A TALE OF TWO SIGNALS: OPEN-VOCABULARY DENSE PERCEPTION WITH FROZEN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) encode spatial semantics in decoder hidden states and provide image-conditioned responses about category presence. We investigate how these signals complement each other for spatial recognition without task-specific fine-tuning. Our analyses show that same-layer visual–text pairing is not necessarily optimal, and selected visual-layer aggregation can improve recognition beyond the strongest single-layer counterpart in our diagnostic. We further distinguish the contextual role of jointly encoded categories from their role as prediction alternatives: textual context affects discrimination, while category competition can alter recognition even with fixed representations. Building on these observations, we introduce TALE (Semantic Aggregation and Selective Verification), which aggregates visual states, jointly constructs contextual category representations, and independently verifies leading candidates in ambiguous regions. Valid negative responses guide similarity refinement without reconstructing the classification representations or restricting the full prediction vocabulary. Experiments on regional classification, open-vocabulary detection, and segmentation demonstrate the utility of TALE, while controlled ablations characterize its recognition gains and inference costs. The anonymous code repository is available at: `https://anonymous.4open.science/r/TALE-6848/`.

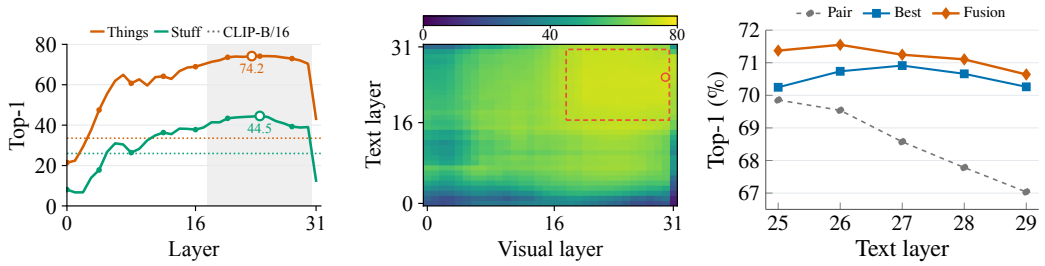
1 INTRODUCTION

Open-vocabulary dense perception extends visual recognition beyond predefined categories by associating textual concepts with image regions or pixels (Gu et al., 2022; Li et al., 2022). It requires models to preserve spatial detail while reliably distinguishing diverse semantic concepts. Similar local appearances and semantic relationships between categories can produce strong responses for incorrect categories, including those absent from the image (Chen et al., 2025). Image-level category information can constrain dense predictions: a category absent from an image should not be assigned to any spatial location (Pathak et al., 2015). Understanding how local visual–text evidence relates to image-level category presence therefore offers a route to more reliable dense perception.

Existing approaches commonly build on vision–language models (VLMs) such as CLIP (Radford et al., 2021) to match local visual features with text embeddings, improving local representations or visual–language alignment through further training (Wu et al., 2024; Wang et al., 2025; Li et al., 2025) or attention refinement (Wang et al., 2024; Lan et al., 2024b). Recent studies have also explored semantic guidance for category selection in dense perception including filtering redundant categories (Chen et al., 2025) and using multimodal large language model (MLLM) responses to assess category presence (Zhang et al., 2026). MLLMs (Liu et al., 2023) offer a natural possibility for connecting these two forms of semantics: their internal representations retain visual semantic information (Tao et al., 2024; Liu et al., 2025), while image-conditioned responses can assess category presence. This motivates us to ask: *can the internal representations and presence inference capabilities of a frozen MLLM complement each other to improve spatial semantic discrimination?*

To investigate this question, we first examine how visual states from different decoder layers interact with contextual category representations. Our analysis reveals cross-layer visual–text compatibility and shows that selected visual states provide complementary information through aggregation. Although this aggregation improves visual discrimination, visual–text associations acquired during pretraining may still favor incorrect categories with similar appearances or related semantics, weak-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Layer-wise recognition (b) Cross-layer compatibility (c) Layer aggregation

Figure 1: Visual semantics across decoder layers. (a) Top-1 accuracy of visual–text pairs, and the gray interval indicates the effective layers identified in (b). (b) Thing-mask Top-1 accuracy across visual–text layer pairs; the red circle marks the best pair (visual L30, text L25), and the dashed box highlights the high-accuracy region. (c) Thing-mask Top-1 accuracy with the jointly encoded 133-class vocabulary: the same-layer pair, the best visual single layer, and the best tested two-branch fusion for each fixed textual layer.

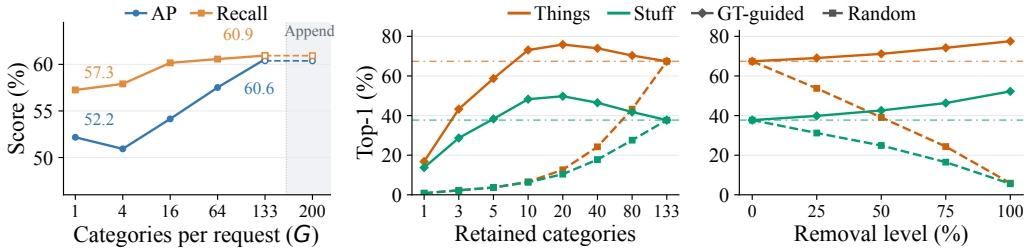
ening alignment between regions and their correct categories. This motivates examining whether these candidates are supported by image-level presence evidence: presence inference explicitly assesses whether a candidate occurs in the image, providing additional evidence for regional discrimination.

Based on these findings, we introduce **TALE** (seman**T**ic **A**ggregation and se**L**ective v**E**rification), which separates contextual matching from image-conditioned category refinement. TALE aggregates selected visual hidden states and matches them with category representations jointly constructed from the image and vocabulary. For regions with closely ranked category responses, TALE nominates the leading categories for independent image-conditioned presence assessment. For candidates receiving valid negative responses, the model's response preference determines a soft penalty on the corresponding regional similarities. The classification representations remain fixed, and all vocabulary categories remain eligible for prediction. Experiments demonstrate improvements across region classification, open-vocabulary object detection, and semantic segmentation. Our contributions are fourfold:

- We uncover cross-layer semantic complementarity in frozen MLLMs: visual states from the same decoder layer are not necessarily the best counterparts to a fixed textual representation, and selected cross-layer aggregation can further improve recognition beyond the strongest single-layer counterpart in our diagnostic.
- We distinguish the contextual and competitive roles of category vocabulary in frozen MLLM recognition. Joint encoding can improve visual–text discrimination depending on context composition, whereas changing prediction alternatives affects recognition even with fixed representations. These controls motivate separating representation construction from the refinement of category competition.
- We introduce TALE, which combines aggregated visual states and jointly encoded category representations with independent presence evidence to refine regional predictions while keeping the recognition representations fixed.
- Experiments across regional recognition, open-vocabulary detection, and segmentation demonstrate the effectiveness of TALE, while controlled ablations quantify the contributions of visual aggregation and presence-guided refinement and their inference costs.

2 Motivation

We investigate how decoder depth and textual context affect spatial visual–text similarity, and how classification choices affect recognition.



(a) Input size (b) Retained categories (c) Selective deletion

Figure 2: Textual context and category selection (%). (a) Image-level AP/Recall versus categories per request; the hollow extension is an append control. (b,c) Things/Stuff Top-1 under retained-category and selective deletion; horizontal lines show full-vocabulary accuracy.

2.1 Spatial Semantic Matching across Decoder Layers

Cross-Layer Compatibility. Prior approaches have probed information in MLLM language layers through object recognition and segmentation (Tao et al., 2024; Liu et al., 2025; Wu et al., 2026), and examined how visual tokens are processed across layers (Fan et al., 2026). Given an image I and vocabulary $\mathcal{C}$, we investigate how visual states interact with contextual category representations by independently varying the decoder layers used to extract visual and textual representations:

$$s_c^{(l_v,l_t)}(i) = \frac{\langle V_i^{(l_v)}, T_c^{(l_t)} \rangle}{\|V_i^{(l_v)}\|_2 \, \|T_c^{(l_t)}\|_2}. \tag{1}$$

Here, $V_i^{(l_v)}$ denotes the visual representation at position i from layer l_v, and $T_c^{(l_t)}$ denotes the textual representation of category c from layer l_t, conditioned on the image and input vocabulary. As shown in Fig. 1(a), MLLM hidden states outperform CLIP-B/16 across multiple decoder depths. The gray interval is the overlap of the visual and textual layer ranges forming the robust high-accuracy block in Fig. 1(b), showing that high-accuracy visual–text pairings span different decoder layers.

