ONE IMAGE, PARALLEL DECISIONS: STRUCTURED QUERYING FOR ZERO-SHOT MULTI-LABEL RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative multimodal large language models (MLLMs) combine visual understanding with language-conditioned reasoning, offering considerable potential for zero-shot multi-label recognition. However, their generative interface does not naturally provide scores for multiple non-exclusive categories: label generation may omit candidates, while independent class-wise querying repeatedly processes shared context. We propose Structured Parallel Querying (SPQ), an inference framework that organizes class-presence questions as isolated suffixes over a shared visual and instructional prefix. class-isolated attention and aligned positions preserve the context of standalone queries, enabling parallel Yes/No scoring without autoregressive answer generation. We systematically evaluate SPQ across MLLM families and scales, examining recognition performance, component interactions, computational scalability, and sensitivity to input resolution and answer verbalizers. Category-level and qualitative analyses reveal pronounced advantages on several categories associated with small objects, alongside backbone- and dataset-dependent limitations. We further investigate parameter-efficient fine-tuning using only seen-category supervision and evaluate its transfer to unseen categories. Together, these experiments establish SPQ as a practical interface for studying and exploiting the potential of generative MLLMs in zero-shot multi-label recognition.

1 INTRODUCTION

Zero-shot multi-label recognition aims to identify multiple non-exclusive categories in an image without labeled examples for the target categories. Contrastive vision–language models such as CLIP provide a natural foundation for this task. Through large-scale contrastive pretraining on image–text pairs, CLIP aligns visual and textual representations, allowing novel category names to be converted into text embeddings and directly scored against an image (Radford et al., 2021; Jia et al., 2021). This interface has supported extensive research on zero-shot multi-label recognition, including prompt adaptation, positive–negative contexts, and label co-occurrence modeling (Sun et al., 2022; Ali & Khan, 2023; Xie et al., 2026).

Multimodal large language models (MLLMs) offer a potentially richer foundation. By connecting visual encoders with language models trained or instruction-tuned on large-scale multimodal data, MLLMs can interpret open-vocabulary concepts, compositional descriptions, visual context, and language-conditioned questions (Li et al., 2023; Liu et al., 2023; Bai et al., 2023; Yao, 2025). However, their generative interface does not naturally provide a complete vector of class-wise scores. As illustrated in Fig. 1, direct label generation produces a variable-length sequence that may omit candidate categories and requires textual parsing. Multi-token category names and previously generated labels also introduce autoregressive dependencies, making generated likelihoods difficult to interpret as independent multi-label confidence scores (Ma et al., 2025). Related MLLM-based recognition studies similarly report difficulties in obtaining consistent outputs and reliable confidence scores from open-ended generation (Wen et al., 2025; Xuan et al., 2026).

Asking one binary question per category provides a more direct scoring interface, but naive independent querying repeatedly processes the same visual and instructional context. Simply concatenating

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.

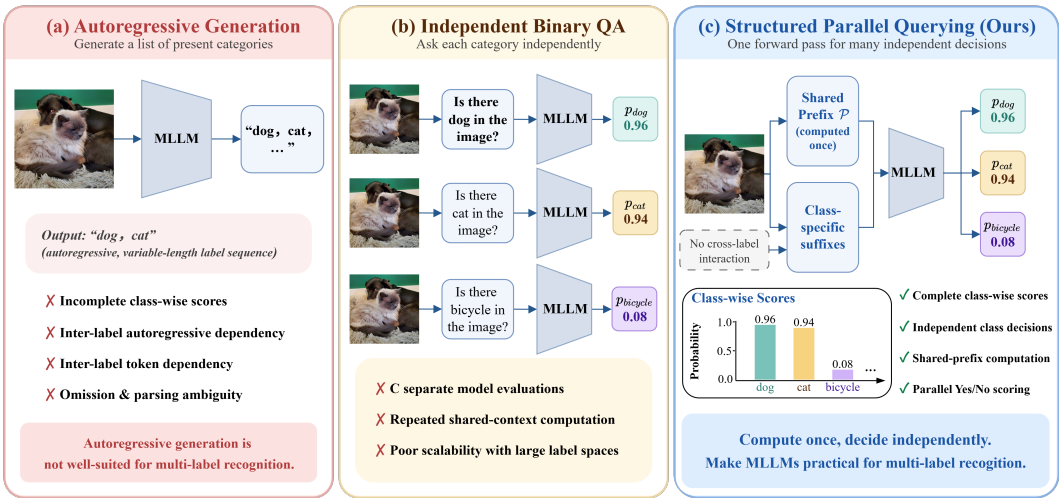


Figure 1: Structured Parallel Querying enables independent multi-label decisions through a shared visual prefix.

all questions is not sufficient either. Under standard causal attention, later queries can access earlier class queries, and their positional context changes with the query order. We propose *Structured Parallel Querying* (SPQ), a general interface for applying generative MLLMs to zero-shot multi-label recognition. SPQ places the shared visual and instructional context in a common prefix and represents each candidate category as an independent suffix. A class-isolated attention mask allows each suffix to access the shared prefix and its own tokens while blocking attention between different suffixes. Aligned positional encodings match the context of standalone queries, and Yes/No next-token logits at the suffix decision positions are read out jointly to produce all class-wise scores in one forward pass.Our contributions are:

- We identify the missing class-wise querying and scoring interface that limits the direct use of generative MLLMs for zero-shot multi-label recognition.
- We propose Structured Parallel Querying, which combines shared-prefix organization, class-isolated attention, aligned positions, and parallel Yes/No logit readout.
- We evaluate this interface across MLLM families and scales, studying recognition quality, component interactions, and computational scalability.

2 Related Work

2.1 Zero-Shot Multi-Label Recognition

Zero-shot multi-label recognition extends zero-shot learning from single-label prediction to the simultaneous recognition of multiple, potentially unseen categories. Early approaches transfer semantic information from seen to unseen categories through attributes, word embeddings, label embeddings, and visual–semantic mappings (Changpinyo et al., 2016; Zhang et al., 2016; Lee et al., 2018). Other methods explicitly model label correlations or exploit semantic relations between categories to improve multi-label prediction (Fu et al., 2015; Lee et al., 2018; Liu et al., 2024).

Large-scale vision–language pretraining introduced a more direct route to zero-shot recognition. CLIP aligns image and text representations through contrastive learning, allowing arbitrary category descriptions to be used as classifier prototypes (Radford et al., 2021). DualCoOp learns positive and negative textual contexts for each category (Sun et al., 2022), while CLIP-Decoder extracts multi-label predictions from CLIP-aligned representations(Ali & Khan, 2023). Language-driven cross-modal classifiers improve the semantic representation of candidate labels (Liu et al., 2024), while recent training-free approaches improve CLIP-based multi-label recognition through local refinement, score prompting, and label co-occurrence modeling (Lin et al., 2024; Miller et al., 2025; Xie et al., 2026). Other work has investigated LLM-generated prompts and hierarchical textual

