dead Tokens	61.8	1495	34	64.1	58.2	31.6	21.6	53.4	68.4	51.4	86	36.3	53.7
−Alive Tokens	60.5	1475	32.2	62.5	55.4	29	18.6	44.4	58.3	42.9	84.2	36.2	50.1

where s_i measures the highest alignment of cluster i from one image to any cluster in another. Averaging s_i over multiple image pairs provides a compact measure of cluster stability and semantic coherence across visual instances.

Persistence of Clusters Across Images. As shown in Fig. 1(right) for the top six most similar clusters, our cross-image analysis reveals that *many visual tokens not only encode identical information within their own image but also occupy nearly the same regions in the representation space across different images*. For example, the largest cluster C_0 of each image remains remarkably stable, with an average cross-image similarity exceeding $\mathbf{S_C} > 0.98$ and variance below 10^{-4}. Several smaller clusters also exhibit strong cross-image consistency, suggesting shared latent semantics or structural priors. These findings highlight the need for further semantic validation of such persistent tokens, as their invariance may stem from modality-specific artifacts rather than meaningful visual cues.

B.3. Impacts of τ

We also examine the effect of the clustering threshold τ. As shown in Fig. a and Fig. b, the overall cluster distribution remains consistent across τ = 0.95, 0.9, and 0.8. This indicates that our clustering results are robust and largely insensitive to the choice of τ.

C. Clustering Analysis on Other Models

Since the projection layer in MLLMs performs only a linear transformation on CLIP ViT features, we hypothesize that

Table c. **Effect of layer decoupling and shallow-layer skipping on multimodal reasoning.** vMHA is crucial for spatially grounded reasoning in CV-centric tasks, while vFFN contributes to knowledge refinement. In contrast, shallow-layer skipping shows minimal degradation, confirming that early visual processing is largely redundant. † denotes training-based method.

Method	General				OCR			CV Centric			Hallucination		**Avg.**
	GQA	MME^P	MMStar	MMBench	VQA_{text}	OCRbench	VQA_{Doc}	$RefCOCO+_{val}$	$RefCOCO_{testA}$	$RefCOCO_{testB}$	POPE	Hallusion	Avg.
LLaVA-v1.5 7B	61.9	1507	33.5	64.1	58.2	31.2	21.5	50.0	64.5	47.4	85.9	36.2	52.7
Layer Decoupling													
-vMHA†	61.9	1472	32.1	63.5	56.9	30.3	19	45.1	62.3	44.5	86.6	36.7	51.4
-vFFN†	62.2	1507	33.7	63.5	56.3	31.2	18.7	33.5	46	32.1	86.6	37.3	48.3
Shallow Layers Skipping													
L_{in}=4†	62.4	1444	34.0	66.0	55.9	32.0	21.8	51.2	67.9	50.0	85.6	36.8	53.2
L_{in}=5†	62.4	1442	33.3	65.8	56.5	31.9	21.6	51.9	68.0	50.6	85.8	36.4	53.3
L_{in}=6†	62.2	1413	33.1	66.2	57.3	31.3	21.2	51.4	67.3	49.3	85.4	36.8	52.9
L_{in}=8†	61.5	1440	32.7	65.4	56.1	31.0	20.5	48.0	64.0	46.1	85.4	36.7	51.9
L_{in}=10†	61.9	1391	33.3	63.1	57.6	30.5	20.0	46.7	63.8	45.5	85.9	34.2	51.3

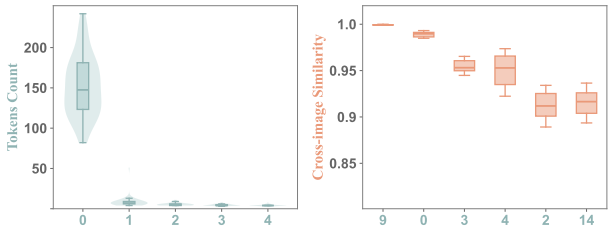

Figure a. **Clusters Distribution when** $\tau = 0.95$. (Left) Number of tokens in the top-5 clusters. (Right) Cluster similarity cross images. The cluster index is ranked by the number of tokens.