Layer Aggregation and Complementarity. Motivated by cross-layer compatibility, we investigate whether visual states from different depths provide complementary semantic information. For a fixed textual layer l_t, we rank visual layers by single-layer recognition accuracy and denote the highest-ranked layer by l_{best}. Selecting the Top-N layers with index set $\mathcal{L}_N$, we aggregate them as:

$$F = \mathcal{A}(V^{(l_{\text{best}})}, \frac{1}{N-1} \sum\nolimits_{l \in \mathcal{L}_N \setminus \{l_{\text{best}}\}} V^{(l)}). \tag{2}$$

The two inputs form the main ($V^{(l_{\text{best}})}$) and auxiliary branches, respectively. The operator $\mathcal{A}$ processes and normalizes each branch separately before weighted fusion. Here, N counts all selected visual layers, including the main layer. Fig. 1(c) compares same-layer pairing, the best single visual layer, and the best evaluated fusion for thing masks at textual layers T25–T29. Fusion outperforms the strongest single layer at all five textual layers, supporting cross-layer visual complementarity.

2.2 Textual Context in Visual–Text Recognition

Effect of Textual Input Group Size. Jointly encoded categories provide textual context for one another through causal attention. To examine its effect on recognition, we vary the number of categories per request, $G \in \{1, 4, 16, 64, 133\}$, while keeping the final prediction vocabulary fixed at 133 categories. Compared with individual encoding ($G = 1$), full-vocabulary encoding ($G = 133$) improves image-level AP from 52.2% to 60.6% and Recall from 57.3% to 60.9%, achieving the highest values among the evaluated group sizes (Fig. 2a). These results

Table 1: Recognition under different textual contexts.

Textual input	AP	Recall	Things	Stuff
Single	52.0	57.1	49.1	31.9
Full	60.4	60.9	60.0	36.3
Nearby (16)	**62.9**	62.4	**66.8**	**40.4**
Random (16)	61.2	**63.5**	61.3	36.1
Distant (16)	59.8	63.0	53.9	31.9

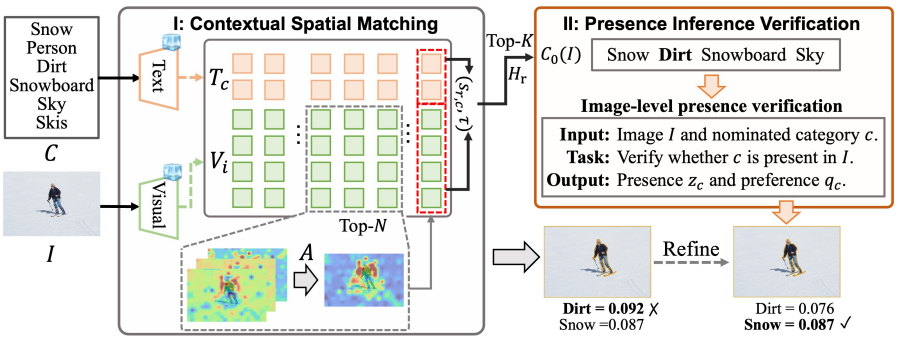


Figure 3: Overview of TALE. Full-vocabulary spatial similarities nominate leading categories for ambiguous regions. Independent image-conditioned presence assessment then refines the corresponding region–category similarities without recomputing the representations.

support **jointly constructing contextual category representations for recognition**. Appending categories after the original vocabulary leaves performance nearly unchanged, consistent with preserving the original textual tokens' visible context.

Effect of Textual Context Composition. With visual representations and the prediction vocabulary fixed, Nearby contexts achieve the highest AP and Things/Stuff Top-1, whereas Random contexts yield the highest Recall (Table 1). These results show that the benefits of joint encoding depend on context composition and the recognition metric.

These findings suggest that jointly constructing contextual category representations can improve recognition, while highlighting the importance of context size and composition. Additional context-composition and vocabulary-order analyses are provided in Appendix A.2.

2.3 Category Selection with Fixed Representations

Category Information beyond the Top Prediction. For classification within ground-truth regions, correct categories often rank among the leading alternatives without being ranked first. On Things and Stuff masks, Top-5 accuracy reaches 88.3% and 68.8%, compared with Top-1 accuracy of 67.4% and 37.7%, respectively. This suggests that correct categories can remain among the leading alternatives even when they do not receive the highest visual–text similarity.

Selecting Categories for Prediction. To examine how prediction alternatives affect recognition, we keep the full-vocabulary visual and textual representations fixed and vary only the competing categories. Prioritizing annotated categories achieves higher accuracy than using the full vocabulary, whereas random selection can remove correct alternatives (Fig. 2(b)). Removing unannotated categories improves recognition, while randomly removing the same number reduces accuracy (Fig. 2(c)). These annotation-guided controls suggest that recognition errors may arise from competing category responses, even when the representations remain unchanged. They reveal an opportunity to refine category competition, rather than establish the effectiveness of model-generated presence evidence. Further ranking and category-competition diagnostics appear in Appendix A.3.

3 Methods

Motivated by the observations, as shown in Fig. 3, we explore two complementary ways to improve recognition: constructing discriminative hidden-state representations through weighted visual-layer aggregation and joint full-vocabulary encoding, and refining regional category competition through image-conditioned presence inference.

3.1 Contextual Spatial Matching

Multi-Layer Visual Semantic Aggregation. Motivated by the aggregation gains in Fig. 1(c), following Eq. 2, TALE uses a reference visual layer as the main branch and the mean of selected addi-

tional visual states as the auxiliary branch. The two branches are spatially aligned and normalized separately, then combined through weighted fusion to obtain F.

Spatial Visual–Text Similarity. With the textual layer fixed, we abbreviate $T_c^{(l_t)}$ as T_c. TALE jointly encodes the full vocabulary to obtain all category representations from a single shared input sequence. For each $c \in \mathcal{C}$, we pool and normalize the hidden states of its corresponding textual tokens, conditioned on the image and vocabulary. These representations are constructed once per image and remain fixed throughout recognition. Let F_i denote the aggregated visual descriptor at spatial position i. We compute its cosine similarity $s_c(i)$ with each category representation:

$$s_c(i) = \frac{\langle F_i, T_c \rangle}{\|F_i\|_2 \, \|T_c\|_2}, \qquad c \in \mathcal{C}. \tag{3}$$

Ambiguity-Guided Candidate Nomination. For each region r in image I, we pool the spatial similarities over its positions to obtain $s_{r,c} = \mathrm{Pool}_{i \in r} \, s_c(i)$. We measure regional ambiguity by the gap between the two highest category similarities, Δ_r. A region nominates categories only when $\Delta_r \leq \tau$, where τ is the similarity-gap threshold. This small gap indicates that the two leading categories receive similar visual–text support. For each triggered region, let H_r denote the set of its Top-K nominated categories. Their union forms the image-level verification set $\mathcal{C}_0(I)$:

$$H_r = \underset{c \in \mathcal{C}}{\arg\,\text{top-K}} \, s_{r,c}, \qquad \mathcal{C}_0(I) = \bigcup\nolimits_{r:\, \Delta_r \leq \tau} H_r. \tag{4}$$

Other regions nominate no categories. Here, K specifies the number of candidates per triggered region, while $\mathcal{C}_0(I)$ contains the distinct candidates across the image. Each category is verified once per image, while its associations with the nominating regions are retained for refinement.

3.2 Verification through Presence Inference

Motivated by the category-competition analysis in Sec. 2.3, we complement hidden-state visual–text similarities with independent image-level presence evidence.

Independent Image-Conditioned Presence Assessment. For each $c \in \mathcal{C}_0(I)$, we independently query the same frozen MLLM about whether c is present in the full image. Each request contains a single category and no previous dialogue. At the first answer-token position, we sum the generation probabilities over accepted `Yes` and `No` token variants, obtaining $p_c^{\texttt{Yes}}$ and $p_c^{\texttt{No}}$, respectively. Their relative `Yes` preference is:

$$q_c = \frac{p_c^{\texttt{Yes}}}{p_c^{\texttt{Yes}} + p_c^{\texttt{No}}}. \tag{5}$$

This quantity measures the model's response preference, rather than a calibrated probability of category presence. Valid `Yes`/`No` answers are mapped to $z_c \in \{\texttt{Present}, \texttt{Absent}\}$; ambiguous, invalid, or incomplete answers are unresolved. Each category is queried once per image, and its judgment and preference are shared by the regions that nominated it.

Presence-Guided Regional Refinement. We use negative presence evidence to suppress unsupported regional candidates. For a nominated region–category pair with a valid negative response, we refine its similarity as:

$$\widetilde{s}_{r,c} = s_{r,c} + \lambda \log q_c, \qquad c \in H_r, \quad z_c = \texttt{Absent}, \tag{6}$$

where $\lambda \geq 0$ controls the refinement strength. A lower `Yes` preference q_c yields a larger penalty. Positive and unresolved responses leave the original similarities unchanged. The refined similarities $\widetilde{s}_{r,c}$ guide subsequent prediction, with all categories in the full vocabulary remaining eligible. This refinement changes category competition without recomputing the visual or textual representations. Alternative refinement rules are evaluated in Appendix B.2.

4 Experiments

4.1 Zero-Shot Semantic Classification

We evaluate category recognition on COCO using ground-truth boxes, thing masks, and stuff masks. Table 2 reports Top-1 and Top-5 accuracy. With frozen LLaVA-1.5-Vicuna-7B, TALE achieves

Table 2: Semantic recognition on COCO. We report Top-1 and Top-5 accuracy (%).