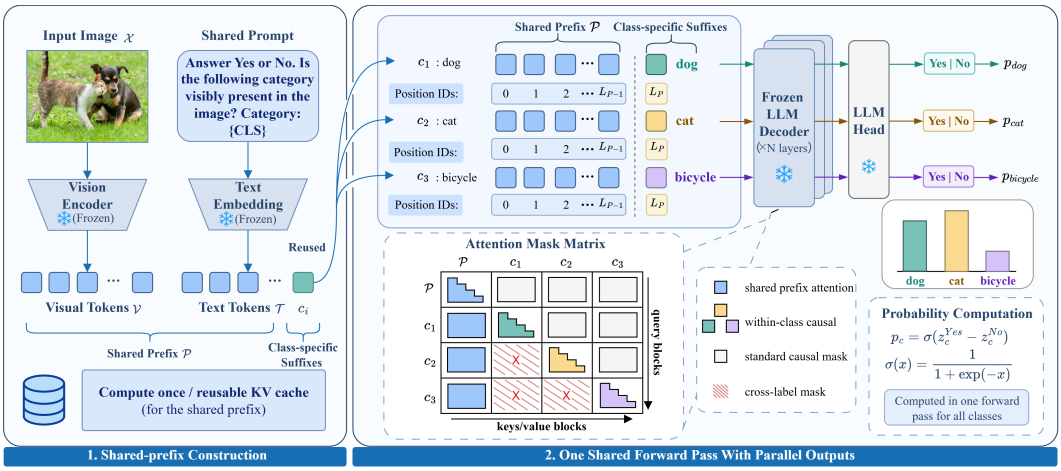


Figure 2: Model Architecture of Structured Parallel Querying.

structures for adapting CLIP to multi-label recognition (Yang et al., 2024; Wang et al., 2025). These methods primarily rely on a discriminative image–text similarity interface, where each candidate category can be assigned a score independently.

2.2 MLLMs for Visual and Multi-Label Recognition

MLLMs extend vision–language modeling by connecting visual encoders with large language models and adapting them to multimodal instruction-following tasks. BLIP-2 uses a lightweight query transformer to connect frozen visual and language models (Li et al., 2023). LLaVA demonstrates that visual instruction tuning can turn a language model into a general visual assistant (Liu et al., 2023). Qwen-VL and InternVL further improve visual resolution, grounding, and multimodal reasoning capabilities (Bai et al., 2023; Chen et al., 2024).

Despite this potential, generative MLLMs remain less explored as general foundations for zero-shot multi-label recognition. Existing studies have investigated graph-augmented multimodal reasoning for multi-label zero-shot learning (Wen et al., 2025), deterministic generation and candidate scoring for zero-shot human–object interaction recognition (Xuan et al., 2026), and the behavior of autoregressive language models in multi-label prediction (Ma et al., 2025). These works demonstrate that MLLMs can provide useful visual and semantic knowledge, but they also reveal a mismatch between open-ended generation and multi-label scoring. Our work focuses on establishing a general class-wise querying and scoring interface that exposes MLLM capabilities for zero-shot multi-label recognition.

2.3 Structured and Parallel Inference

A separate line of work studies how multiple candidate decisions or reasoning branches can share computation while remaining conditionally independent. Unified multiple-choice models use option-specific masks and position assignments to prevent different candidate options from influencing one another (Wang et al., 2023). Parallel decoding methods construct isolated branches that reuse a shared non-parallel prefix and align the positions of parallel tokens (Yao et al., 2026). Other systems reuse or decompose computation over shared prefixes and distinct suffixes to improve structured or high-throughput inference (Juravsky et al., 2024; Athiwaratkun et al., 2024; Gim et al., 2024; Zheng et al., 2024; Yao et al., 2025; Zheng et al., 2026).

These methods target text classification, reasoning, verification, or inference systems. SPQ instead combines shared-context computation with multimodal class-presence questions and parallel Yes/No scoring. Its contribution is a practical multi-label inference interface, rather than a new generic attention primitive.

3 METHOD

3.1 PROBLEM SETUP

Let $\mathcal{X}$ denote an input image and let $\mathcal{C} = \{c_i\}_{i=1}^{C}$ denote a set of C candidate categories. The objective of zero-shot multi-label recognition is to determine whether each candidate category is visibly present in $\mathcal{X}$, without training a category-specific classifier. The desired output is a class-wise score vector

$$\mathbf{p}(\mathcal{X}) = [p_1(\mathcal{X}), \ldots, p_C(\mathcal{X})],$$

where $p_i(\mathcal{X})$ denotes the presence score of category c_i.

SPQ evaluates binary class-presence questions jointly while retaining the attention dependencies and positional context of standalone queries. Figure 2 summarizes the construction.

3.2 STRUCTURED CLASS-PRESENCE QUERIES

For each candidate category c_i, we formulate recognition as a binary class-presence question:

```
Answer Yes or No.
Is the following category visibly present in the image?
Category: {CLS},
```

where $\{\texttt{CLS}\}$ is replaced by the category name c_i. The instruction requests answers using the two tokens `Yes` and `No`, converting open-ended generation into a well-defined binary decision while retaining the MLLM's visual understanding and language-conditioned reasoning.

The category name is placed at the end of the query so that the image and instruction can be shared by all candidate categories. The category-specific tokens determine which class is being evaluated, whereas the model does not need to generate a variable-length list of labels. Therefore, every candidate category has an explicit decision position and produces a directly comparable score.

Let $\mathcal{V}(\mathcal{X})$ denote the visual tokens produced by the vision encoder and let $\mathcal{T}$ denote the tokens of the shared instruction. We construct the shared prefix

$$\mathcal{P} = [\mathcal{V}(\mathcal{X}); \mathcal{T}].$$

For category c_i, let $\mathcal{B}_i$ denote its complete class-specific suffix. This suffix includes the category tokens and any template tokens required by the corresponding MLLM to reach the answer-prediction position, such as separators or an assistant-start marker. The packed input is

$$\mathcal{S} = [\mathcal{P}; \mathcal{B}_1; \ldots; \mathcal{B}_C].$$

Figure 2 illustrates the decomposition into a reusable shared prefix and class-specific suffixes.

3.3 CLASS-ISOLATED ATTENTION

Simply concatenating all category suffixes is insufficient. Under the standard causal attention pattern, tokens in an earlier suffix can be attended to by tokens in later suffixes. Consequently, the probability assigned to a later category may depend on which categories appeared before it. Changing the order of the candidate categories can then change the predicted scores, even though the image and candidate set remain unchanged. Such order dependence contradicts the independent decision structure of multi-label recognition.

We eliminate this dependency with a class-isolated attention mask. For a query position q belonging to category block $\mathcal{B}_i$, the decoder may attend to the shared prefix and to preceding tokens within the same block, but cannot attend to tokens from another category block. Let M denote the attention mask, where $M_{qk} = 0$ indicates an allowed connection and $M_{qk} = -\infty$ indicates a blocked connection. We define

$$M_{qk} = \begin{cases} 0, & q, k \in \mathcal{P},\ k \le q, \\ 0, & q \in \mathcal{B}_i,\ k \in \mathcal{P}, \\ 0, & q, k \in \mathcal{B}_i,\ k \le q, \\ -\infty, & \text{otherwise.} \end{cases}$$

The first case preserves causal attention within the shared prefix. The second lets each suffix access the common visual and instructional evidence, and the third retains causal attention within that suffix. All remaining connections are blocked, including attention between different category suffixes. Consequently, a category score does not depend on the content or ordering of other suffixes when their native positional context is also preserved.

3.4 Unified Position Assignment

Packing the suffixes consecutively creates a positional discrepancy. If the physical positions in the packed sequence were used directly, the second and subsequent categories would receive different positional contexts from the first category. Their scores could then depend on their locations in the packed sequence rather than only on their category content.

To remove this effect, we copy each suffix's native position identifiers from its standalone query. For one-dimensional positional encodings, this assigns each suffix the same starting position relative to the shared prefix. Let $L_{\mathcal{P}} = |\mathcal{P}|$ and $L_i = |\mathcal{B}_i|$. Regardless of its physical location in $\mathcal{S}$, the position identifiers of $\mathcal{B}_i$ are assigned as

$$\mathrm{pos}(\mathcal{B}_i) = [L_{\mathcal{P}}, L_{\mathcal{P}} + 1, \ldots, L_{\mathcal{P}} + L_i - 1].$$

Therefore, every category query starts immediately after the same shared prefix from the perspective of positional encoding. For multidimensional positional encodings, we retain the original coordinate vectors. Combined with attention isolation, this preserves standalone-query positions and removes suffix-order dependence in the mathematical computation, up to implementation-level numerical differences.