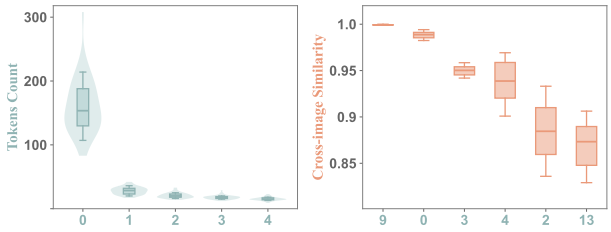

Figure b. **Clusters Distribution when** $\tau = 0.8$. (Left) Number of tokens in the top-5 clusters. (Right) Cluster similarity cross images. The cluster index is ranked by the number of tokens.

the redundant clusters originate primarily from the CLIP visual encoder itself. As shown in Fig. c, LLaVA-v1.5-13B exhibits nearly identical clustering behavior to the 7B variant (Fig. 1), supporting this assumption.

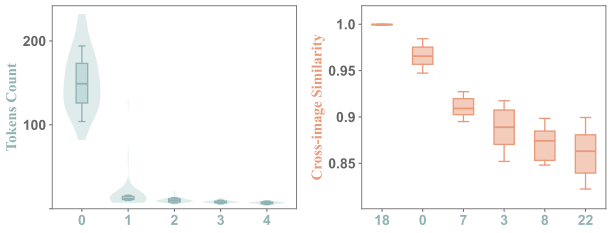

Figure c. **Clusters Distribution of LLaVA-v1.5 13B**. (Left) Number of tokens in the top-5 clusters. (Right) Cluster similarity cross images. The cluster index is ranked by the number of tokens.

C.1. Other Vision Encoders

We evaluate Qwen2.5-VL, InternVL3, and Qwen3-VL, and find that none display the same repetitive clustering pattern. Their visual tokens show notably higher diversity, suggesting that more recent encoders embed richer and less redundant visual information.

D. Effects of LLM Visual Sinks Manipulation

D.1. LLM Sinks Pruning

As shown in Fig. d(left), pruning visual sink tokens at the embedding stage leads to no measurable drop in performance. To understand how the attention previously focused on these tokens is redistributed, we visualize the average attention flow after pruning. The results show that attention originally directed toward visual sinks is reallocated to textual sinks within the system prompt, suggesting that these

tokens act primarily as structural placeholders for attention normalization rather than meaningful information carriers.

D.2. Skipping MLP 2

We further test the effect of skipping MLP-2 for sink tokens, the layer identified as the key stage for sink formation. As illustrated in Fig. d(right), bypassing MLP-2 substantially reduces the similarity between sink tokens and the ⟨bos⟩ representation, confirming that MLP-2 drives representational collapse toward the ⟨bos⟩ direction. Moreover, the inset plot reveals that skipping MLP-2 also markedly decreases the total attention allocated to visual sinks, suggesting that weakened sink alignment diminishes their ability to attract residual attention across layers.

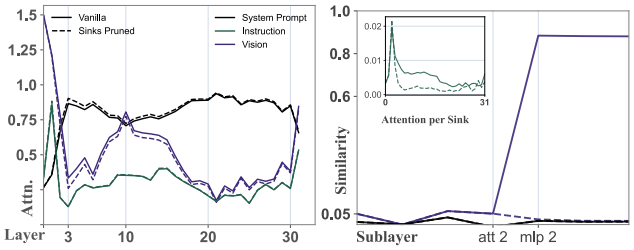

Figure d. **Effect of LLM sink manipulation.** (Left) Attention redistribution after pruning visual sinks at the input: attention shifts from visual sinks to textual sinks in the system prompt. (Right) Skipping MLP-2 weakens alignment between visual sinks and BOS, confirming its role in sink formation, while overall attention patterns remain stable.