Method	Model	Boxes		Thing Masks		Stuff Masks	
		Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
CLIP	ViT-B/16	29.2	51.6	33.6	56.1	25.9	51.0
CLIP	ViT-L/14	22.3	47.9	29.4	53.1	14.5	32.6
TALE	LLaVA-1.5-Vicuna-7B	67.3	88.9	68.9	88.7	40.2	72.4
	LLaVA-1.6-Vicuna-7B	71.0	90.1	72.2	89.6	42.7	71.8
	LLaVA-1.6-Vicuna-13B	**80.0**	**94.5**	**82.5**	**94.4**	**50.9**	**81.8**
	LLaVA-1.6-Mistral-7B	60.1	78.6	64.7	81.6	39.1	63.9

Table 3: Cross-dataset zero-shot classification (Top-1, %).

Configuration	VOC20	VOC21	PC59	PC60	ADE150	Stuff171	Object81
Single layer	95.4	92.9	60.7	55.2	37.3	40.9	69.5
TALE	**96.1**	**93.7**	**65.5**	**61.5**	**43.8**	**46.2**	**73.2**

67.3%, 68.9%, and 40.2% Top-1 accuracy, exceeding the CLIP ViT-B/16 baseline by +38.1, +35.3, and +14.3 points, respectively. Top-5 results show that correct categories frequently remain among the leading predictions. TALE also performs strongly across the evaluated LLaVA variants; LLaVA-1.6-Vicuna-13B obtains the highest Top-1 results at 80.0%, 82.5%, and 50.9%. Using ground-truth regions isolates category discrimination from proposal generation and boundary estimation, allowing us to evaluate the semantic utility of frozen MLLM representations directly. Additional model comparisons, transfer results, and fusion-weight controls appear in Appendix B.1.

Transfer to Other Datasets. We further evaluate zero-shot regional classification on seven dataset protocols using the layer configuration selected on COCO, without target-dataset layer tuning. As shown in Table 3, TALE consistently outperforms the single-layer baseline by $+0.6$–$+6.5$ points. The largest gains occur on ADE150, PC60, and Stuff171, reaching $+6.5$, $+6.2$, and $+5.2$ points, respectively. These results support the transferability across the evaluated protocols.

4.2 Applications to Open-Vocabulary Dense Perception

Open-Vocabulary Object Detection. We evaluate OV-COCO following (Zareian et al., 2021; Gu et al., 2022) and report novel AP_{50}^{N}. Table 4 groups methods by whether detector training is used; EVA denotes EVA-CLIP (Sun et al., 2023). Under identical SAM proposals and detection post-processing, TALE with frozen LLaVA-1.5-Vicuna-7B achieves 31.8 AP_{50}^{N}, exceeding our CLIP-B/16 baseline by 7.7 points. It performs best among the listed training-free methods. The strongest trained method reaches 46.2, leaving a gap to detector-adapted systems.

Table 4: Open-vocabulary object detection on OV-COCO.

Method	Model variant	AP_{50}^{N}
With training		
F-VLM [ICLR 2023]	CLIP-RN50	28.0
F-ViT [ICLR'24]	EVA-L/14	24.7
CLIPSelf [ICLR'24]	EVA-B/16	37.6
CLIPSelf [ICLR'24]	EVA-L/14	44.3
DeCLIP [CVPR'25]	EVA-L/14	46.2
Zip [ICLR 2024]	CLIP-RN50	40.3
Training-free		
Zip [ICLR 2024]	CLIP-RN50	23.5
Zip [ICLR 2024]	CLIP-RN50x64	27.1
CLIP	CLIP-B/16	24.1
TALE	LLaVA	**31.8**

Open-Vocabulary Semantic Segmentation. We evaluate OVSS on VOC21 (Everingham et al., 2010), PC60 (Mottaghi et al., 2014), ADE150 (Zhou et al., 2017), Object81 (Lin et al., 2014), and Stuff171 (Caesar et al., 2018). Table 5 compares TALE with CLIP-based methods (Lan et al., 2024a; Wang et al., 2024; Lan et al., 2024b; Stojnić et al., 2025; Yang et al., 2025; Chen et al., 2025) and FSeg-LLaVA (Zhang et al., 2026). Using frozen LLaVA-1.5-Vicuna-7B, TALE achieves 69.0, 38.2, 23.1, 42.3, and 26.8 mIoU, respectively—the highest results on all five benchmarks. Relative to FSeg-LLaVA with the same LLaVA variant, the gains are +1.0, +7.6, +6.2, +0.3, and +5.6, respectively. These results show that TALE performs consistently across different datasets.

Table 5: Comparison on open-vocabulary semantic segmentation benchmarks.

Method	Models	VOC21	PC60	ADE150	Object81	Stuff171
ClearCLIP [ECCV'24]	CLIP	51.8	32.6	16.7	33.0	23.9
SCLIP [ECCV'24]	CLIP	61.7	31.5	17.8	32.1	23.9
ProxyCLIP [ECCV'24]	CLIP	61.3	35.3	20.2	37.5	26.5
LPOSS [CVPR'25]	CLIP	62.9	34.1	22.2	35.1	26.5
ResCLIP [CVPR'25]	CLIP	61.1	33.5	18.0	35.0	24.7
FreeCP [ICCV'25]	CLIP	65.8	35.3	18.4	37.2	24.9
FSeg-LLaVA [CVPR'26]	1.5-Vicuna-7B	68.0	30.6	16.9	42.0	21.2
FSeg-LLaVA [CVPR'26]	1.6-Vicuna-7B	65.9	33.4	20.0	41.6	23.1
FSeg-LLaVA [CVPR'26]	1.6-Vicuna-13B	64.1	32.9	18.7	39.1	23.3
FSeg-LLaVA [CVPR'26]	1.6-Mistral-7B	65.1	32.4	17.9	40.0	23.2
TALE (Ours)	1.5-Vicuna-7B	**69.0**	**38.2**	**23.1**	**42.3**	**26.8**

Table 6: Effects of visual aggregation and presence refinement on Top-1/Top-5 accuracy (%).

Variants	Visual		Verification	Top-1		Top-5		Queries/img.
	Main	Aux.		Boxes	Things	Boxes	Things	
1	✓	–	–	62.9	65.8	86.8	87.7	0.0
2	–	✓	–	63.9	67.1	87.2	88.4	0.0
3	✓	✓	–	65.1	67.0	88.7	88.6	0.0
4	✓	–	τ	64.9	67.9	86.6	87.6	8.2
5	✓	✓	w/o τ	**67.4**	**69.2**	88.8	**88.8**	15.4
TALE	✓	✓	τ	67.3	68.9	**88.9**	88.7	8.2

4.3 Ablation Study

Component Ablation. Table 6 varies visual aggregation and verification while fixing the category representations. Without verification, combining the L30 main branch with the auxiliary mean of L20–L24 (variants 1–3) improves Boxes Top-1 and both Top-5 metrics over either branch, supporting complementary information across visual layers. With aggregation fixed, presence refinement (variant 3 vs. TALE) improves Boxes/Things Top-1 by 2.1/2.0 points, showing that image-level presence evidence complements the aggregated visual representation. Variant 4 uses the main branch for prediction while sharing the aggregation-based nominations and presence responses. Variant 5 removes selective triggering and verifies every region. Compared with variant 5, Ours requires 46.8% fewer queries with 0.2/0.2-point lower Top-1.

Aggregation across Textual Depths. We fix each textual representation from H18–H32 and augment its same-layer visual counterpart with the highest-ranked visual layer when it differs from the main layer. On 5,000 COCO images with the joint 133-category vocabulary, aggregation improves Top-1 at all 15 textual depths for Boxes and Things, and at 12 of 13 applicable depths for Stuff (Fig. 4). Refinement further improves every applicable fused configuration in this comparison. The gains across multiple fixed pairings support complementary visual and presence evidence. Full Top-1/Top-5 curves across ranking budgets appear in Appendix A.1.

Error-Type Analysis. We group triggered regional errors by whether the predicted category is unannotated (U) or annotated (A) in the image (Table 7). Both groups contain incorrect regional predictions; correction requires matching regional ground truth. All rates are computed within each triggered-error group. Under No-only refinement, U receives more negative responses and higher correction rates despite lower ground-truth Top-3 coverage. This supports com-

Table 7: Presence verification on triggered misclassifications.