3.5 Parallel Yes/No Scoring

As illustrated in Figure 2, SPQ uses the MLLM's original decoder and language-model head to obtain class-presence scores. For each category c_i, let r_i denote the final input-token position of its complete query suffix $\mathcal{B}_i$, immediately before the answer begins. This suffix includes any answer-prefix tokens required by the model's chat template. Under the MLLM's next-token prediction mechanism, the representation at r_i is used to predict the first answer token.

Let $h_{r_i} \in \mathbb{R}^d$ be the final decoder representation at this position, including the model's native output normalization, where d is the hidden dimension. The original language-model head projects this representation onto the token vocabulary as follows.

$$\mathbf{z}_i = W_{\mathrm{LM}} h_{r_i} + b_{\mathrm{LM}}, \qquad \mathbf{z}_i \in \mathbb{R}^{|\Omega|}, \tag{1}$$

where Ω is the vocabulary, $W_{\mathrm{LM}} \in \mathbb{R}^{|\Omega| \times d}$ is the output projection matrix, and b_{LM} is its bias (set to zero when absent in the original model). Each entry of $\mathbf{z}_i$ is an unnormalized score for a possible next token.

Let v_{Yes} and v_{No} denote the vocabulary indices of the single-token answers `Yes` and `No`, respectively, as determined by the model's tokenizer in the answer context. We then select the corresponding logits.

$$z_i^{\mathrm{Yes}} = \mathbf{z}_i[v_{\mathrm{Yes}}], \qquad z_i^{\mathrm{No}} = \mathbf{z}_i[v_{\mathrm{No}}]. \tag{2}$$

These two values express the model's preference for an affirmative or negative answer to the query for c_i; they are not themselves probabilities.

To obtain the class-presence score, we apply a two-way softmax to the selected logits as follows.

$$\begin{aligned} p_i(\mathcal{X}) &= \frac{\exp(z_i^{\mathrm{Yes}})}{\exp(z_i^{\mathrm{Yes}}) + \exp(z_i^{\mathrm{No}})} \\ &= \frac{1}{1 + \exp\left[-\left(z_i^{\mathrm{Yes}} - z_i^{\mathrm{No}}\right)\right]} = \sigma\left(z_i^{\mathrm{Yes}} - z_i^{\mathrm{No}}\right), \end{aligned} \tag{3}$$

where $\sigma(u) = 1/(1 + \exp(-u))$ is the sigmoid function. Thus, $p_i(\mathcal{X})$ is the probability assigned to `Yes` after renormalizing the next-token distribution over the two answers $\{\texttt{Yes}, \texttt{No}\}$.

The normalization is performed separately for each category, rather than across categories, so multiple categories can simultaneously receive high scores. Reading the logits at all decision positions

yields $\mathbf{p}(\mathcal{X}) = [p_1(\mathcal{X}), \ldots, p_C(\mathcal{X})]$ in one structured forward pass, without generating answer tokens or introducing an additional classification head.

The packed construction computes the shared prefix once and reads all suffix endpoints jointly.

4 EXPERIMENTS

4.1 DATASETS

We evaluate SPQ on four multi-label recognition benchmarks, PASCAL VOC2007, MS-COCO2014, NUS-WIDE, and OpenImages V4. All experiments follow the zero-shot learning (ZSL) setting, where predictions and evaluation are restricted to the unseen category set, rather than the union of seen and unseen categories used in generalized zero-shot learning.

For PASCAL VOC2007 (Everingham et al., 2010), we use a partition of 16 seen and 4 unseen categories and evaluate the four unseen categories on the 4,952 images in the official test set. For MS-COCO2014, we adopt the 48-seen/17-unseen category partition used in prior multi-label ZSL work and evaluate only the 17 unseen categories (Ben-Cohen et al., 2021). For NUS-WIDE, we follow the established protocol that treats 925 tags as seen categories and the 81 human-annotated concepts as unseen categories (Huynh & Elhamifar, 2020; Ben-Cohen et al., 2021). For OpenImages V4, we use the 7,186-seen/400-unseen partition and evaluate recognition over the 400 unseen categories (Huynh & Elhamifar, 2020; Ben-Cohen et al., 2021). These benchmarks cover unseen label spaces ranging from 4 to 400 categories, enabling evaluation of SPQ at different query scales.

4.2 EXPERIMENTAL DETAILS

We instantiate SPQ with four MLLM configurations: Qwen2.5-VL-3B, Qwen2.5-VL-7B, LLaVA-1.5-7B, and InternVL3.5-4B. This selection covers three model families and multiple model scales, allowing us to assess whether the proposed querying interface generalizes across backbones. In the main ZSL experiments, the pretrained model parameters remain unchanged, and no task-specific training is performed. Here, "unseen" refers to the benchmark category partition and does not imply that the corresponding concepts were absent from MLLM pretraining.

The main evaluation uses 224×224 inputs and the class-presence construction in Section 3. Supplementary prompt variants are documented in Appendix A. Efficiency comparisons with independent binary querying use the same backbone, prompt, and scoring rule.

We report mean average precision (mAP), computed by averaging the average precision of individual unseen categories. We additionally report precision, recall, and F1 at fixed prediction budgets, denoted by P@k, R@k, and F1@k. For each image, the k highest-scoring unseen categories are selected as predictions. We use $k \in \{1, 2\}$ for PASCAL VOC2007, $k \in \{3, 5\}$ for MS-COCO2014 and NUS-WIDE, and $k \in \{10, 20\}$ for OpenImages V4. All metrics are evaluated against annotations restricted to the corresponding unseen category set. Additional implementation details and the protocols for the supplementary experiments are provided in Appendix A.

Comparison with baseline methods. Table 1 compares SPQ with vanilla CLIP and two training-free CLIP-based methods, TagCLIP (Lin et al., 2024) and SPARC (Miller et al., 2025). All baseline results are locally reproduced using the same benchmark splits and evaluation protocol as SPQ. All methods use frozen pretrained models without task-specific training. SPQ achieves the highest mAP on VOC and COCO: Qwen2.5-VL-3B reaches 94.45% on VOC, exceeding TagCLIP by 3.15 percentage points, while Qwen2.5-VL-7B reaches 91.30% on COCO, surpassing SPARC by 5.89 points. On COCO, all four SPQ backbones also outperform the CLIP-based baselines in both Top-3 and Top-5 F1, indicating that the benefits extend beyond a single MLLM configuration.Results on NUS-WIDE and OpenImages reveal a more metric-dependent comparison. On NUS-WIDE, SPARC achieves the highest mAP of 53.18%, compared with 50.75% for SPQ with Qwen2.5-VL-7B. However, the latter achieves Top-3 and Top-5 F1 scores of 54.87% and 50.15%, exceeding the strongest CLIP-based baseline for each metric by 17.04 and 15.52 points, respectively. On OpenImages, SPQ with Qwen2.5-VL-7B reaches 73.23% mAP, trailing SPARC by 0.95 points, whereas Qwen2.5-VL-3B achieves the highest Top-10 F1 of 19.85%; CLIP retains the best Top-20 F1. These differences reflect the distinct evaluation axes: mAP measures image ranking within each class, whereas Top-k

Table 1: Zero-shot multi-label recognition on four datasets. All results are reported in percentages over unseen categories.