E. Multi-Trajectory Diagnostic Benchmark

To better illustrate the construction of our diagnostic benchmark introduced in Sec. 7.1, we provide detailed examples in Fig. e. Each image is designed such that the target object or character occupies exactly one visual patch after resizing. For every image, we formulate three types of questions probing different semantic axes: (1) Object recognition, verifying whether the model can correctly identify the target entity; (2) Color identification, assessing grounding of visual color semantics; and (3) Counting, testing numerical reasoning within the patch.

The benchmark comprises three subgroups:

1. **Object**, object images with color and count variations;
2. **OCR**, isolated characters placed on white padding, evaluating fine-grained recognition and color binding;
3. **OCR with background**, characters placed on colored backgrounds, designed to measure contextual color bias and test whether the model grounds color to the object itself or to the dominant surrounding region.

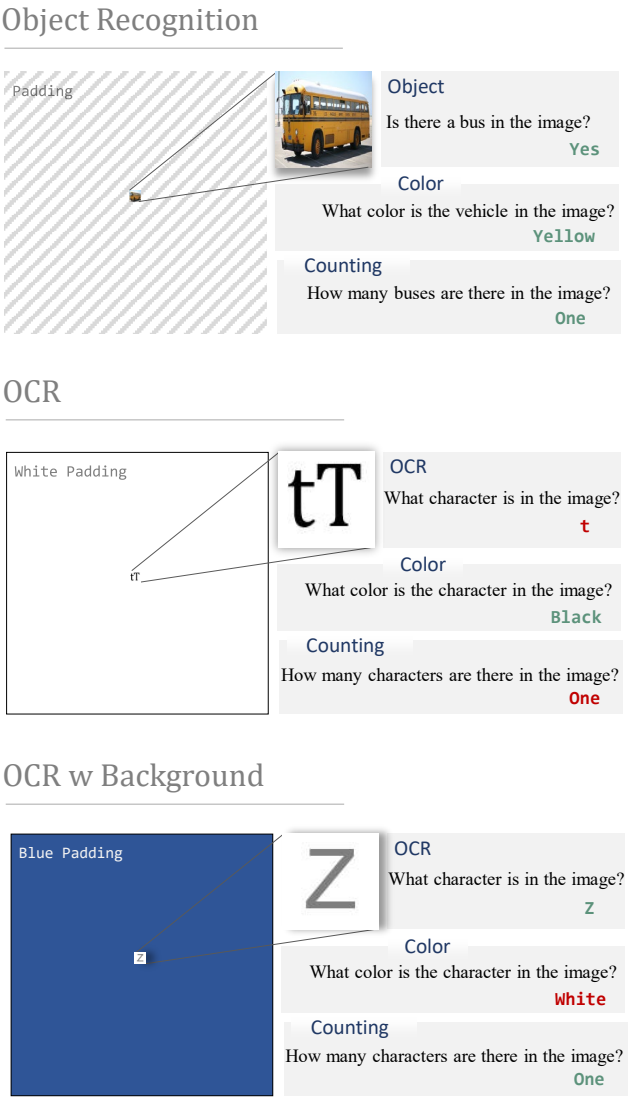

Figure e. **Examples from the multi-trajectory benchmark.** (Top) **Object recognition:** a bus image with associated object, color, and counting questions. (Middle) **OCR:** a single character on white padding for text–color grounding. (Bottom) **OCR with background:** same setup with a colored background, testing contextual bias in color reasoning.

F. Qualitative Analysis on Contextual Color Bias

While our quantitative benchmark (see Sec. 7.1) reveals that MLLMs often confuse object and background colors, the underlying cause of this phenomenon remains unclear. In this section, we provide qualitative evidence demonstrating that such errors stem from *contextual color bias*—a tendency for models to associate color semantics with dominant or surrounding visual regions rather than the object itself. Through controlled visual manipulations and *Em-*

bedLens layer-wise tracing, we show that MLLMs infer color primarily from context statistics instead of grounded object cues.