Region	Negative response		GT in Top-3		Corrected	
	U	A	U	A	U	A
Boxes	74.2	24.3	45.3	54.0	23.2	6.9
Things	84.5	36.8	31.4	34.0	20.3	8.4
Stuff	83.0	50.9	42.1	42.8	13.1	6.4

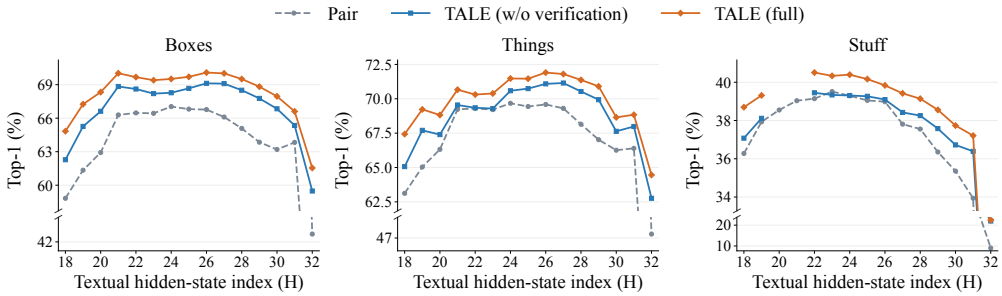


Figure 4: Regional Top-1 across textual depths at $B = 1$: same-layer Pair, TALE (w/o verification), and TALE (full). The highest-ranked auxiliary candidate is used. For Stuff at H20/H21, the top-ranked visual layer coincides with the same-layer counterpart, leaving no additional auxiliary layer.

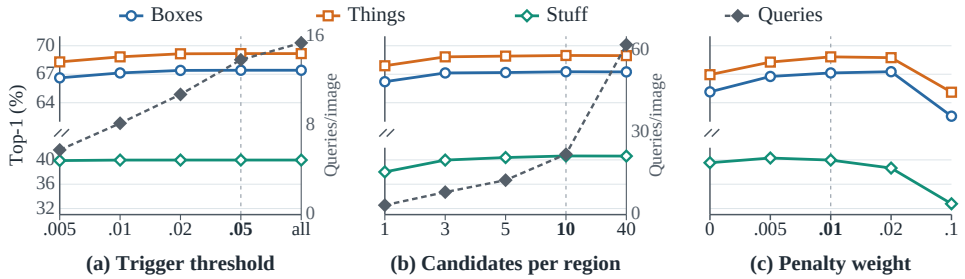


Figure 5: Regional Top-1 accuracy and mean category queries per image (right axes) as τ, K, and λ vary individually. Vertical dashed lines mark the highest equal-weight mean Top-1 across Boxes, Things, and Stuff; query counts are shown in (a,b).

plementary image-level evidence for suppressing unsupported competitors, although presence alone cannot establish which region contains a category. Annotations serve only for evaluation; unannotated does not necessarily mean absent. Further diagnostics appear in Appendix B.2.

Regional Refinement Sensitivity. We vary τ (triggering), K (regional candidate count), and λ (penalty strength) individually around $(0.01, 3, 0.01)$ (Fig. 5). Broader triggering yields diminishing gains. Increasing K from 3 to 10 improves Top-1 across region types but raises queries from 8.2 to 22.0 per image. Excessive penalties particularly harm Stuff: Top-1 falls from 40.3% at $\lambda = 0.005$ to 32.8% at 0.1. The reference setting therefore balances recognition and query cost rather than maximizing every metric.

Inference Overhead. *(1) Selective Verification.* Selective verification uses 8.2 rather than 15.4 queries/image, reducing query counts by 46.8%. At 346.7 ms/query on average, this favors querying ambiguous regions. *(2) Aggregation and Verification Overhead.* On one RTX 5090 over 100 matched images, aggregation adds 0.6 s/image; verification takes 3.0 of 3.9 s/image. We reuse responses across regions to avoid repeated queries (Appendix C.2). *(3) Context Performance–Cost Trade-off.* With visual representations fixed and verification disabled, Nearby gains 4.3/6.6 points on Things/Stuff but incurs 31.2-fold latency under sequential, batch-one execution without caching, with similar peak memory. Full avoids repeated encoding, favoring latency over Nearby's higher accuracy (Table 8; Appendix C.1).

Table 8: Context construction costs.

Context	Top-1 (%)		Latency (ms/image)	
	Things	Stuff	Encoding	Total
Full	64.0	30.1	**146.8**	**366.8**
Nearby (16)	**68.3**	**36.7**	11227.9	11451.1

4.4 VISUALIZATION

Figure 6 compares activation maps from individual decoder layers with their aggregation. The individual layers differ in spatial focus and show scattered responses outside the highlighted re-

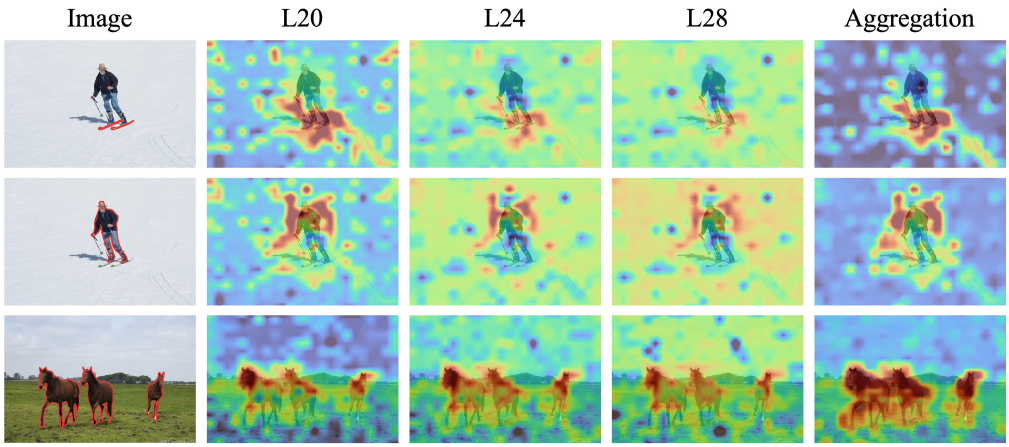


Figure 6: Activation maps from individual layers and their aggregation.

gions. Aggregation produces more coherent activation around the target regions while reducing some background responses. These examples qualitatively illustrate the benefit of combining visual information across depths.

5 Related Works

Open-vocabulary dense perception. VLMs enable text-defined detection (Gu et al., 2022; Minderer et al., 2022; Liu et al., 2024) and segmentation through trained local alignment (Wu et al., 2024; Wang et al., 2025; Li et al., 2025), training-free CLIP adaptation (Zhou et al., 2022; Wang et al., 2024; Lan et al., 2024a), spatial cues (Lan et al., 2024b; Zhang et al., 2025), or reference embeddings (Xuan et al., 2025). We instead study spatial visual–text discrimination in frozen MLLMs.

Visual representations in MLLMs. MLLMs support region understanding (You et al., 2024; Yuan et al., 2024) and segmentation (Lai et al., 2024; Rasheed et al., 2024; Li et al., 2026). Prior analyses reveal visual semantics in decoder states (Tao et al., 2024; Liu et al., 2025; Wu et al., 2026) and sensitivity to input order (Tan et al., 2024). Without fitting probes, we independently vary visual and textual depths to examine cross-layer compatibility and complementary information beyond the strongest single-layer readout.

Category selection and presence assessment. Category constraints (Pathak et al., 2015), vocabulary purification (Chen et al., 2025), and automatic vocabulary construction (Ülger et al., 2025) support segmentation. FSeg-LLaVA (Zhang et al., 2026) uses category queries and generated responses to construct layer activation maps. We distinguish contextual representation construction from prediction competition: TALE retains jointly constructed representations and separately uses presence responses to refine ambiguous regional similarities over the full vocabulary.

6 Conclusion

Frozen MLLM hidden states support cross-layer spatial visual–text discrimination, with performance shaped by shared textual context. TALE builds on this evidence by aggregating selected visual states and independently assessing leading alternatives in ambiguous regions. It refines their similarities while preserving contextual representations and allowing the full vocabulary to compete. Controlled regional experiments show recognition gains from this combination without task-specific training, at the cost of additional queries.

Limitations. Ambiguity-triggered verification may miss confidently wrong predictions, while independent category queries increase inference cost. Dense tasks also rely on external spatial refinement, so our results do not establish precise localization from MLLM hidden states alone.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for [Assist with translation]. We have not used generative AI tools for [Interpret results, Support qualitative and thematic data analysis, Implement methods, Design or provide feedback on research methodology or experiments, Propose or refine hypotheses], and [Clean and reformat dataset, Assist in the writing of proofs, Provide critical ingredients for proving mathematical claims, Formulate mathematical claims, Generate synthetic data sets, Help develop theoretical models or conceptual frameworks] are not applicable to this work. Additionally, we used generative AI tools for [Propose a title for a research paper, Format references, Create or modify scientific figures or images]. We have reviewed all AI-assisted work. Specifically, we checked translations of literature and text for accuracy, reviewed suggested keywords for consistency with the paper's scope, verified reference details against the cited sources, and checked the accuracy of the data presented in figures and tables. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly available datasets and pretrained models.

REPRODUCIBILITY STATEMENT

The main paper and appendix describe the method formulation, inference settings, evaluation protocols, and implementation details for reproducing our experiments. TALE uses frozen pretrained models without task-specific fine-tuning. Our experimental environment comprises four NVIDIA GeForce RTX 5090 GPUs (32 GB each), Python 3.10.20, PyTorch 2.7.0, CUDA 12.8, cuDNN 9.7.1, and Transformers 4.46.3. Inference latency is measured on a single GPU, as detailed in Appendix C.2.