Method	VOC							COCO						
	mAP	Top-1			Top-2			mAP	Top-3			Top-5		
		P	R	F1	P	R	F1		P	R	F1	P	R	F1
CLIP (ViT-L/14)	82.43	32.89	89.40	48.09	17.93	97.47	30.29	73.36	32.27	80.18	46.02	20.89	86.52	33.66
CLIP (ViT-B/16)	85.70	33.40	90.78	48.83	18.15	98.68	30.67	74.73	32.89	81.73	46.91	21.51	89.07	34.65
CLIP (ViT-B/32)	81.63	32.71	88.91	47.83	18.08	98.30	30.55	69.55	32.38	80.46	46.18	21.10	87.38	33.99
CLIP (RN50)	83.17	33.52	91.11	49.01	18.11	98.46	30.60	68.97	32.45	80.64	46.28	21.41	88.68	34.50
TagCLIP (ViT-B/16)	91.30	34.55	93.90	50.51	18.34	99.72	30.99	84.89	34.44	85.58	49.12	22.56	93.44	36.35
SPARC (ViT-B/16)	89.74	32.99	89.68	48.24	18.01	97.91	30.42	85.41	34.56	85.88	49.29	22.20	91.95	35.77
SPQ (Qwen2.5-VL-3B)	**94.45**	35.00	95.12	51.17	18.36	99.78	31.01	88.78	37.74	93.78	53.82	23.48	97.24	37.83
SPQ (Qwen2.5-VL-7B)	93.26	34.31	93.25	50.16	18.24	99.12	30.80	**91.30**	**38.82**	**96.45**	**55.36**	**23.86**	**98.79**	**38.43**
SPQ (LLaVA-1.5-7B)	87.56	**35.06**	**95.28**	**51.26**	**18.40**	**100.00**	**31.08**	87.38	38.31	95.18	54.63	23.77	98.42	38.29
SPQ (InternVL3.5-4B)	91.30	34.35	93.36	50.22	18.22	99.01	30.77	82.54	36.68	91.14	52.31	23.15	95.87	37.30

Method	NUS-WIDE							Openimages						
	mAP	Top-3			Top-5			mAP	Top-10			Top-20		
		P	R	F1	P	R	F1		P	R	F1	P	R	F1
CLIP (ViT-L/14)	43.18	33.84	41.94	37.46	25.70	53.07	34.63	72.74	10.16	79.19	18.01	5.98	93.21	11.24
CLIP (ViT-B/16)	46.09	33.78	41.86	37.39	25.32	52.30	34.13	70.78	10.38	80.87	18.40	**6.02**	**93.86**	**11.32**
CLIP (ViT-B/32)	43.62	34.18	42.36	37.83	25.66	53.00	34.58	69.22	9.69	75.48	17.17	5.79	90.21	10.88
CLIP (RN50)	40.62	29.75	36.86	32.93	22.57	46.61	30.41	68.70	9.15	71.33	16.23	5.58	87.01	10.49
TagCLIP (ViT-B/16)	43.07	32.46	40.23	35.93	24.28	50.15	32.72	69.95	7.21	56.21	12.78	4.60	71.69	8.64
SPARC (ViT-B/16)	**53.18**	28.28	35.04	31.30	22.54	46.55	30.37	**74.18**	6.16	48.04	10.93	3.83	59.68	7.19
SPQ (Qwen2.5-VL-3B)	49.29	48.33	59.89	53.49	36.35	75.06	48.98	71.49	**11.20**	**87.28**	**19.85**	5.91	92.02	11.10
SPQ (Qwen2.5-VL-7B)	50.75	**49.58**	**61.44**	**54.87**	**37.21**	**76.86**	**50.15**	73.23	10.89	84.82	19.30	5.76	89.81	10.83
SPQ (LLaVA-1.5-7B)	34.60	18.86	23.37	20.87	16.58	34.25	22.35	66.97	9.84	76.63	17.43	5.56	86.57	10.44
SPQ (InternVL3.5-4B)	40.30	35.93	44.52	39.76	30.71	63.42	41.38	65.96	10.70	83.40	18.97	5.75	89.60	10.81

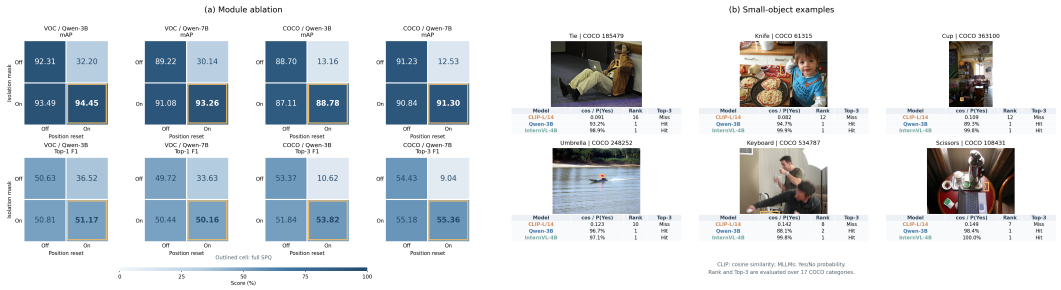
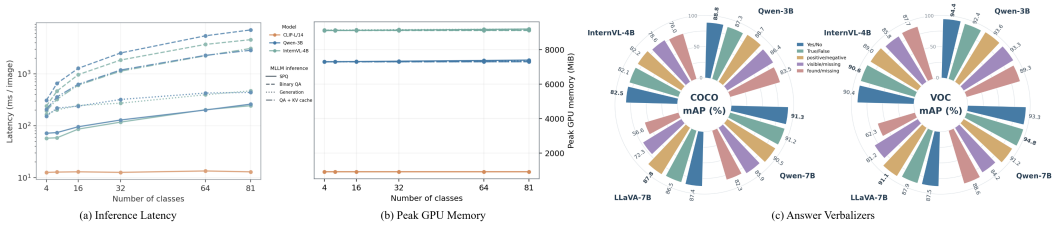


Figure 3: Module ablation and qualitative examples of small-object recognition.

metrics assess category selection within each image. Overall, SPQ demonstrates the potential of generative MLLMs for zero-shot multi-label recognition, while the variation across backbones and datasets indicates that their advantages are not universal.

4.3 Component Ablation and Qualitative Analysis

Interaction between attention isolation and position alignment. Figure 3(a) evaluates all four combinations of attention isolation and position resetting with Qwen-3B and Qwen-7B. Removing isolation while retaining position resetting causes severe degradation. COCO mAP drops from 88.78%/91.30% to 13.16%/12.53%, and Top-3 F1 falls from 53.82%/55.36% to 10.62%/9.04%. By contrast, retaining isolation without resetting positions yields substantially stronger results, although mAP remains below full SPQ in all four model–dataset settings. This asymmetry indicates that position resetting must be interpreted together with the attention structure. Suffix-local positions are reused under the assumption that different suffixes cannot attend to one another.



(a) Inference Latency (b) Peak GPU Memory (c) Answer Verbalizers

Figure 4: Inference Efficiency and Answer Verbalizer Ablation.

Removing both components largely recovers aggregate performance on COCO, reaching 88.70% and 91.23% mAP, but still yields lower Top-3 F1 than full SPQ. Full SPQ achieves the highest mAP in all four settings, although it does not maximize every F1 score. Similar aggregate accuracy after removing both components does not imply computational equivalence to standalone queries. Cross-suffix attention and packed positions change their original computation.