F.1. Bias Toward Dominant Patch Color

As discussed in the "OCR w Background" examples of Fig. e, MLLMs frequently predict the dominant patch color rather than the actual object color when the two differ. This confirms that color reasoning within a patch is often driven by surface-level color statistics rather than grounded object understanding.

F.2. Bias Toward Surrounding Color Context

In the general OCR subset, most color errors correspond to the color of the surrounding padding instead of the target object. To verify this, we visualize the layer-wise decoding of *EmbedLens* for the case shown in Fig. f. Initially, the character "9" contains no green-related semantics at any layer, yet the model correctly answers "green." When we recolor the digit itself (to blue), the model still predicts "green," showing that it does not ground color to the digit. However, once we modify the color of the surrounding background, the prediction immediately follows the new context, confirming that *color reasoning is largely inferred from nearby visual context rather than intrinsic object appearance*.

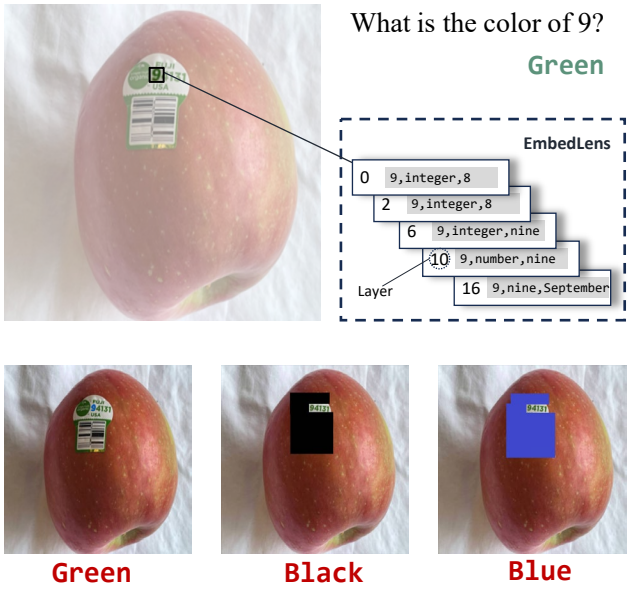

Figure f. **Bias Toward Surrounding Color Context**

G. Semantics Grounding for Vision Late Entry MLLMs

To examine how delayed visual input affects semantic grounding, we analyze *vision late-entry* MLLMs—models where visual tokens are introduced only at intermediate layers of the LLM backbone. One might expect that delaying visual injection allows the projector to take on stronger alignment responsibilities, yielding more semantically grounded visual tokens upon entry. However, our findings show that this is not the case.

As shown in Fig. g, the proportion of semantically grounded visual tokens at the entry point (left) remains nearly identical across entry layers (0–10), indicating that the projector itself does not inherently enhance semantic alignment. Instead, we observe that models with later visual entry continue to refine visual semantics over a broader range of layers, whereas dense models reach a stable alignment stage much earlier. This suggests that late-entry architectures shift the burden of cross-modal alignment deeper into the LLM, distributing semantic fusion across subsequent layers rather than concentrating it near the input.

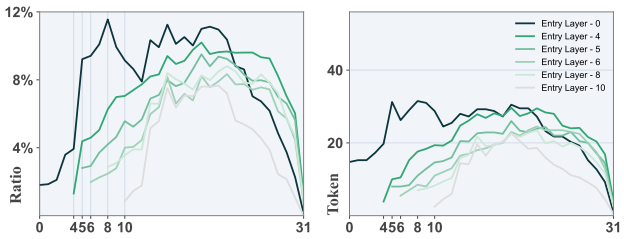

Figure g. **Semantic grounding in vision late-entry models.**