REFERENCES

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1209–1218, 2018.

Qi Chen, Lingxiao Yang, Yun Chen, Nailong Zhao, Jianhuang Lai, Jie Shao, and Xiaohua Xie. Training-free class purification for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23124–23134, 2025.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

Yingqi Fan, Junlong Tong, Anhao Zhao, and Xiaoyu Shen. What do visual tokens really encode? uncovering sparsity and redundancy in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11987–11997, 2026.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022.

Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-VLM: Open-vocabulary object detection upon frozen vision and language models. In *International Conference on Learning Representations*, 2023.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9579–9589, 2024.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ClearCLIP: Decomposing CLIP representations for dense vision-language inference. In *European Conference on Computer Vision*, 2024a.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ProxyCLIP: Proxy attention improves CLIP for open-vocabulary segmentation. In *European Conference on Computer Vision*, 2024b.

Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022.

Yi Li, Hongze Shen, Lexiang Tang, Xin Li, Xinpeng Ding, Yinsong Liu, Deqiang Jiang, Xing Sun, and Xiaomeng Li. DenseMLLM: Standard multimodal LLMs for dense prediction. In *Forty-third International Conference on Machine Learning*, 2026.

Yunheng Li, Yuxuan Li, Quan-Sheng Zeng, Wenhai Wang, Qibin Hou, and Ming-Ming Cheng. Unbiased region-language alignment for open-vocabulary dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23795–23805, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Benlin Liu, Amita Kamath, Madeleine Grunde-McLaughlin, Winson Han, and Ranjay Krishna. Visual representations inside the language model. In *Conference on Language Modeling*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 2024.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. In *European Conference on Computer Vision*, 2022.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 891–898, 2014.

Deepak Pathak, Philipp Krähenbühl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1796–1804, 2015.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. GLaMM: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13009–13018, 2024.

Cheng Shi and Sibei Yang. The devil is in the object boundary: Towards annotation-free instance segmentation using foundation models. In *International Conference on Learning Representations*, 2024.

Vladan Stojnić, Yannis Kalantidis, Jiří Matas, and Giorgos Tolias. LPOSS: Label propagation over patches and pixels for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9794–9803, 2025.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved training techniques for CLIP at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Zhijie Tan, Xu Chu, Weiping Li, and Tong Mo. Order matters: Exploring order sensitivity in multimodal large language models. *arXiv preprint arXiv:2410.16983*, 2024.

Mingxu Tao, Quzhe Huang, Kun Xu, Liwei Chen, Yansong Feng, and Dongyan Zhao. Probing multimodal large language models for global and local semantic representations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pp. 13050–13056. ELRA and ICCL, 2024.

Osman Ülger, Maksymilian Kulicki, Yuki Asano, and Martin R. Oswald. Auto-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24266–24275, 2025.

Feng Wang, Jieru Mei, and Alan Yuille. SCLIP: Rethinking self-attention for dense vision-language inference. In *European Conference on Computer Vision*, 2024.

Junjie Wang, Bin Chen, Yulin Li, Bin Kang, Yichi Chen, and Zhuotao Tian. DeCLIP: Decoupled learning for open-vocabulary dense perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14824–14834, 2025.

Boyong Wu, Sanghwan Kim, and Zeynep Akata. From drop-off to recovery: A mechanistic analysis of segmentation in MLLMs. *arXiv preprint arXiv:2603.17228*, 2026.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. CLIPSelf: Vision transformer distills itself for open-vocabulary dense prediction. In *International Conference on Learning Representations*, 2024.

Xiwei Xuan, Ziquan Deng, and Kwan-Liu Ma. ReME: A data-centric framework for training-free open-vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20954–20965, 2025.

Yuhang Yang, Jinhong Deng, Wen Li, and Lixin Duan. ResCLIP: Residual attention for training-free dense vision-language inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29968–29978, 2025.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *International Conference on Learning Representations*, 2024.

Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28202–28211, 2024.

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14393–14402, 2021.

Bingfeng Zhang, Siyue Yu, Hui Li, Jiahua Lin, Wenwu Wang, and Jimin Xiao. The power of prior: Training-free open-vocabulary semantic segmentation with LLaVA. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6336–6345, 2026.

Dengke Zhang, Fagui Liu, and Quan Tang. CorrCLIP: Reconstructing patch correlations in CLIP for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24677–24687, 2025.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 633–641, 2017.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In *European Conference on Computer Vision*, 2022.

APPENDIX

A REPRESENTATION AND CATEGORY-COMPETITION DIAGNOSTICS

These diagnostics extend the Motivation experiments. Sample sets and representation settings are specified separately, so absolute accuracies should be compared within each experiment. Annotation-guided category controls use ground truth explicitly as diagnostic references; TALE uses annotations only for evaluation, not candidate nomination or presence verification. Unless stated otherwise, Top-1 and Top-5 average accuracy over categories. Percentages and differences are computed before rounding to one decimal place.

A.1 VISUAL-LAYER COMPATIBILITY AND AGGREGATION

Cross-Layer Compatibility. Figure 7 supplements the layer-pair diagnostics in the main text.

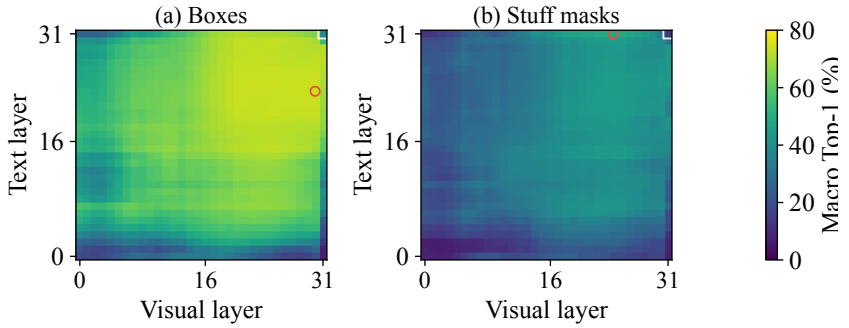
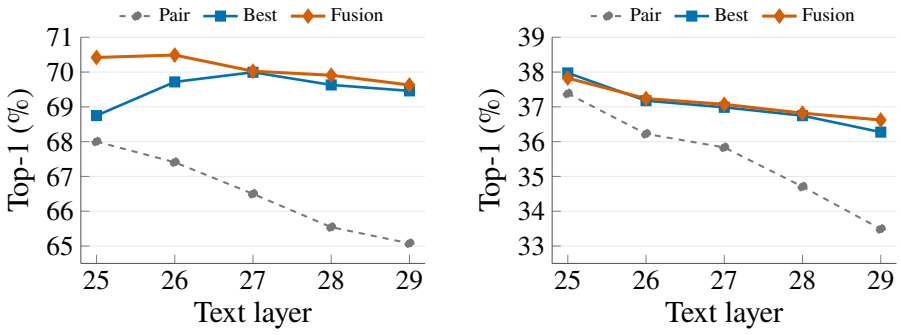


Figure 7: Additional split-vocabulary visual–text compatibility matrices omitted from Fig. 1. We report Top-1 accuracy over all decoder-layer pairs for (a) boxes with Things80 and (b) stuff masks with Stuff53. The white square marks the final–final pair; the red circle marks the strongest pair (V30–T23 for boxes and V24–T31 for stuff).

Pair, Best Single Layer, and Fusion. Figure 8 reports the corresponding three-step comparison for Boxes and Stuff.



(a) Boxes (b) Stuff

Figure 8: Boxes and Stuff Top-1 with the jointly encoded 133-class vocabulary across textual layers T25–T29. Curves compare Pair, Best, and Fusion as in Fig. 1(c). The fused configuration is selected from the tested Top-2–Top-6 groups and strong-layer subsets on the evaluation results.

Aggregation and Verification across Textual Depths. This scan uses frozen LLaVA-1.5-Vicuna-7B, 336×336 image inputs, the jointly encoded 133-category vocabulary, and GT boxes, thing

masks, and stuff masks. For textual hidden-state index H18–H32, the main visual layer is the same decoder block, indexed L(H-1). Visual-layer rankings are fixed separately for each textual depth and region type using 100 development images. Each branch is RMS-normalized separately, and the main state and mean auxiliary state are combined with equal weights.

The legends in Figs. 9 and 10 follow the Top-B selection used in Fig. 4: we select the highest-ranked B visual layers and remove the main layer if present, without replacement. Here, B denotes the selection budget before excluding the main layer. The auxiliary branch contains $B - 1$ layers when the main layer is among the selected layers, and B otherwise. $B = 0$ uses only the main layer. The two empty auxiliary sets, Stuff H20/$B = 1$ and H21/$B = 1$, are left blank. All other configurations are included; identical layer sets at adjacent budgets give repeated points.