Qualitative recognition of small objects. Figure 3(b) presents six COCO examples with annotated small objects under matched 448×448 whole-image inputs. The highlighted categories fall outside CLIP's top three predictions but appear within the top three for both MLLMs. For example, the tie ranks sixteenth for CLIP but first for both Qwen and InternVL, with Yes/No scores of 93.2% and 98.9%, respectively. Similar differences occur for the distant umbrella and the scissors amid surrounding objects. These cases illustrate how class-conditioned MLLM querying can prioritize objects that receive low rankings under global image–text similarity. Additional examples and selection details are provided in Figure 9 and Appendix A.

4.4 Inference Efficiency

We compare SPQ, independent binary querying, and candidate-list generation on Qwen-3B and InternVL-4B, with CLIP ViT-L/14 as a reference. Measurements use 16 fixed NUS images, 4–81 classes, 224×224 inputs, and three repetitions after warm-up on an A800. Timing boundaries, preprocessing, and caching policies are specified in Appendix A.

As shown in Figure 4, SPQ scales substantially better than independent binary querying. At 81 classes, it reduces latency from 7,051 to 261 ms for Qwen and from 4,570 to 243 ms for InternVL, corresponding to $27.0\times$ and $18.8\times$ speedups, respectively. It is also $1.68\times$ and $1.93\times$ faster than candidate-list generation while providing a score for every candidate class. Total peak memory varies little across inference strategies within each backbone, although SPQ incurs additional memory as the candidate set grows. CLIP remains faster and more memory-efficient; SPQ therefore improves the practicality of MLLM-based recognition rather than eliminating its computational gap with contrastive models.

4.5 Adaptation and Sensitivity Analyses

Answer verbalizer ablation. We compare five answer pairs across four backbones on VOC and COCO, updating both prompt instructions and scoring tokens (Figure 4). Yes/No leads in four of eight settings, while True/False and positive/negative each lead in two. Both pairs containing missing underperform Yes/No throughout; found/missing reduces LLaVA's COCO mAP from 87.45% to 56.56%. Thus, verbalizers are not interchangeable, consistent with prior findings in prompt-based classification (Schick & Schütze, 2021; Gao et al., 2021; Zhao et al., 2021). We retain Yes/No across settings rather than selecting pairs on the test set; additional results appear in Appendix E.

Transfer through seen-category adaptation. For each backbone and dataset, we train three LoRA (Hu et al., 2021) variants: vision-only, language-only, and joint vision–language adaptation. As shown in Figure 5, all three scopes improve COCO mAP and Top-3 F1. Language-only adaptation achieves the highest mAP in five of eight model–dataset settings, while joint adaptation leads in the remaining three, including both InternVL settings. The effectiveness of language-only adaptation suggests that improving the interpretation and readout of pretrained visual features can benefit unseen-category recognition without modifying the vision encoder. One possible explana-

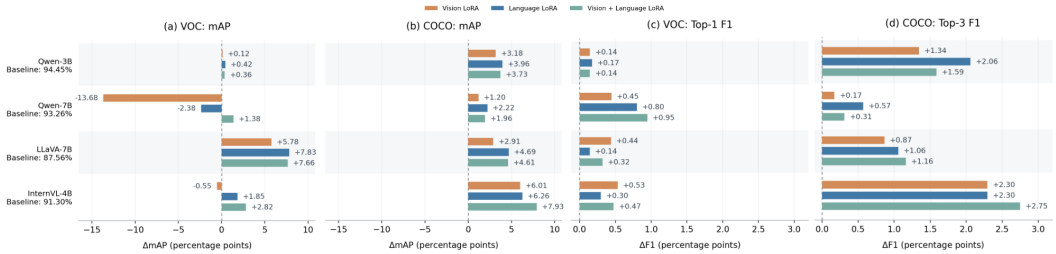
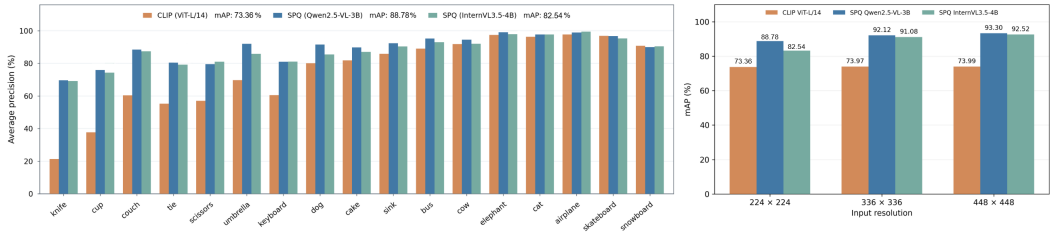


Figure 5: Effect of LoRA adaptation scope on VOC and COCO.



(a) Per-class AP on COCO

(b) Effect of Input Resolution

Figure 6: Per-class performance and input resolution on COCO.

tion is that freezing visual representations preserves their broad semantic coverage, whereas visual adaptation may overfit to seen-category supervision. The 13.68-point VOC mAP decrease under vision-only adaptation of Qwen-7B is consistent with this possibility, although it does not establish the mechanism. Overall, these results show that SPQ supports further gains in zero-shot transfer, with language-only adaptation offering a competitive, but not universally optimal, choice. Full results are reported in Appendix D.

Where does SPQ improve recognition? We examine per-class AP to identify where SPQ-based MLLMs outperform CLIP. The three evaluated vision encoders use 14×14 patches, and Figure 6(a) shows pronounced gains on knife, cup, and scissors, categories often associated with small image regions. In contrast, the gaps are smaller on categories such as elephant and airplane. Together with the qualitative examples, this pattern suggests an advantage in recognizing less visually prominent objects. A plausible explanation concerns the recognition interface: global image–text contrastive alignment may favor dominant scene content, whereas explicit class-presence queries can direct MLLM predictions toward evidence for an individual category. Category-level AP supports this interpretation but does not isolate object size or training objective as its cause.

Does higher resolution close the gap? Figure 6(b) tests whether additional input detail reduces this difference. Increasing resolution from 224×224 to 448×448 raises Qwen mAP from 88.68% to 93.30% and InternVL from 83.14% to 92.52%, whereas CLIP changes only from 73.64% to 73.99%. Both MLLMs retain their advantage at every matched resolution, and the category-level gaps persist in Figure 8. Thus, lower input resolution alone does not explain CLIP's weaker performance in these settings. The results are consistent with differences in how the models use object-level evidence, beyond the availability of additional pixels. This comparison uses positional interpolation of the same CLIP checkpoint; further per-class analysis appears in Appendix B.

5 Conclusion

SPQ provides a structured interface for zero-shot multi-label recognition with generative MLLMs. Shared context, isolated attention, aligned positions, and parallel Yes/No readout yield complete category scores without answer generation. Experiments demonstrate competitive recognition and substantial acceleration over sequential querying, while revealing backbone- and setting-dependent benefits. SPQ supports further adaptation and provides a practical basis for investigating MLLMs in multi-label recognition.

REFERENCES

Muhammad Ali and Salman Khan. Clip-decoder: Zeroshot multilabel classification using multimodal clip aligned representations. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 4677–4681. IEEE, 2023.

Ben Athiwaratkun, Sujan Kumar Gonugondla, Sanjay Krishna Gouda, Haifeng Qian, Hantian Ding, Qing Sun, Jun Wang, Jiacheng Guo, Liangfu Chen, Parminder Bhatia, et al. Bifurcated attention for single-context large-batch sampling. In *Forty-first International Conference on Machine Learning*, 2024.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization. *Text Reading, and Beyond*, 2(1):1, 2023.

Avi Ben-Cohen, Nadav Zamir, Emanuel Ben-Baruch, Itamar Friedman, and Lihi Zelnik-Manor. Semantic diversity learning for zero-shot multi-label classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 640–650, 2021.

Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5327–5336, 2016.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Yanwei Fu, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Shaogang Gong. Transductive multi-class and multi-label zero-shot learning. *arXiv preprint arXiv:1503.07884*, 2015.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pp. 3816–3830, 2021.

In Gim, Guojun Chen, Seung-seob Lee, Nikhil Sarda, Anurag Khandelwal, and Lin Zhong. Prompt cache: Modular attention reuse for low-latency inference. *Proceedings of Machine Learning and Systems*, 6:325–338, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Dat Huynh and Ehsan Elhamifar. A shared multi-attention framework for multi-label zero-shot learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8773–8783. IEEE, 2020.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Jordan Juravsky, Bradley Brown, Ryan Ehrlich, Daniel Y Fu, Christopher Ré, and Azalia Mirhoseini. Hydragen: High-throughput llm inference with shared prefixes. *arXiv preprint arXiv:2402.05099*, 2024.

Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. Multi-label zero-shot learning with structured knowledge graphs. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1576–1585. IEEE, 2018.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023.

Yuqi Lin, Minghao Chen, Kaipeng Zhang, Hengjia Li, Mingming Li, Zheng Yang, Dongqin Lv, Binbin Lin, Haifeng Liu, and Deng Cai. Tagclip: A local-to-global framework to enhance open-vocabulary multi-label classification of clip without training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3513–3521, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Yicheng Liu, Jie Wen, Chengliang Liu, Xiaozhao Fang, Zuoyong Li, Yong Xu, and Zheng Zhang. Language-driven cross-modal classifier for zero-shot multi-label image recognition. In *ICML*, pp. 32173–32183, 2024.

Marcus Ma, Georgios Chochlakis, Niyantha Maruthu Pandiyan, Jesse Thomason, and Shrikanth Narayanan. Large language models do multi-label classification differently. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2472–2495, 2025.

Kevin Miller, Aditya Gangrade, Samarth Mishra, Kate Saenko, and Venkatesh Saligrama. Sparc: Score prompting and adaptive fusion for zero-shot multi-label recognition in vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4313–4321. IEEE, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume*, pp. 255–269, 2021.

Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *Advances in Neural Information Processing Systems*, 35:30569–30582, 2022.

Junjie Wang, Ping Yang, Ruyi Gan, Yuxiang Zhang, Jiaxing Zhang, and Tetsuya Sakai. Zero-shot learners for natural language understanding via a unified multiple-choice perspective. *IEEE Access*, 11:142829–142845, 2023.

Shouwen Wang, Qian Wan, Junbin Gao, and Zhigang Zeng. Language-driven multi-label zero-shot learning with semantic granularity. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1968–1978. IEEE, 2025.

Xiang Wen, Haobo Wang, Ke Chen, Tianlei Hu, and Gang Chen. Gmcot: a graph-augmented multimodal chain-of-thought reasoning framework for multi-label zero-shot learning. *Frontiers of Information Technology & Electronic Engineering*, 26(12):2623–2637, 2025.

Ming-Kun Xie, Zhiqiang Kou, Zhongnian Li, Gang Niu, and Masashi Sugiyama. Unlocking the power of co-occurrence in clip: A dualprompt-driven method for training-free zero-shot multi-label classification. In *International Conference on Learning Representations*, volume 2026, pp. 125344–125357, 2026.

Shiyu Xuan, Dongkai Wang, Zechao Li, and Jinhui Tang. Zero-shot hoi detection with mllm-based detector-agnostic interaction recognition. *arXiv preprint arXiv:2602.15124*, 2026.

Shuo Yang, Zirui Shang, Yongqi Wang, Derong Deng, Hongwei Chen, Qiyuan Cheng, and Xinxiao Wu. Data-free multi-label image recognition via llm-powered prompt tuning. *arXiv preprint arXiv:2403.01209*, 2024.

Jinwei Yao, Kaiqi Chen, Kexun Zhang, Jiaxuan You, Binhang Yuan, Zeke Wang, and Tao Lin. Deft: Decoding with flash tree-attention for efficient tree-structured llm inference. In *International Conference on Learning Representations*, volume 2025, pp. 66657–66688, 2025.

Yanyu Yao. A survey on multimodal large language models. In *Computer Science Undergradaute Conference 2025@ XJTU*, 2025.

Yuncheng Yao, Yuxuan Xia, Shengjie Wang, and Danyang Zhuo. Parallel prefix verification for speculative generation. *arXiv preprint arXiv:2605.04263*, 2026.

Yang Zhang, Boqing Gong, and Mubarak Shah. Fast zero-shot image tagging. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5985–5994. IEEE, 2016.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pp. 12697–12706. Pmlr, 2021.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody H Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, et al. Sglang: Efficient execution of structured language model programs. *Advances in neural information processing systems*, 37:62557–62583, 2024.

Zhen Zheng, Xin Ji, Taosong Fang, Fanghao Zhou, Chuanjie Liu, and Gang Peng. Batchllm: Optimizing large batched llm inference with global prefix sharing and throughput-oriented token batching. *Proceedings of Machine Learning and Systems*, 8:1742–1756, 2026.

A Reproducibility Details

AI use statement

Generative AI tools were used only to assist with the writing of this paper, including language polishing, translation, and improving the clarity and organization of the manuscript. All research-related work, including problem formulation, method development, implementation, experiment design and execution, data analysis, and interpretation of the results, was performed by the authors. The authors reviewed and verified all AI-assisted text and take full responsibility for the content of the paper.

A.1 Evaluation Protocol

Predictions are evaluated over the benchmark-specific unseen category set. We compute AP separately for each category and average the resulting values to obtain mAP. Top-k evaluation selects the k highest-scoring candidate categories for each image. We use $k \in \{1, 2\}$ for VOC, $k \in \{3, 5\}$ for COCO and NUS-WIDE, and $k \in \{10, 20\}$ for OpenImages.

For the locally reproduced COCO and NUS-WIDE supplementary experiments, precision and recall are computed from globally accumulated true positives, with F1 given by their harmonic mean. Specifically, $P@k = \mathrm{TP}_k/(Nk)$, $R@k = \mathrm{TP}_k/\sum_{n,c} y_{nc}$, and $F1@k = 2(P@k)(R@k)/(P@k + R@k)$, where N is the number of evaluated images and y_{nc} is the binary ground-truth annotation restricted to the evaluated categories.

The supplementary COCO experiments use the same 2,729 images and 17 categories across models and resolutions. The supplementary NUS-WIDE comparison uses 46,848 images and 81 categories. Image identifiers, category order, and ground-truth arrays are checked for consistency before comparing predictions.

A.2 Input Processing and Query Construction

In the controlled resolution experiments, each RGB image is resized directly to 224×224, 336×336, or 448×448 using bicubic interpolation, without cropping or dynamic tiling. Each backbone retains its native image normalization. The same checkpoint is used at all resolutions. For CLIP ViT-L/14, the spatial positional embeddings of the original 224-resolution checkpoint are bicubically interpolated, leaving the class-token embedding unchanged. InternVL uses its native positional interpolation, while Qwen preserves its native multimodal positional encoding. Both MLLMs use BF16 for this resolution comparison.

The resolution, efficiency, and qualitative experiments use the same class-presence query as the main evaluation:

```
Answer Yes or No.
Is the following category visibly present in the
image?
Category: {CLASS}
```

where {CLASS} is replaced by the candidate category name. The textual query is incorporated into each backbone's native chat template. The shared prefix consists of the visual tokens and the instruction tokens preceding the category-specific content. Each class-specific suffix retains the tokens required by the corresponding chat template to reach the answer-prediction position. SPQ blocks attention between different class-specific suffixes and assigns each suffix the positional identifiers of its corresponding standalone query. Class-presence scores are computed from the `Yes` and `No` logits according to Eq. 3.