Verification uses each configuration's ambiguity-triggered regional Top-3 candidates, with $\tau = 0.01$ and $\lambda = 0.01$. Valid Yes and No responses contribute $\lambda \log q_c$ to nominated similarities; unresolved responses leave them unchanged. All 133 categories remain eligible.

Across the 268 configurations, 223 include auxiliary layers. After collapsing the 14 equivalent configurations, aggregation improves Top-1 over the corresponding Pair in 70/70 Boxes, 66/69 Things, and 63/70 Stuff configurations. Increasing the ranking budget does not yield monotonic gains. At $B = 1$, mean Top-1 gains over Pair are 3.7, 2.4, and 1.6 percentage points for Boxes, Things, and Stuff; excluding the 100 development images gives 3.7, 2.4, and 1.6 points, with the same improvement counts. The figures report all 5,000 images, including development images; neither layers nor budgets are reselected on these results.

Summary. Selected cross-layer aggregation improves over Pair in most evaluated configurations. The Top-B curves are not monotonic: adding more auxiliary layers does not guarantee higher recognition accuracy.

A.2 Textual Context and Vocabulary Order

Context Composition. Table 1 reports the context-composition controls. AP/Recall@10 are image-level metrics; Things/Stuff report regional Top-1. For each target, we select 15 other categories using nearest, random, or farthest similarity in frozen text-only representations, then place the target last. Only its representation is collected; all 133 target representations participate in prediction with fixed visual features. Each configuration is evaluated on 500 images. Token lengths and absolute token positions are not matched, and no repeated-seed uncertainty estimates are available. The full-vocabulary reference here has AP 60.4%, whereas the earlier group-size experiment has AP 60.6%. We retain the original references and compare configurations within each experiment. The target-specific settings construct 133 input contexts per image, whereas Full collects all category representations from one context. We additionally compare Full and Nearby on a separate, matched 100-image evaluation, measuring recognition and runtime within the same implementation (Table 16). These results are reported separately from the 500-image context-composition diagnostic.

Vocabulary Order. Keeping images and vocabulary fixed, changing only textual order yields Things Top-1 between 57.6% and 65.3% and Stuff Top-1 between 33.0% and 38.3% across ten permutations (Fig. 11a). Moving a target from last to first increases its mean similarity and margin over the strongest alternative (Fig. 11b). Figure 12 provides a visual example.

Summary. Context composition and textual order both affect visual–text discrimination; joint encoding can outperform isolated encoding, but Full is not the most accurate context in these comparisons. The results establish contextual sensitivity without isolating semantic composition from token length and position.

A.3 Vocabulary Scale and Category Competition

Regional Ranking Coverage. Table 9 provides the complete regional Top-1/5/10 results for the ranking diagnostic discussed in Sec. 2.3.

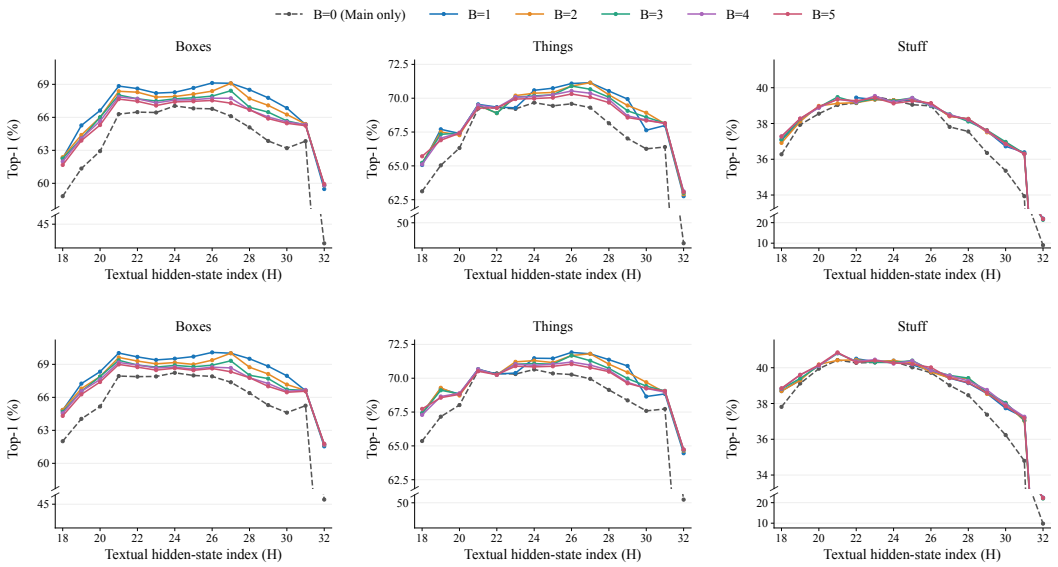


Figure 9: Regional Top-1 across textual depths and ranking budgets B. Top: no refinement; bottom: TALE. Y-axes are broken, with identical scales before and after refinement within each region type.

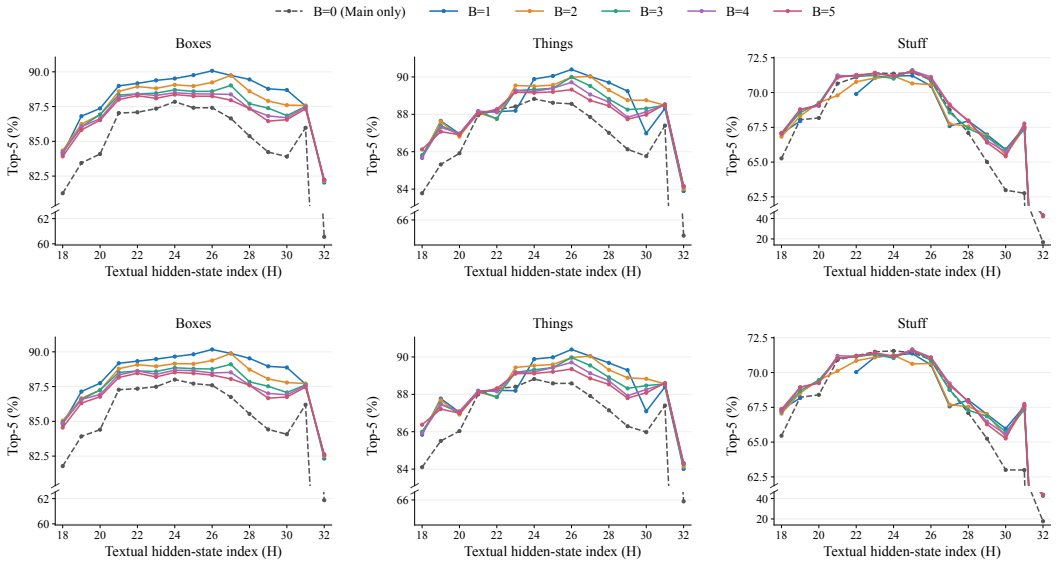
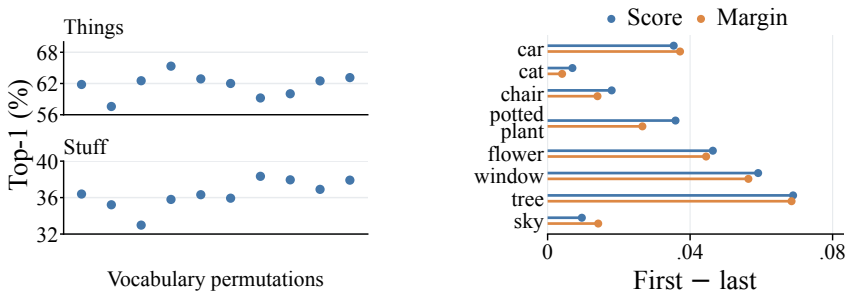


Figure 10: Regional Top-5 for the same configurations as Fig. 9. Top: no refinement; bottom: TALE. Y-axes are broken.

Table 9: Regional Top-1/5/10 accuracy (%).

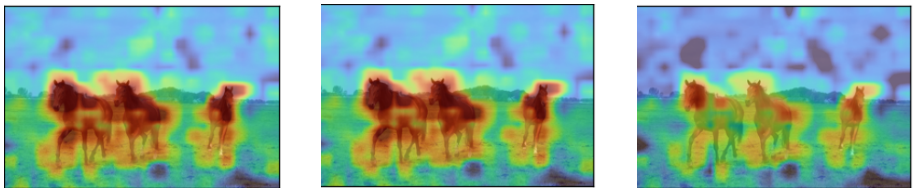
Region	Top-1	Top-5	Top-10
Boxes	65.1	88.5	93.0
Things	67.4	88.3	92.0
Stuff	37.7	68.8	78.3

Vocabulary Scale. The vocabulary-size diagnostic varies the number of input categories over 5, 20, 60, and 133 on the fixed 500-image development set. Fig. 13a reports the annotated-target group, using equal target weights across eight supported target categories. Textual representation drift is measured relative to the canonical order at the same target position and vocabulary composition.



(a) Vocabulary order (b) Score and margin shifts

Figure 11: Order sensitivity with fixed images and vocabulary. Each point in (a) represents one ordering; (b) reports mean first-minus-last score and margin changes.



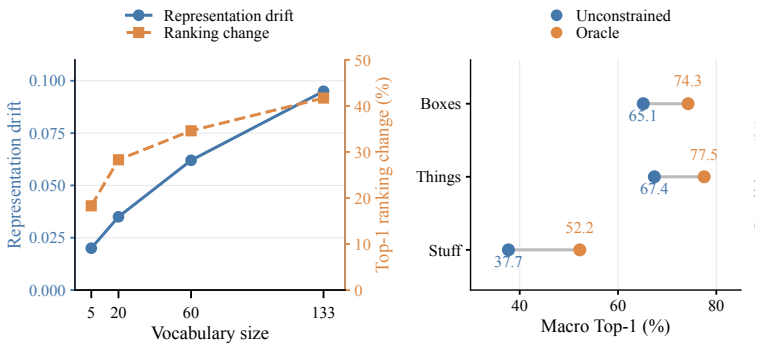
(a) Four-class context
Response: 0.3

(b) Tokens appended
Response: 0.3

(c) Tokens interleaved
Response: 0.2

Figure 12: Textual context changes spatial responses. (a) A diagnostic prompt listing four categories annotated in the image. The vocabulary is expanded to 20 categories by appending their textual tokens after those of horse (b) or interleaving them before those of horse (c). Appending leaves the response nearly unchanged, whereas changing the preceding text alters the spatial activation.

Top-1 changes refer to image-level category ranking, not regional recognition. These experiments assess vocabulary-size sensitivity: changing vocabulary size also changes category content and the available output set, so the results do not isolate token length alone. Confidence intervals and other target groups are retained in the underlying diagnostic reports.



(a) Vocabulary sensitivity (b) Category constraints

Figure 13: Context sensitivity and category constraints. (a) Textual representation drift and image-level Top-1 ranking changes in the vocabulary-size diagnostic. (b) Regional Top-1 with and without oracle category constraints.

Oracle Category Constraints. The oracle experiment in Fig. 13(b) uses image-level annotations only to constrain output categories; visual and textual representations and original similarities remain fixed. Annotation absence is not equivalent to confirmed visual absence. The oracle results therefore quantify a diagnostic opportunity for category selection rather than the performance of a deployable system.

Selective Output Protocol. Full-vocabulary representations and raw scores are reused. We retain 1, 3, 5, 10, 20, 40, 80, or 133 categories per image. GT-guided retention prioritizes annotated categories and fills remaining slots with other categories; it is not an accuracy-optimal upper bound. Regions whose correct category is excluded remain errors. Random controls use five seeds. In the removal diagnostic, random deletion matches the GT-guided deletion count per image. Unannotated categories are not necessarily visually absent. Removing only unannotated channels cannot invalidate an originally correct argmax; its improvement quantifies available selection gains. These are regional classification diagnostics, not segmentation mIoU evaluations. No confidence intervals are reported.

Controlled Context Expansion. The context-expansion experiment uses 100 calibration images and 500 disjoint evaluation images under the 133-category regional classification protocol. Results are point estimates without confidence intervals.

To distinguish adding textual context from expanding the prediction set, we vary either factor or both while keeping visual descriptors fixed (Table 10). Interleaving textual tokens corresponding to 80 additional categories while restricting predictions to annotated categories raises Boxes/Things Top-1 from 68.9%/74.2% to 73.9%/76.4%. Including those categories among the classification choices reduces accuracy to 62.0%/64.4%. Expanding the prediction set with fixed textual representations also reduces accuracy across all three region types. These controls distinguish contextual effects from the effects of adding classification choices; they do not establish the effectiveness of model-inferred category selection.

Visual descriptors are fixed for scoring. Context-only changes the input category vocabulary while restricting outputs to annotated categories; competition-only keeps the full-vocabulary category representations fixed while expanding eligible outputs. Combined controls vary both. Table 10 averages five fixed seeds. The GT-only starting vocabulary is an oracle diagnostic, not a deployable input. Competition-only starts from different category representations than the other controls, so their differences cannot be added as an effect decomposition. Under fixed scores, extra unannotated output classes can displace correct predictions; non-increasing accuracy is partly structural, not evidence of representation damage. Separate forwards do not have bitwise-identical visual prefixes, but all scoring conditions reuse the same stored visual field.

Table 10: Top-1 as extra category count grows from 0 to 80. Textual tokens for the additional categories are appended or interleaved; output competition is controlled separately.

Control	Boxes	Things	Stuff
Context only, append	$68.9 \rightarrow 68.9$	$74.2 \rightarrow 74.3$	$56.4 \rightarrow 56.4$
Context only, interleave	$68.9 \rightarrow 73.9$	$74.2 \rightarrow 76.4$	$56.4 \rightarrow 56.2$
Competition only	$74.3 \rightarrow 67.5$	$77.5 \rightarrow 69.9$	$52.2 \rightarrow 41.0$
Combined, append	$68.9 \rightarrow 65.9$	$74.2 \rightarrow 70.9$	$56.4 \rightarrow 52.3$
Combined, interleave	$68.9 \rightarrow 62.0$	$74.2 \rightarrow 64.4$	$56.4 \rightarrow 37.9$

Summary. Textual context and prediction alternatives have distinct effects: additional context can help even when additional competing categories hurt recognition. Ranking coverage and oracle selection reveal room for refining category competition, but do not by themselves establish the reliability of model-generated presence evidence.

B Regional Recognition and Presence Refinement

B.1 Recognition and Component Ablations

Recognition across MLLM Variants. Table 11 provides additional configurations for the MLLM variants evaluated in Table 2. For LLaVA-1.5-Vicuna-7B, the single-layer reference uses L31, whereas the main branch in our component ablation uses L30.

Table 11: Zero-shot semantic classification across LLaVA variants (Top-1/Top-5, %).

Model	Configuration	Boxes		Things		Stuff	
		Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
LLaVA-1.5-Vicuna-7B	Single layer	42.3	60.6	47.2	64.9	9.0	16.8
	Aggregation	65.1	88.7	67.0	88.6	39.5	**72.7**
	TALE	**67.3**	**88.9**	**68.9**	**88.7**	**40.2**	72.4
LLaVA-1.6-Vicuna-7B	Single layer	68.1	89.8	69.8	89.2	42.5	**73.7**
	Aggregation	69.8	**90.2**	71.4	**89.6**	42.0	71.9
	TALE	**71.0**	90.1	**72.2**	**89.6**	**42.7**	71.8
LLaVA-1.6-Vicuna-13B	Single layer	41.9	61.7	48.3	68.7	49.1	81.6
	Aggregation	78.4	**95.1**	81.6	**94.9**	49.6	**82.2**
	TALE	**80.0**	94.5	**82.5**	94.4	**50.9**	81.8
LLaVA-1.6-Mistral-7B	Single layer	33.3	47.7	38.1	52.6	34.4	60.1
	Aggregation	57.4	78.0	62.5	81.1	37.1	63.7
	TALE	**60.1**	**78.6**	**64.7**	**81.6**	**39.1**	**63.9**

Cross-Dataset Classification. Table 12 reports Top-5 accuracy

Table 12: Cross-dataset zero-shot classification (Top-5, %).

Configuration	VOC20	VOC21	PC59	PC60	ADE150	Stuff171	Object81
Single layer	**99.6**	99.2	85.1	81.6	68.0	66.2	89.7
Aggregation	**99.6**	**99.3**	**90.1**	**88.3**	73.6	**73.8**	**91.3**
TALE	99.4	**99.3**	89.9	88.0	**73.8**	**73.8**	**91.3**

Fusion-Weight Control. Table 13 compares branch weights for Things before presence verification. Both settings use the same H32 category representations, L30 main branch, mean(L20–L24) auxiliary branch, and 133-category prediction range. Equal weighting gives higher regional Top-1 and Top-5 accuracy than the 3:1 weight used in the reported configuration. Boxes and Stuff use 1:1 in both configurations.

Table 13: Aggregation weights with verification disabled (Things Top-1/Top-5, %).

Aggregation weight		Things	
Main	Aux.	Top-1	Top-5
1	1	**67.6**	**89.3**
3	1	67.0	88.6

Summary. The cross-model and transfer results support the broader applicability of aggregation and presence-guided refinement, while the fusion-weight control shows that performance depends on the branch weights. Benefits remain configuration-dependent: the auxiliary branch exceeds fusion on Stuff Top-1 (41.6% vs. 39.5%) in the fixed-configuration ablation, and Top-1 gains need not improve Top-5 accuracy.

B.2 Presence Verification and Refinement

Presence-Query Implementation. We independently query the full image using the model's conversation template:

> Does this image contain the category "{category}"? If no, answer "No." If yes, answer "Yes, it contains {category}." Do not add a description or explanation.

Valid completed responses match `No` or `Yes, it contains {category}` after substituting the queried display name. Matching ignores case and surrounding whitespace and allows an optional final period or exclamation mark. Other responses remain unresolved and unadjusted. Before applying $\log q_c$, we clip the Yes preference to $[10^{-6}, 1 - 10^{-6}]$.