A.3 Efficiency Measurements

Efficiency measurements use an NVIDIA A800 GPU with 80 GB of memory. Models are evaluated sequentially with one image per request. For the scaling experiment, we select 16 fixed NUS-WIDE

images and nested candidate subsets of 4, 8, 16, 32, 64, and 81 classes. Each configuration receives two warm-up calls followed by three repetitions per image, yielding 48 timed observations.

CUDA synchronization is performed around each measurement. Timing includes image preprocessing, device transfers, model execution, and output transfer or text decoding, but excludes checkpoint loading and image decoding from disk. Static binary-query templates and CLIP text features are prepared before timing; the Qwen generation path includes processor tokenization. We report mean latency and PyTorch peak allocated GPU memory.

Independent binary querying evaluates categories sequentially while reusing the shared visual and instructional prefix through key-value caching. The shared prefix is computed once for each image, and only the category-specific suffix is evaluated separately for each candidate category. Candidate-list generation uses greedy decoding with autoregressive KV caching and a limit of 256 new tokens. SPQ uses a dense class-isolated attention mask and reads all decision positions jointly. The reported speedups therefore compare SPQ against a shared-prefix-cached independent-querying baseline rather than a naive implementation that recomputes the shared context for every category. SPQ uses a dense class-isolated attention mask and does not assume a specialized sparse-attention kernel.

A.4 Qualitative Example Selection

The qualitative examples use the 448-resolution COCO predictions from all three models. Candidate examples are restricted to ground-truth-positive categories for which all non-crowd annotated instances have segmentation areas below 32^2 pixels in the original image. We select cases where the target lies outside CLIP's top three predictions but within the top three predictions of both MLLMs, and present six examples spanning different categories. Displayed boxes are ground-truth annotations, not model-generated localizations. The original images are shown at their native aspect ratios for inspection.

These examples are intentionally selected to illustrate recognition differences and are not used to estimate small-object accuracy. The displayed CLIP scores are cosine similarities, whereas MLLM scores are two-answer-normalized Yes/No probabilities.

A.5 LoRA Configuration

We compare adaptation of the vision component, the language component, and both components across four backbones on VOC and COCO. Figure 5 reports changes relative to the corresponding untuned results in Table 1; Table 3 provides the absolute metrics.

We investigate whether seen-category supervision can further improve zero-shot transfer while preserving the SPQ inference interface. For each backbone and dataset, we train three LoRA variants: vision-only, language-only, and joint vision–language adaptation. All remaining components, including the vision–language projector or merger, token embeddings, MLP layers, and language-model head, remain frozen.

Across all experiments, LoRA uses rank $r = 8$, scaling parameter $\alpha = 16$, and dropout 0.05, with no bias adaptation, DoRA, or RS-LoRA. Models are trained for one epoch using AdamW with a fixed learning rate of 10^{-4}, zero weight decay, no learning-rate scheduler or warm-up, and no gradient accumulation. Training is performed with BF16 autocast. The batch size is 16 for Qwen2.5-VL-3B and 8 for Qwen2.5-VL-7B, LLaVA-1.5-7B, and InternVL3.5-4B.

For each training image, we construct exactly one binary class-presence question from the seen-category set. We first identify the seen categories that are present and absent in the image, and approximately alternate between positive and negative questions according to the image index. A category is then sampled from the corresponding positive or negative set using a fixed sampling seed of 2026. Consequently, each image contributes one binary supervision instance per epoch, with an approximately balanced ratio of positive and negative questions. We do not use class-frequency reweighting, focal loss, or a class-balanced sampler.

All models use the same class-presence query as in the training-free setting:

```
Answer Yes or No.
Is the following category visibly present in the
```



Figure 7: Per-class AP comparison on NUS.

```
image?
Category:  {CLASS}
```

where {CLASS} is replaced by a sampled seen-category name. The query is incorporated into the backbone's native chat template, and training uses the original causal language-modeling objective. Tokens corresponding to the image and question are masked from the loss, while cross-entropy is applied only to the assistant answer tokens (*Yes* or *No*). No additional classification head, binary cross-entropy objective, or explicit Yes–No margin loss is introduced.

All training images are converted to RGB and resized directly to 224×224 using bicubic interpolation, without preserving aspect ratio, cropping, dynamic tiling, or data augmentation. At inference time, class-presence scores are obtained from the original language-model head as

$$p_c = \sigma\left(z_c^{\mathrm{Yes}} - z_c^{\mathrm{No}}\right),$$

identically to the training-free SPQ setting.

LoRA is applied only to attention projections. For Qwen2.5-VL and InternVL3.5, the visual attention QKV and output projections and the language-model query, key, value, and output projections are adapted. For LLaVA-1.5, LoRA is applied to the query, key, and value projections of the CLIP vision encoder and to the query, key, value, and output projections of the language model. Multimodal projectors and visual merger modules remain frozen.

B Additional Recognition Results

Category-level results on NUS-WIDE. Figure 7 reports per-class AP for CLIP ViT-L/14, Qwen2.5-VL-3B, and InternVL3.5-4B under the supplementary 224-resolution protocol. Their mAP values are 43.62%, 48.49%, and 45.11%, respectively. Both MLLMs exceed CLIP in aggregate, but their relative performance varies across categories. These results complement the main comparison by exposing variation that is obscured by a single dataset-level average. This controlled supplementary comparison is reported separately from the main-table configurations.

Category-level sensitivity to resolution. Figure 8 expands the resolution analysis to all 17 COCO categories. The three panels share the same category order and AP scale, enabling comparisons



Figure 8: Per-Class AP Comparison on COCO at Different Input Resolutions.

Table 2: Comparison between independent binary querying and SPQ across different MLLM backbones and datasets. Both methods use identical inputs, prompts, and Yes/No scoring rules. Results are mAP (%).

Backbone	Inference	VOC	COCO	NUS-WIDE	OpenImages
Qwen2.5-VL-3B	Independent	94.43	88.80	49.29	71.48
	SPQ	94.45	88.78	49.29	71.49
Qwen2.5-VL-7B	Independent	93.26	91.31	50.75	73.23
	SPQ	93.26	91.30	50.75	73.23
LLaVA-1.5-7B	Independent	87.56	87.38	34.60	66.96
	SPQ	87.56	87.38	34.60	66.97
InternVL3.5-4B	Independent	91.30	82.55	40.30	65.96
	SPQ	91.30	82.54	40.30	65.96

both across models at a fixed resolution and across resolutions for a fixed model. Recognition gaps on categories such as knife, cup, and tie remain visible at matched resolutions. However, individual categories do not improve monotonically as resolution increases, even when aggregate mAP rises. The additional results therefore support persistent category-dependent differences rather than a uniform benefit from higher resolution.

C Equivalence To Independent Querying

Table 2 compares SPQ with standalone independent binary querying across four MLLM configurations and four datasets. The two inference strategies produce nearly identical recognition performance: the maximum absolute mAP difference is only 0.02 percentage points, and the mean absolute difference across all 16 model–dataset settings is 0.005 points. These results empirically support that class-isolated attention and aligned positional assignments preserve the behavior of standalone class-wise queries, up to minor implementation-level numerical differences.

Table 3: Effect of different LoRA fine-tuning scopes on VOC and COCO. We compare adapting both the vision and language components, the vision component only, and the language component only.