Presence-Response Refinement. We compare how the same presence responses affect regional recognition while holding the visual readout, triggered regions, and nominated categories fixed (Table 14). All rules reuse the same representations, triggered regions, Top-3 nominations, and independent presence responses on 5,000 images. The visual configurations were selected separately for each region type on 100 development images; they differ from the fixed L30 main branch and mean(L20–L24) auxiliary branch used in the main component ablation. Yes-only, No-only, and Yes/No apply $\lambda \log q_c$ to nominated similarities after the indicated valid responses; q_c is the Yes preference defined in Methods. The fixed-penalty control subtracts λ after a valid No. Here $\tau = \lambda = 0.01$; unresolved responses leave similarities unchanged.

Table 14: Presence-guided refinement with fixed representations and candidate nominations (Top-1/Top-5, %).

Refinement	Boxes		Things		Stuff	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
None	69.1	90.1	67.3	87.4	39.3	**71.3**
Yes-only	68.8	90.1	66.8	87.4	38.7	71.2
Yes/No	70.1	**90.2**	69.1	**87.5**	40.3	71.2
No-only (TALE)	**70.2**	**90.2**	**69.4**	**87.5**	**40.5**	71.2
No-only, fixed penalty	**70.2**	90.1	69.3	**87.5**	40.2	71.2

Verification by Original Error Type. We analyze 5,000 COCO images, including the 100 development images, using the frozen development-selected configuration and compare No-only and Yes/No refinement (Table 15). Both rules reuse the same representations, triggered regions, regional Top-3 nominations, and cached presence responses, with $\tau = \lambda = 0.01$; no additional inference is required. Image-level category annotations are the union of all Things and Stuff panoptic annotations before region filtering or capping.

Regional correctness is determined by comparing the Top-1 prediction with the ground-truth category of the evaluated box or mask. Among originally incorrect regions that trigger verification, U denotes predictions outside the image-level annotation set, and A denotes predictions inside that set. Both groups are regionally incorrect; annotation membership alone does not establish regional correctness. A correction occurs only when the refined Top-1 matches the regional ground truth. Originally correct predictions that become incorrect are counted separately as damage.

For Table 7, negative response measures whether the original predicted category receives a valid No; GT in Top-3 measures whether the region's ground-truth category is among its original nominations. All three rates use the triggered errors in the corresponding U/A group as their denominator, including unresolved responses. Category annotations are used only for diagnostic grouping and evaluation, not as inputs to verification or similarity refinement. Unannotated categories are not necessarily visually absent.

For Table 15, Regions counts all regions in each original prediction group, and Triggered reports the fraction that undergo verification. The remaining four columns use all triggered regions in the corresponding group as their denominator, including unresolved responses. The No-only and Yes/No columns report wrong-to-correct transitions for incorrect groups and correct-to-wrong transitions

for the correct group. Correction rates are not restricted to regions with GT in Top-3. These rates weight regions equally, rather than averaging accuracy over categories. No-only penalizes valid No responses, whereas Yes/No penalizes both valid Yes and No responses; invalid or unresolved responses remain neutral.

Table 15: Verification outcomes by original prediction group and refinement rule (%).

Region	Original prediction	Regions	Triggered	Negative response	GT in Top-3	Corrected / damaged	
						No-only	Yes/No
Boxes	Incorrect (U)	5,218	45.4	74.2	45.3	23.2	23.1
	Incorrect (A)	7,287	30.8	24.3	54.0	6.9	10.5
	Correct	22,885	8.9	–	–	4.5	5.1
Things	Incorrect (U)	9,765	51.8	84.5	31.4	20.3	19.3
	Incorrect (A)	6,502	36.5	36.8	34.0	8.4	8.7
	Correct	19,123	10.7	–	–	4.8	12.5
Stuff	Incorrect (U)	6,171	45.4	83.0	42.1	13.1	12.1
	Incorrect (A)	4,263	31.3	50.9	42.8	6.4	5.3
	Correct	9,739	14.7	–	–	27.3	23.1

Under No-only refinement, errors involving unannotated predicted categories are corrected more often despite lower GT Top-3 coverage. The corrected/harmed counts are 705/92 for Boxes, 1,225/98 for Things, and 453/391 for Stuff, compared with 783/105, 1,184/256, and 411/331 under Yes/No refinement. Removing the penalty on Yes responses reduces damage for Boxes and Things but increases it for Stuff. No-only also yields fewer net corrected regions for Boxes and Stuff, despite their higher Top-1 averaged over categories in Table 14; region counts and accuracy averaged over categories weight errors differently.

Summary. Negative responses provide useful evidence for regional refinement, especially when the original predicted category is unannotated in the image. No-only refinement improves Top-1 averaged over categories over Yes/No refinement in this comparison, but its higher Stuff damage rate shows that it is not uniformly safer at the region level.

C Inference Cost

C.1 Textual Representation Construction

Matched Recognition and Runtime Comparison. Table 16 compares Full and Nearby with identical visual representations and readout, the same final 133-category vocabulary, and no presence-guided refinement. Measurements use a single GPU, sequential execution, batch size one, and no inference caching. Recognition and timing are measured on the same 100 images; Top-1 and Top-5 are both averaged over categories. Nearby takes 76.5 times as long for textual representation construction and 31.2 times as long overall as Full. Peak allocated memory remains almost unchanged because the target-specific contexts are processed sequentially, rather than held in memory simultaneously. Nearby improves Top-1 but reduces Stuff Top-5 by 1.3 points. This comparison supports Full as a lower-latency construction scheme under the measured implementation; it does not establish an optimal accuracy–cost trade-off, and does not evaluate batching or inference-cache optimizations for Nearby.

Summary. Full offers substantially lower construction and total latency than target-specific Nearby contexts under the measured sequential implementation, with similar peak memory. This motivates Full as a practical construction choice, rather than an accuracy optimum or an established best accuracy–cost trade-off.

C.2 Runtime, Memory, and Verification Workload

Stage-Wise Latency. Table 17 reports total and stage-wise inference latency.

Table 16: Matched Full–Nearby comparison on 100 images.

Metric	Full	Nearby (16)
Things Top-1 (%)	64.0	68.3
Things Top-5 (%)	86.2	87.6
Stuff Top-1 (%)	30.1	36.7
Stuff Top-5 (%)	71.3	70.1
Textual representation construction (ms/image)	146.8	11227.9
Total latency (ms/image)	366.8	11451.1
Peak allocated memory (GiB)	13.7	13.7

Each configuration uses the same 100 measured images after 10 warm-up images. Main uses L30 and Aux. averages L20–L24, with separate normalization; Main:Aux. weights are 1:1, 3:1, and 1:1 for Boxes, Things, and Stuff, respectively. Representation construction includes visual processing and spatial visual–text similarity. Total additionally includes input preparation (5.8–6.0 ms) and regional pooling (19.9–23.5 ms), so displayed stages do not sum to Total. Measurements include stage-boundary CUDA synchronization and exclude model loading, logging, and evaluation. The nomination-only row does not restrict prediction categories; its lower total than Aggregation reflects timing variation in other stages.

Table 17: Inference latency on one RTX 5090 (ms/image).

Configuration	Total	Representation construction	Candidate nomination	Verification	Prediction
Main only	437.9	409.1	0.0	0.0	0.2
Aggregation	1001.5	971.8	0.0	0.0	0.2
Aggregation + nomination	929.9	902.3	0.2	0.0	0.1
Aggregation + refinement	3902.2	918.6	0.2	2957.3	0.4

Memory and Verification Workload. Table 18 reports memory, query workload, and estimated computation for the same timing experiment.

Memory is peak GPU memory; other entries are averages over the same 100 images. Verification makes 853 distinct image–category queries and generates 3,648 tokens, including end-of-sequence tokens. Every configuration uses one representation-construction forward pass per image; verification adds 36.5 generation forward passes on average. Reserved memory includes allocated memory. FLOPs are analytical estimates of major matrix and convolution operations, not profiler measurements; the range reflects unknown per-query generation lengths, not a confidence interval.

Table 18: Memory and verification workload for the latency comparison.

Configuration	Allocated (GiB)	Reserved (GiB)	Queries/img.	Generated tokens/img.	Estimated TFLOPs/img.
Main only	13.5	13.6	0.0	0.0	15.2
Aggregation	13.5	13.6	0.0	0.0	15.2
Aggregation + nomination	13.5	13.6	0.0	0.0	15.2
Aggregation + refinement	13.5	13.6	8.5	36.5	353.8–364.8

Measurement Scope. These timings execute representation construction and independent presence verification without feature or response caches. Query counts in the 5,000-image accuracy experiments and latency on the 100-image timing set have different denominators. Query-count reductions alone do not establish proportional latency reductions; the results above report measured end-to-end inference.

Summary. Presence verification dominates latency without substantially increasing peak allocated memory. Query counts and estimated FLOPs cannot replace measured end-to-end latency.