Model	LoRA Scope	VOC						
		mAP	Top-1			Top-2		
			P	R	F1	P	R	F1
Qwen2.5-VL-3B	Vision + Text	94.805	35.097	95.390	51.314	18.376	99.890	31.042
	Vision Only	94.571	35.097	95.390	51.314	18.366	99.835	31.025
	Text Only	**94.875**	**35.117**	**95.445**	**51.343**	18.376	99.890	31.042
Qwen2.5-VL-7B	**Vision + Text**	**94.642**	**34.956**	**95.005**	**51.107**	18.336	99.671	30.974
	Vision Only	79.579	34.612	94.072	50.605	18.316	99.561	30.940
	Text Only	90.876	34.855	94.731	50.960	18.336	99.671	30.974
LLaVA-1.5-7B	Vision + Text	95.215	35.279	95.884	51.580	**18.397**	**100.000**	**31.076**
	Vision Only	93.342	**35.359**	**96.103**	**51.698**	18.387	99.945	31.059
	Text Only	**95.386**	35.158	95.554	51.402	**18.397**	**100.000**	**31.076**
InternVL3.5-4B	**Vision + Text**	**94.119**	34.673	94.237	50.694	18.296	99.451	30.906
	Vision Only	90.749	**34.713**	**94.347**	**50.753**	18.215	99.012	30.769
	Text Only	93.154	34.552	93.908	50.517	**18.316**	**99.561**	**30.940**

Model	LoRA Scope	COCO						
		mAP	Top-3			Top-5		
			P	R	F1	P	R	F1
Qwen2.5-VL-3B	Vision + Text	92.515	38.854	96.540	55.408	23.862	98.816	38.442
	Vision Only	91.961	38.683	96.115	55.165	23.818	98.634	38.371
	Text Only	**92.743**	**39.184**	**97.360**	**55.879**	**23.877**	**98.877**	**38.465**
Qwen2.5-VL-7B	Vision + Text	93.261	39.037	96.995	55.670	**23.972**	**99.272**	**38.619**
	Vision Only	92.498	38.940	96.753	55.530	23.899	98.968	38.501
	Text Only	**93.524**	**39.221**	**97.451**	**55.931**	23.957	99.211	38.595
LLaVA-1.5-7B	Vision + Text	91.987	**39.123**	**97.208**	**55.792**	**23.994**	**99.363**	**38.654**
	Vision Only	90.286	38.915	96.692	55.496	23.921	99.059	38.536
	Text Only	**92.065**	39.050	97.026	55.687	23.921	99.059	38.536
InternVL3.5-4B	**Vision + Text**	**90.473**	**38.610**	**95.933**	**55.060**	**23.914**	**99.029**	**38.524**
	Vision Only	88.552	38.292	95.144	54.607	23.701	98.149	38.182
	Text Only	88.801	38.292	95.144	54.607	23.738	98.300	38.241

D Additional Adaptation Results

Table 3 reports the complete results for all three LoRA scopes, including top-k precision, recall, and F1. The rankings of adaptation scopes differ across metrics. For example, on VOC, language-only adaptation gives LLaVA its highest mAP of 95.386%, whereas vision-only adaptation achieves its highest Top-1 F1 of 51.698%. On COCO, language-only adaptation gives Qwen-7B its highest mAP of 93.524%, while joint adaptation achieves its highest Top-5 F1 of 38.619%.

These differences show why adaptation should be evaluated using both per-class ranking across images and label selection within images. They also qualify the apparent advantage of language-only adaptation: it achieves the highest mAP in five of the eight model–dataset settings, but does not consistently maximize top-k F1. The full results thus support choosing adaptation scope according to the backbone and evaluation objective, using validation data rather than test-set performance.

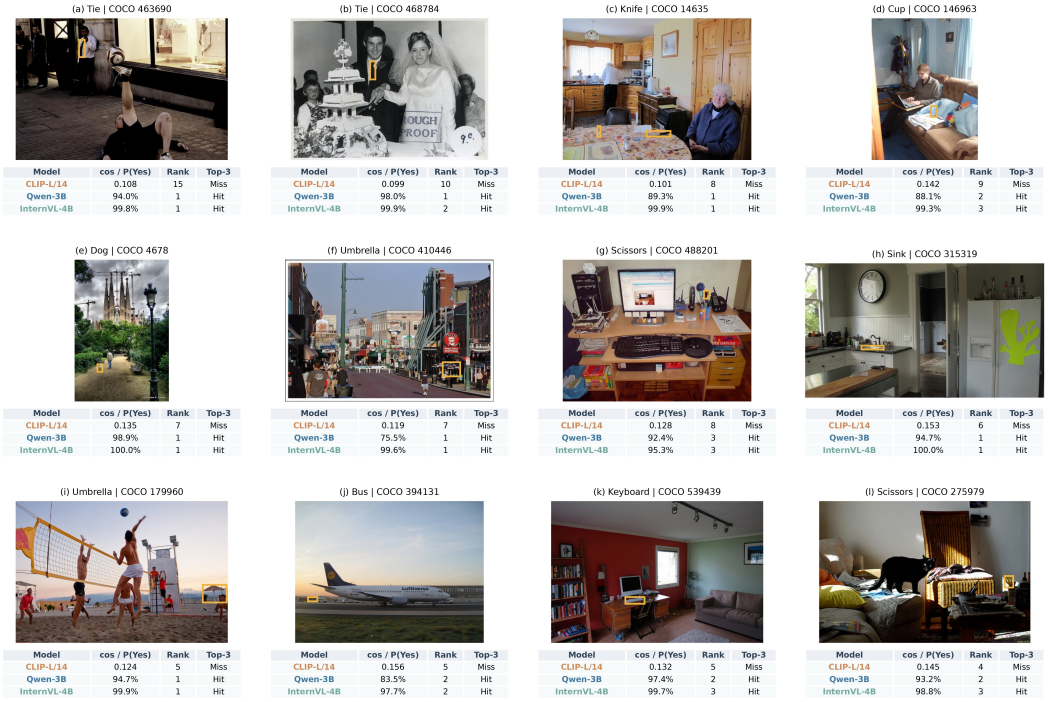


Figure 9: Additional qualitative examples of small-object recognition on COCO.

E Additional Verbalizer Analysis

The best answer pair varies with the backbone and dataset. On VOC, True/False increases Qwen-7B mAP from 93.26 to 94.82, while positive/negative increases LLaVA from 87.50 to 91.09. In contrast, visible/missing and found/missing underperform Yes/No in all eight settings. These observations motivate using a fixed default for comparisons and reporting sensitivity, rather than selecting an answer pair on the test set. They do not identify whether tokenization, pretraining frequency, or instruction alignment causes the differences.

Additional qualitative examples. Figure 9 presents 12 additional COCO examples spanning nine categories, including small foreground objects and distant background instances. All models receive the same 448×448 whole-image input. Both MLLMs rank the highlighted categories within the top three, whereas CLIP ranks them between fourth and fifteenth among the 17 candidates. For example, the tie in panel (a) ranks fifteenth for CLIP but first for both MLLMs. In panel (j), both MLLMs retain the distant bus among their top two predictions despite the prominent airplane. Similar differences occur for the scissors amid desktop clutter in panel (g) and the keyboard in panel (k).

These cases illustrate how MLLM-based class-presence querying can prioritize annotated objects that global image–text matching ranks lower, across variations in scale, clutter, and scene composition. CLIP cosine similarities and MLLM Yes/No scores are not directly comparable; category ranks and Top-3 inclusion provide the common evaluation criterion. Because the examples are selected to illustrate successful recognition differences, they complement the quantitative results rather than establish a dataset-wide advantage on small objects. Selection details are provided in Appendix A.